



Synergetic **Content Creation & Communication**

User Evaluation Report

Project Reference No.	FP7 - 231854
Deliverable No.	D7.4.1
Workpackage No.	WP 7: User requirements, user evaluation and specifications
Nature:	R (Report)
Dissemination Level:	PU (public)
Document version:	1.0
Date:	07/02/2011
Editors(s):	Eric Karstens, Liliana Bounegru (EJC)
Document description:	The final version of the first User Evaluation Report documents the user evaluation plan and methodology, the results of the first focus group on the user interface, and the results of the four evaluation rounds of the first SYNC3 prototype with representatives of all target user groups: journalists, media analysts and bloggers.

History

Version	Date	Reason	Revised by
0.1	02/09/2010	Internal review	EJC
0.2	07/09/2010	Internal review	EJC
0.3	24/12/2010	Intermediate user evaluation results	EJC
0.4	29/12/2010	Internal review	EJC
0.5	01/02/2011	Final user evaluation results	EJC
0.6	06/02/2011	Peer review	RIAN, NCSR'D'

Authors List

Organisation	Name
EJC	Liliana Bounegru (author)
EJC	Eric Karstens (author and internal reviewer)

Executive Summary

This document is the final version of the first user evaluation report which was due M20 (30 November 2010), two months after the planned release date of the first SYNC3 prototype in M18 (September 2010). On grounds of the delay in the development of the first SYNC3 prototype and its release on 23 November 2010, the release of the final version of this report was shifted by about two months to the end of January 2011 (M22). The aim the User Evaluation Report is to document the evaluation process and its results as a way to enable dissemination of the data collected during the user evaluation rounds and offer the developers the opportunity to properly analyze the user feedback data and use it as documentation for ongoing improvement of functionality and usability aspects of the SYNC3 system and for identifying bugs in the system.

This report outlines the plan for the first round of user evaluation of the SYNC3 system and describes the concrete steps taken during the user evaluation process. To this end, it establishes the background of the methodology applied for the first round of user evaluation of the SYNC3 system. It includes the overview and conclusions of the focus group on the user interface organized at the EJC together with L3S on 30 March 2010, and the complete results of the four rounds of user evaluations of the first SYNC3 prototype, organized at the EJC headquarters in Maastricht on 25 November 2010, Report International headquarters in London on 16 December 2010, Bortun-Olteanu PR Agency in Bucharest on 19 January 2011, and the news agency RIA Novosti on 20-21 January 2011. It will be extended into a second and third user evaluation report as soon as testing of the second and third prototype will have taken place. The second prototype is planned to be released in M25, April 2011, and the third prototype in M31, October 2011.

The SYNC3 concept and intention were met with the approval of the test participants. Particularly well-received was the capacity to enable better understanding of the dynamics between traditional and social media by linking together news articles with blog posts that relate to them. The alternative ways of visualization of generated news and blog results, the Overview and the Map View, were met with great interest, although the Overview particularly produced some confusion as to what it represents. A series of suggestions to increase their user-friendliness has been collected during the evaluations. Although well-received at conceptual level, the areas most often indicated by the test users as requiring further development and improvement in order to become usable were the relevance and ranking of generated results to user queries, and the Workspace. Furthermore, missing functions or limited functionalities were noted by participants, yet condoned in the context of the beta version of the SYNC3 system on hand.

Table of Contents

- Executive Summary3
- List of Figures6
- List of Tables7
- List of Terms and Abbreviations8
- 1. Introduction.....9
 - 1.1. Goals and objectives of WP7, task 7.49
 - 1.2. Task 7.4 description and responsible contributors9
 - 1.3. Task 7.4 testing and validation stages 10
 - 1.4. Specific objectives of task 7.4 10
 - 1.5. Relationship between user evaluation and other WPs 11
 - 1.6. Scope of the user evaluation 11
- 2. Principles of Validation..... 13
 - 2.1. Definition of user validation 13
 - 2.2. Relevant validation standards..... 14
 - 2.3. The role and relevance of validation 14
 - 2.4. Planning of the validation process 18
 - 2.4.1. Requirements validation..... 18
 - 2.4.2. Validation and Verification 19
 - 2.4.3. Level of confidence 20
 - 2.4.4. Test early and often 20
 - 2.4.5. How many tests? 21
 - 2.4.6. Identifying future problems 24
- 3. SYNC3 user evaluation methodology 25
 - 3.1. SYNC3 objectives to be evaluated..... 25
 - 3.2. Prototypes and evaluation scenarios 26
 - 3.3. User evaluation methodology..... 28
 - 3.4. Survey of exploitation opportunities..... 29
- 4. SYNC3 user evaluation by use case..... 30
- 5. The user evaluation process 35
 - 5.1. Recruiting user evaluators..... 35
 - 5.2. Expected number of tests..... 35
 - 5.3. Qualification of the user evaluators 36
 - 5.4. Test environment and technical prerequisites 36
 - 5.5. Managing user evaluation risks..... 36
- 6. Intermediate conclusion..... 38
- 7. Results of the first user interface focus group..... 39
 - 7.1. Set-up..... 39
 - 7.1.1. Agenda..... 39
 - 7.1.2. Participants..... 39
 - 7.2. Outcome of the Focus Group..... 40
 - 7.2.1. Search Function 40

7.2.2.	Workspace Function	41
7.2.3.	Workspace Personalization and Recommendations.....	42
7.2.4.	General Comments	42
7.3.	Outcome from the individual testing.....	43
7.4.	Conclusions of the focus group on user interface	43
8.	Results of the 1 st prototype user evaluations	45
8.1.	Evaluation framework.....	45
8.2.	Questionnaire results.....	49
8.3.	Evaluation session results.....	60
8.4.	Table of issues raised by users.....	65
9.	Conclusions from 1st prototype user evaluations.....	69
10.	References	70

List of Figures

Figure 1: Projects Show Steady Improvement.....	15
Figure 2: Static software verification.....	19
Figure 3: 3 users will discover 85% of an application’s usability problems	23
Figure 4: The V-model of development	24
Figure 5: SYNC3 approach.....	26
Figure 6: Visual representation of SYNC3 use cases related to the search function	32
Figure 7: Visual representation of SYNC3 use cases related to search results visualization	32
Figure 8: Screenshot of the SYNC3 Workspace	42

List of Tables

Table 1: IEEE Standards Overview	14
Table 2: Success or failure of software projects.....	15
Table 3: Recipe for Success: CHAOS 10.....	16
Table 4: Traditional vs. Simple testing.....	22
Table 5: Evaluation-related deliverables.....	26
Table 6: Correlation of SYNC3 use cases with the first integrated prototype	27
Table 7: Summary of SYNC3 use cases and the status of their implementation in the 1 st prototype	30
Table 8: Overview of user evaluation dates and locations	45

List of Terms and Abbreviations

Abbreviation	Description
--------------	-------------

1. Introduction

1.1. Goals and objectives of WP7, task 7.4

The project DoW describes the general intention of the SYNC3 user evaluation activities in section B1.3.1.VII WP7 as follows:

This task will deal with the planning and coordination as well as with the actual process of user evaluation of all system prototypes. Three prototypes are foreseen in the duration of the project, so three cycles of user evaluation are definitely foreseen, although more evaluation workshops may be organised upon demand of the WP leaders in order to receive feedback from users on specific aspects of the SYNC3 tool (e.g. interface 'look and feel', etc.). EJC will be responsible for the evaluation of the prototypes, coordinate the whole process and will provide users for the testing.

1.2. Task 7.4 description and responsible contributors

The three prototypes are foreseen in the duration of the project in M18 (September 2010), M25 (April 2011), and M31 (October 2011). EJC will be responsible for organising all the evaluation workshops and gathering the necessary professional users who will test the system prototypes. EJC will be responsible for the definition of the evaluation scenario of the system prototypes, i.e. the process the users will go through and the means of recording their reactions and comments.

The three integrated prototypes of SYNC3, as foreseen in the DoW, will thus go through validation and testing. Where necessary and sensible for the process of development, intermediary integrated prototypes or user interface prototypes will be submitted to user testing as well. Results of the validation will be fed into subsequent development cycles.

The work is to be carried out in an iterative way, based on proven methodologies to be selected at the start of task 7.4. It will follow the "design - validate - redesign - validate & redesign" model. The process of validation will be performed using standard validation techniques that conform to industrial practice.

Following this model, a first focus group on the user interface was already organized by EJC and L3S on March 30, 2010. Its outcome and conclusions were communicated to the SYNC3 consortium in an internal report¹. WP6 leader L3S took corresponding actions to address the results of the validation in the following development cycles of the user interface. A second focus group on the user interface is planned for mid October.

7 types of users will be called upon to test and validate the system, representing the target groups of SYNC3:

- Professional journalists and editors, both freelance and permanently employed, interested in getting an overview of news items from traditional media and following how they were received in the blogosphere;
- Citizen journalists and bloggers (non-professional journalists) who play an active role in the process of collecting, analysing and disseminating news;
- Communication, Marketing, and PR professionals within corporations and communication agencies, interested in brand monitoring and product perception in online news sources and blogs;
- Policy makers as well as influence groups, such as lobby groups, interested in following the public debate on current issues in the news and blogosphere;
- Media monitoring/media intelligence professionals, a subcategory of PR professionals specialized particularly in documentation and analysis of media content;

- Academic researchers specializing in comparative media studies interested in comparative analysis of content in traditional media and new media (blogs);
- General public (media consumers).

1.3. Task 7.4 testing and validation stages

User evaluation depends on the technical SYNC3 partners delivering the two preliminary prototypes: basic and intermediate, and the final prototype, which are foreseen for month 18 (September 2010), month 25 (April 2011), and month 31 (October 2011), respectively. The actual validation sessions must be scheduled as soon as possible after the prototype releases, since user evaluation reports are supposed to be due within two months after, specifically in month 20 (November 2010), month 27 (June 2011), and month 33 (December 2011), respectively. Insights from these reports are supposed to help developers refine the different modules as well as to gain additional insights into user needs and requirements.

On grounds of the delay in the development of the first prototype, it was necessary to shift the release of his report by about two months to the end of January 2011 (M22). While the delay did not entail time constraints in performing the planned evaluation sessions and no quality concessions in the collection of user feedback, it did determine the partial overlap of the evaluation period with the winter holidays season, which led to several weeks of inactivity in terms of organizing evaluation sessions, in the middle of the evaluation period.

The procedure in the three evaluation cycles is essentially similar: to validate the SYNC3 system, groups of users from each of the seven target groups are presented with the integrated prototypes by way of the graphical user interface (GUI) in order to collect information on usability, functionality and stability of the SYNC3 system from the user perspective. This includes (contingent on respective prototype functionality) factors such as: search capabilities: search facilities, search/retrieval performance, proportion and quality of sources coverage, 'freshness' of listed results, meaningfulness/relevance and precision of retrieved records, response time, user effort (e.g. in designing queries to obtain satisfactory results); output form: quality/value/relevance of displayed content from extracted records; user interface and documentation: user friendliness, intuitiveness, user effort in learning and using the interface, comprehensiveness, quality, usefulness of documentation.

Collecting this external view is supposed to create a flow of valuable, independent and non-blinkered insights into how to optimally tailor the single modules into a working system and how to optimise the user interface.

1.4. Specific objectives of task 7.4

In tune with the overall SYNC3 vision, and based on the general objectives above, the specific goals of task 7.4 can be described as follows:

- Develop a detailed and appropriate testing strategy based on current research and best practices to provide the SYNC3 project with professional validation plan;
- Develop a plan to validate the SYNC3 product with users from each target group and implement this plan;
- Establish a proactive identification of possible problems, e.g. a "fail fast" indicator in order to avoid follow-up problems in the case of usability and functionality issues of a single module or inadmissible user interaction with the system as a whole;
- Provide insights from validation and testing to help developers create an innovative system, to steer the project as whole, and ensure the high quality of the final product;
- Create a validation and testing process that is pragmatic, yet simple and effective;
- Ensure high visibility of the testing results in order to provide developers with insights for the finalisation of single modules as well as for overall system configuration;

- Quickly disseminate evaluation results, in order to make the testing meaningful and timely for developers;
- Document the results of the validation process for future reference and as a basis for user training;
- Establish a proper basis for initiating the next phase of the development cycle;
- Ensure that the completed end product complies with established software and system requirements;
- Satisfy non-technical standards, policies, practices, procedures, and conventions, where applicable;
- Help create a usable and useful final product.

1.5. Relationship between user evaluation and other WPs

The task of user evaluation is closely intertwined with other activities of the SYNC3 project. In general, it is assumed that the findings from these WPs and the documented status of the different work packages are stable and remain basically unchanged. Every member of the project able to identify significant changes that are relevant to testing and validation is asked to immediately send notice to the EJC in the case of wrong or outdated assumptions in this test and validation document.

The most important basis for task 7.4 is the User Requirements and Specifications Report (D7.1), which has laid down the relevant framework for the SYNC3 system that was agreed upon within the consortium, in particular:

- Potential user groups, i.e., professional journalists, citizen journalists and bloggers (non-professional journalists), communication professionals, policy makers, and the broader public (media consumers). Two other potential user groups were later identified, i.e., media monitoring professionals and academic researchers specializing in comparative media studies;
- Scenarios of use for each user group, including basic description of planned functions, aiming to provide potential users with a good understanding of the intention, functionality and use of SYNC3;
- User requirements overview, collected via user surveys (questionnaires), focus groups (interviewing), monitoring of existing systems and example scenarios (use cases).

Furthermore, all work packages related to technical development, namely WP3 *News Clustering and Excerpt Extraction*, WP4 *Blog processing*, and WP5 *News event labelling and relations*, WP6 *System architecture, interfaces and integration*, as well as WP8, *System architecture and integration*, have foreseen their own evaluation metrics as part of a self assessment plan.

The assessment plan outlines the performance measures and evaluation metrics for each technical work package. The evaluations in WPs 3-5 are supposed to make sure that the respective system component is fully functional, both in terms of technology and programming, and as far as the respective processing of content is concerned.

WP8 will carry out systematic evaluations to measure the overall system performance. The coordination of the testing activities to ensure that the integration meets the documented requirements is established as phase three in this work package.

1.6. Scope of the user evaluation

The scope of this user evaluation plan is to ensure functionality and usability from a user perspective. To reach this goal, validation must be combined with verification to some extent. Validation is not a method to identify software defects or bug tracking. This is a practice generally described as verification. This verification will be executed as part of the individual development practices.

The abovementioned evaluation tasks must consequently be considered separately from the user evaluation efforts discussed in the present document. They are of a purely technological nature and must be carried out by engineers and software developers, whereas the user evaluation will be performed by journalistic organizations with limited engineering expertise, yet a clear understanding of the practical benefits to be reaped from SYNC3 in the context of everyday workflow.

However, the successful completion of the technical testing is a prerequisite for any kind of user evaluation. If user interaction with the prototypes founders on error messages, poor performance and system stops or breakdowns which are not caused by the human interaction itself, test users will become frustrated early on in the process, confidence in the SYNC3 tool as a whole will be undermined, and scarce or not even any meaningful user feedback will be elicited. Therefore, even the basic prototype must be as technically robust as possible.

Also, the technical partners should be on standby at the time of user testing to resolve technical issues quickly and to assist with questions arising on usability and operation of the system. The EJC will co-ordinate the testing dates with the technology partners and organize telecommunications (such as Skype) to make sure that timely assistance can be provided when needed.

2. Principles of Validation

This chapter provides an overview of why and how testing in software development is necessary, as well as of which procedures create value, ensure quality and enhance visibility of progress towards the final product.

Questions to be answered are:

- What is the benefit of testing and validation in software development?
- Which best practices can be identified?
- How can the findings be applied to test and validate the SYNC3 system?

Current research reviewed in this chapter shows that a number of insights can help significantly if applied to this particular project. The following pages therefore provide an overview of the factors relevant to the software development process in general and particularly for testing and validation.

At the end of each sub-chapter the relevant insights for SYNC3 are highlighted, in order to provide an overview of the practices that have been applied in developing the actual validation plan for SYNC3, which is described in Section 4.

2.1. Definition of user validation

Validation: assuring that a software system meets the user's needs

This statement is a short and rememberable definition of validation. A clear understanding of the term is important, otherwise assumptions what it means could lead to some confusion regarding the specific scope of testing and validation.

For example, the following definition from a technical standpoint is too narrow as it describes validation as “the process of evaluating software at the end of the software development process to ensure compliance to software requirements”. Starting to validate once the project nears completion would be too late to uncover flaws, usability issues and integration issues. Therefore the following paragraphs will discuss the meaning of validation from different angles.

Broader definitions and meanings of the term “validation”:

“In general, validation is the process of checking if something satisfies a certain criterion. Examples would be: checking if a statement is true (validity); if an appliance works as intended; if a computer system is secure, or if computer data are compliant with an open standard. Validation implies one is able to testify that a solution or process is correct or compliant with set standards or rules.”ⁱⁱ

A technical definition of the term “validation”:

“In Computer Programming Terminology, validation refers to the process of controlling that data inserted into an application satisfies predetermined formats or complies with stated length and character requirements and other defined input criteria. (Verification usually requires human judgement).”ⁱⁱⁱ

The definition used in the course of the SYNC3 project:

The definition that we use to describe and understand the scope, the desired outcome of the testing and validation process is the one below, differentiating between the objectives from two points of view:

Validation: “Are we building the right product?”

e.g. is the software usable for the end users?

Verification: “Are we building the product right?”

e.g. are there defects or bugs in the code?

As for the difference between “validation” and “verification”, it is important to note that verification is a related, yet very different concept and requires specific procedures to be applied properly. Verification essentially means to ensure that the software has no serious defects or flaws, such as software bugs.

It should be noted that tracking and resolving defects is not in the scope of the particular task discussed in this document. As mentioned above, the user validation tests will be used to look at the modules from a user perspective. Ensuring that the modules work from a technical view will be a task of their respective technical developers.

However, should any bugs be uncovered during the validation process, they will be immediately communicated to all stakeholders affected.

2.2. Relevant validation standards

As a formalized process, standards defined by either IEEE or ISO describe processes to ensure quality standards. Proper validation is a tool in the development process in order to reach both a level of confidence that the product fulfils the users' needs as well as meeting so-called “minimal” criteria of quality standards.

The most relevant standard is IEEE 1012 (Software Validation & Verification Plan, SVVP), which provides an outline covering relevant aspects of validation. This standard is part of a whole system of standards applied to software development from different perspectives and at different stages of development.

Table 1: IEEE Standards Overview

Acronym	Full name	Identification
SQAP	Software Quality Assurance Plan	IEEE 730
SCMP	Software Configuration Management Plan	IEEE 828
STD	Software Test Documentation	IEEE 829
SRS	Software Requirements Specification	IEEE 830
SVVP	Software Validation & Verification Plan	IEEE 1012
SDD	Software Design Description	IEEE 1016
SPMP	Software Project Management Plan	IEEE 1058

2.3. The role and relevance of validation

Software Development is an area of knowledge still gaining importance. The reason for this is dependency of businesses, consumers and public services on software-based services working without major flaws. Another driver is the need to ensure quality of ever more complex software systems. As a result, any knowledge that supports the development of innovative, productive and easy to use software is becoming an important differentiator in a competitive economic environments as well as public services.

Against this background it is easier to understand why the software creation process has been the target of many studies and research projects that tried to gather knowledge of how to achieve desired project outcomes. Statistically, many software projects run into trouble - the

number of failed or challenged software projects runs up to 70 %. There are numerous reasons for this, the main issue being complexity.

The most complete set of data tracking the success of software projects was supplied by the Standish Group, a US research firm that focuses on mission-critical projects, that has conducted extensive research in order to determine success rates of small or large software projects.

Since 1994 the Group publishes an annual report called the CHAOS Study, which is based on statistical data from over 30.000 software projects tested. The key question was how many projects succeeded or failed. From this extensive set of data the Standish Group developed general recommendations to ensure software development success.

The Standish Group study categorizes projects into three resolution types:

Table 2:Success or failure of software projects

Category	Description
Successful	The project is completed on time and on budget, with all features and functions originally specified.
Challenged	The project is completed and operational, but over budget, late, and/or with fewer features and functions than initially specified.
Failed	The project is cancelled before completion, or never implemented.

Source: Standish Group, 2001

The good news is that in recent years success rates are up across the board, while cost and schedule overruns are declining. The CHAOS research timeline provides evidence of steady improvement in IT project management. Quote: "In 1994, only 16% of application development projects met the criteria for success/completed on time, on budget, and with all features/functions originally specified. In 2000, 28% of all projects were in the successful column."^{iv}



Figure 1: Projects Show Steady Improvement

“Tracking U.S. project outcomes showed that in 1994, 28,000 projects were successful. In the year 2000, that number increased to 78,000 – almost a threefold climb. Conversely, failed projects amounted to 54,000 in the 1994 study vs. 65,000 in the 2000 study. This constituted an 18% increase, while overall project growth exceeded 60%. Challenged projects grew at a rate of 62%, to equal 137,000 over the 1994 number of 93,000.”

Success factors in software development

The first so-called CHAOS study, conducted in 1994, identified 10 success factors, called the “CHAOS 10.” This list provides an overview of which factors have a high positive or negative impact on software projects. Used proactively, these factors can be viewed as early indicators to avoid failure and as a means to define priorities early on.

Based on new data from the annual studies, the Standish Group has updated the CHAOS 10 and published a new ranking of the factors in the year 2000. Although no project requires all 10 factors to be successful, the more factors present in the project strategy, the higher the confidence level.

User involvement, which is the key task of validation, is ranking high on the list. It is the second most important factor to ensure successful software development and has been the most important factor in past studies.

Table 3: Recipe for Success: CHAOS 10

Recipe for Success: CHAOS 10 Each factor was weighted according to its influence on project success. The more points achieved, the lower the project risk.		Success factor
1	Executive support	18
2	User involvement	16
3	Experienced project manager	14
4	Clear business objectives	12
5	Minimized scope	10
6	Standard software infrastructure	8
7	Firm basic requirements	6
8	Formal methodology	6
9	Reliable estimates	5
10	Other criteria	5

Source: The Standish Group, 2001^{vi}

The following list is an excerpt from a 2001 article in *Softwaremag* and describes the role of individual factors in more detail. It is important to understand that not one single factor is the key to success. Rather, an appropriate balance considering all factors must be found for a given project.

Executive support

Traditionally, executive support occupied the No. 2 spot; however, it is now the No. 1 factor in project failure. Executive support influences a project's process and progress. Lack of executive input can jeopardize a project.

User involvement

Lack of user involvement traditionally has been the No. 1 reason for project failure. Conversely, it has been the leading contributor to project success. Even when delivered on time and on budget, a project can fail if it doesn't meet user needs or expectations. However, this year user involvement has moved to the No. 2 position. Despite how that may sound, user involvement hasn't decreased in importance; it's just that IT professionals have, in effect, solved this major problem.

Experienced project manager

Ninety-seven percent of successful projects have an experienced project manager at the helm.

Clear business objectives

This factor has moved down one spot because evidence shows that experienced project managers increase success rates.

Minimized scope

Wrapping up the top five is minimized scope. Time is the enemy of all projects, and since scope affects time, or project duration, they are linked. Clearly then, minimizing scope increases a project's chances of success. Minimized scope has replaced small milestones. While these two factors are similar, the act of minimizing scope leads to greater success than does creating small milestones. Concentrating on the top five will result in 70 success points.

Standard software infrastructure

Requirements are in a state of constant flux, but infrastructure needs stability. The Standish Group's research shows that 70% of application code is infrastructure. Some of this code is unique to the application; nonetheless, much of this code could be purchased from an infrastructure vendor. By using standard infrastructure, the application development team can concentrate on business rules rather than on technology. Many application development projects fail not in stand-alone application development, but in existing application integration. Standard infrastructures can shortcut application integration.

Firm basic requirements

The word "basic" refers to base-level requirements. Creating minimal, obtainable base requirements and then developing those features will reduce the effect of change. Delivering minimal features allows users and executive sponsors to see quick results. As a result, project managers are better prepared to articulate the needs and priorities of the next project phase.

Formal methodology

This provides a realistic picture of the project and resources committed to it. And it results in steps and procedures the team can reproduce and reuse. It also enables the team to maximize consistency. And it incorporates lessons learned into active projects. The process encourages a go or no-go decision checkpoint. It also helps the project team proceed with a higher level of confidence, or halt or alter steps to fit changing requirements. CHAOS research shows that 46% of successful projects use a formal project management methodology, compared with 30% of challenged and failed projects. So, this factor should increase success rates by about 16%.

Reliable estimates

Systematic project estimating must be approached realistically, because estimating is just plainly hard. Then add to that the difficulty of developing, purchasing, and integrating components into existing and packaged applications, and outside services. IT managers must use all their collective knowledge and experience to come up with estimates that reflect the true effort required.

Other criteria

In last place is a collection of other factors. These factors include small milestones, proper planning, competent staff, and ownership. In the past, each of these factors was given its own category.^{vii}

Relevant finding for the SYNC3 user validation plan:

The study confirms the paramount importance of user involvement in the software development process. Well-implemented rounds of user evaluation will therefore not only serve to assist the engineers in optimising software functionality and usability, but also enhance the no less crucial end-user acceptance of the eventual product.

2.4. Planning of the validation process

There are many different ways in which validation tests can be planned and applied. A short overview helps to identify the testing process best applicable to SYNC3.

2.4.1. Requirements validation

Designing a system test forces a development team to deeply understand the requirements. The better these requirements are visible, the earlier incompleteness, ambiguity, and inconsistency can be identified. Correcting such problems early will speed up development and reduce the number of late requirements changes.

There are three basic methods of evaluation:

Ad-Hoc testing

- “Just see if you can break it”
- Make up test cases “on the fly”
- Human interpretation of requirements

Systematic testing

- Driven by explicit quality assurance goals
- Test designed for comprehensive coverage
- Tests specify expected output as a benchmark

Automated testing

- Driven by explicit quality assurance goals
- Test suite designed for comprehensive coverage
- Scripts need no human judgment

Relevant finding for the SYNC3 test and validation plan:

Requirements validation has been covered mostly in earlier stages of the SYNC3 project. Still, as this is an ongoing concern, this topic will be part of the validation process as well. Also, using a systematic testing approach is needed given the innovative character of SYNC3 with a focus on the integration of innovative modules such as sentiment analysis. Automated testing will not work with SYNC3, because the entire system

usability and benefits hinge on user interaction.

2.4.2. Validation and Verification

There is a strong dependency between validation and verification. Only a full static and dynamic verification plus a structured validation process will provide full coverage in the end.

Verification must be done regularly to identify and eliminate flaws and defects of the software. This task is usually very complicated when a large project is cut into different modules that need to be integrated later in the process.

Validation can identify whether a user can execute a task with or without training. It can answer the question whether the graphical user interface (GUI) is intuitive or needs detailed and sophisticated help documentation. For example, a test can verify whether a user is able to perform a search and retrieve the news event that he is interested in, and use the various options that the system provides for exploring a news event: exploring relations between events, getting on overview of the specific items referring to them from traditional media sources and blogs, exploring the news event based on time, location, causal relations, or sentiment of the authors of blog entries that discuss the news event of interest, use the workspace feature to select news events or news items, create relations between them and add own comments and opinions on them and use the **sharing and collaboration features to create their own profiles, groups of friends and themed groups, and sign up for updates.**

Validation therefore is not the tool to create user-friendly interfaces only. Complex tasks might require complex software with many features which are not necessarily intuitive and self-explanatory. Validation is also a very effective way to learn which training and documentation is needed in order to gain acceptance.

Verification, on the other hand, is needed to ensure that the software runs without major defects, e.g. that a video starts when a button is pressed, that large files can be processed without stalling the system, or that stored data is secure and can be reliably found when a search is initiated. Verification can be executed as a static or a dynamic process. Static verification usually refers to a software inspection or a code analysis. Dynamic verification is executed with test data to check how the system is working under load.

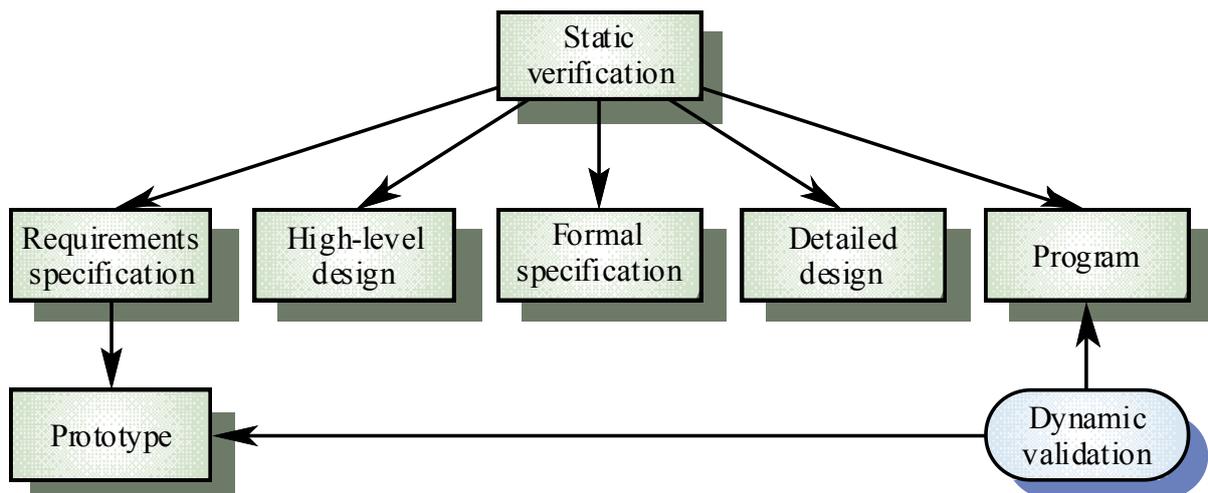


Figure 2: Static software verificationviii

Relevant finding for the SYNC3 test and validation plan:
 The dependency between validation and verification is an issue that must be covered in the test plan. Validation and verification are two sides of the same coin. Verification

must be completed to a high degree before each round of validation, while validation will provide feedback for modifications which will require verification in turn before user testing.

2.4.3. Level of confidence

The “level of confidence” describes a status when many foreseeable factors are under control. This relates to different aspects such as functionality, integration, time, budget, etc. The term is also helpful to keep the development confined to the key features and avoid the uncertainty that is caused by too many changes.

Finishing a project when it is “good enough”

Verification and validation (V&V) establish confidence that the software is fit for its purpose. This term applies strictly to a particular version or release of software in the development lifecycle. Therefore, this process will normally not result in software free of defects or including every conceivable feature. More to the point, the goal is to be able to finish a project and get to a product that is “good enough”. This level of quality is usually the goal of standards such as IEEE or ISO which describe the minimum level of quality that is needed.

Releasing a system when it is “fit for purpose”

The required level of confidence can differ from system to system. High security software needs a higher level, while broadly used systems must ensure that, e.g., user data or stored content cannot be altered. Therefore it is important to define a set of criteria that describe what to test during a validation and verification process. The specification usually referred to in addition to “good enough” is usually to test whether a system is “fit for purpose.”

Both terms are important to ascertain a certain level of flexibility in the development process. And they open productive ways to turn ideas into features from release to release.

Relevant finding for the SYNC3 test and validation plan:

The goal of creating a system that is “good enough” and “fit for purpose” is important for user evaluation as well. These concepts are actually a warning not to use testing to search for features that could be included. This would create the risk that the project becomes open-ended. In order to avoid this, a “wish list” will be one of the tools to collect possible features for future releases of SYNC3 without compromising the ongoing development.

2.4.4. Test early and often

The CHAOS 10 identifies “user involvement” as a key success factor. It does not define how it is supposed to be initiated, planned and managed. What can be done to really uncover user requirements while they shift or change, particularly in the field of annotation of multimedia?

The point is that the only way to get to a level of certainty is to involve users in different stages of development and to use a set of tools to ensure that the needs of these users are understood in the best possible way. This still is an iterative process. Results from interviews, use cases and use models can result in surprises and new directions, which can shift the demands regarding development. The less momentous these new insights are, the easier it is to finish a project in time and on budget.

Validation and verification therefore must be applied in each stage of the development process. Simply testing an end product in a late stage will statistically often result in major rework demands.

Relevant finding for the SYNC3 test and validation plan:

Constant user involvement is the key, testing early and often can create additional benefits. The test plan should answer the question how many tests are to be planned and how often and at what stage of development tests can be done.

2.4.5. How many tests?

How many users must take part in tests in order to create a system that is “fit for purpose”?

Usability expert Steve Krug^{ix} presents compelling data and findings that testing does not need to be overly complex in order to get results which are helpful for the development process and the programmers.

His recommendations plus suggestions from other authors such as Jacob Nielsen are helpful as they provide background information on how the test process can be set up and how many users must be tested to get trustable results. Additionally, simplifying testing as much as possible helps to reach the goal of testing “early and often”, which – as discussed above – is pretty important to ensure project success. Krug’s book, however, is about usability, not validation. Although validation goes further and is more formal than usability testing there are still many relations relevant to both disciplines.

A distinction to be made is that validation judges interface design based on formal requirements as well as from a functional perspective, not from a marketing point of view, where, for instance, an optimum of usability is highly important for eCommerce. Nonetheless, the recommendations from the usability experts are highly useful in this context.

The following excerpts from *Don’t make me think* provide relevant information on which aspects need to be considered:

The importance of recruiting representative users is overrated

It’s good to do testing with people who are like the people who will use the site (or software), but it’s much more important to test early and often. The motto is: “Recruit loosely and grade on a curve”. The point of testing is not to prove or disprove something. Instead it’s to inform the judgement of everyone involved with development. Quote: “People like to think that they can use testing to prove whether navigation system A is better than navigation system B but you can’t. No one has the resources to set up a controlled experiment you’d need. What testing can do is provide you with invaluable input, which, taken together with your experience, professional judgement, and common sense, will make it easier to choose wisely - and with greater confidence - between A and B.”

Testing is an iterative process

Testing isn’t something you can do once. You make something, test it, fix it and test it again.

Nothing beats a live audience reaction

Testing slightly different versions of a software results in gaining even more insight on what might be the best solution to a given task.

Simple test set-up

In the beginning, usability testing was a very expensive proposition. You had to have a usability lab with an observation room behind a one-way mirror, and at least two video cameras so you could record the users’ reactions and the thing they were using. You had to recruit a lot of people so you could get results that were statistically significant. It was science. It cost \$20,000 to \$50,000 a shot. It didn’t happen very often. But in 1989, Jakob Nielsen wrote a paper titled “Usability Engineering at a Discount”^x and pointed out that it

didn't have to be that way. You didn't need a usability lab, and you could achieve the same results with much fewer users.

Table 4: Traditional vs. Simple testing

	Traditional testing	Simple testing
Number of users per test	Usually eight or more to justify set-up costs	Three to four
Recruiting effort	Select carefully to match target audience	Grab some people. Almost anybody who uses the Web (or a computer) will do
Where to test	A usability lab, with an observation room and a one-way mirror.	Any office or conference room
Who does the testing	An experienced usability (or validation) professional	Any reasonably patient human being
Advance planning	Tests have to be scheduled weeks in advance to reserve the usability lab and allow time for recruiting	Tests can be done almost any time, with little advance scheduling
Preparation	Draft, discuss, and revise a test protocol	Decide what you're going to show
What/When do you test?	Unless there is a huge budget, put all your eggs in one basket and test once when the site is nearly complete	Run small tests continually throughout the development process
Cost	€5.000 to €15.000 (or more)	About €300
What happens afterwards	A 20-page written report appears a week later, then the development team meets to decide what changes to make	Each observer writes one page of notes for the day of test. The development team can debrief the same day.

Source: Steve Krug^{xi}

Number of users

Steve Krug suggests the ideal number to be three or four users for each round of testing. The first three users are very likely to encounter the most significant problems.^{xii} Statistically there is proof that 3-4 users will discover 85 per cent of a software or site's usability problems, and there are diminishing returns for testing additional users.^{xiii}

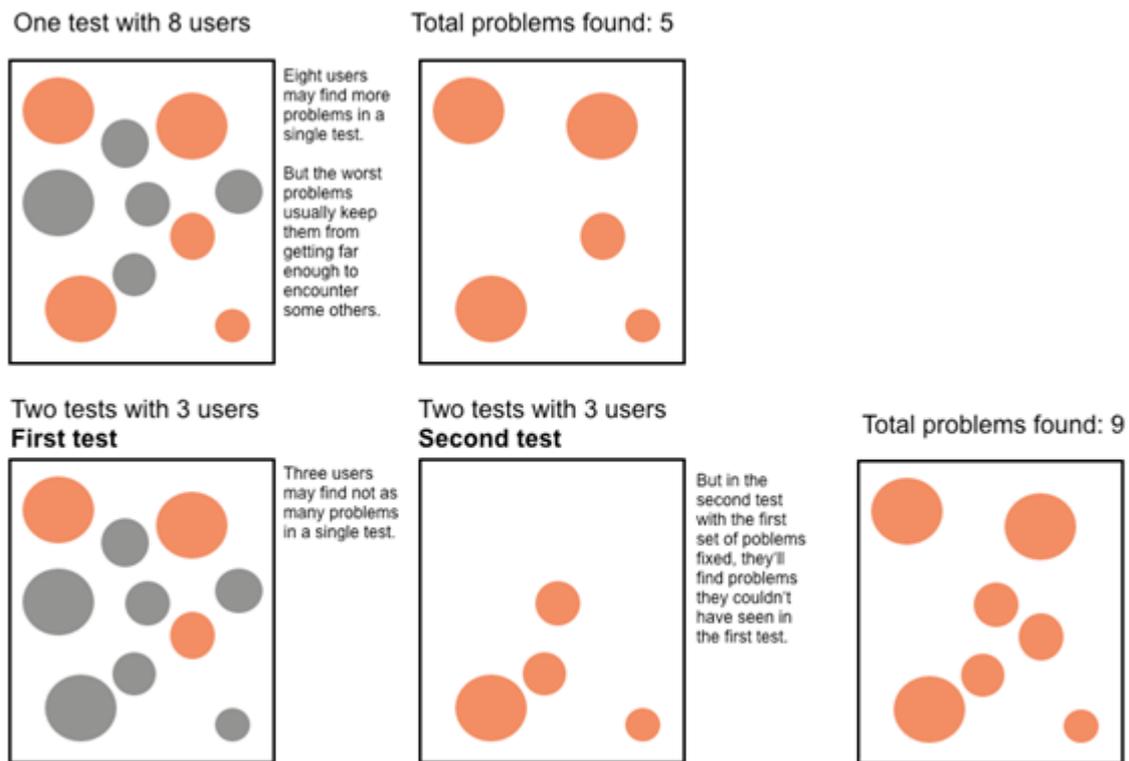


Figure 3: 3 users will discover 85% of an application’s usability problems

Many discussions have taken place about the sample sizes for usability evaluation since the studies of Steve Krug and Jacob Nielsen. The most popular rule in the usability evaluation field at the present moment is the “4±1” or “magic number five” rule.^{xiv} Nielsen and Molich studied the issue of sample size for usability testing in the case of the heuristic evaluation method (HE), in which a small number of evaluators inspect the user interface design of a system in order to detect usability problems in comparison with known usability principles (heuristics). By means of empirical studies and mathematical methods the authors demonstrated that 2/3 of usability problems are detected by five evaluators and that tests with additional evaluators are not likely to expose new problems.^{xv} Similar findings were indicated by Virzi in the case of determining the optimal sample size for usability testing using the TA (think aloud) method. He demonstrates that 80% of usability problems are detected by four or five users.^{xvi} These findings have been challenged by several subsequent studies.^{xvii} A meta-analysis of sample size issues in usability evaluation conducted in 102 usability evaluation experiments extracted from online academic databases, including ACM Digital Library, IEEE Xplore, and ScienceDirect, and offline sources since 1990 indicated that in order to obtain reliable/optimal results using the TA method the size of the user sample must be nine in order to detect 80% of the usability problems. In spite of this, the “4±1” or “magic number five” rule maintains itself as the most frequently invoked rule in determining sample sizes in usability evaluation.

Relevant finding for the SYNC3 test and validation plan:

Wherever possible and appropriate we will apply the recommendations provided above to simplify the test process. The main reason is that this will result in more testing

within a shorter time period and test results that can be used quickly by all developers. A complicated test process with high user numbers and a formal test report would also not fit the dynamic nature of this particular, journalists- and news-driven project.

2.4.6. Identifying future problems

The SYNC3 project aims to combine innovative modules to deliver the final product. Therefore the issue of system integration will gain importance from step to step. Starting with the first prototype, the task of integrating all the systems components must be considered in every step forward. Otherwise, testing and validation might not be able to identify problems that can arise once the modules have to work with each other to provide functionality.

Traditional “V-models”^{xviii} of software development assume that testing can be done as build-up process: Once a phase is finished, every module that passed the tests is supposed to work on the next level as well. However, this was frequently not the case in real environments.

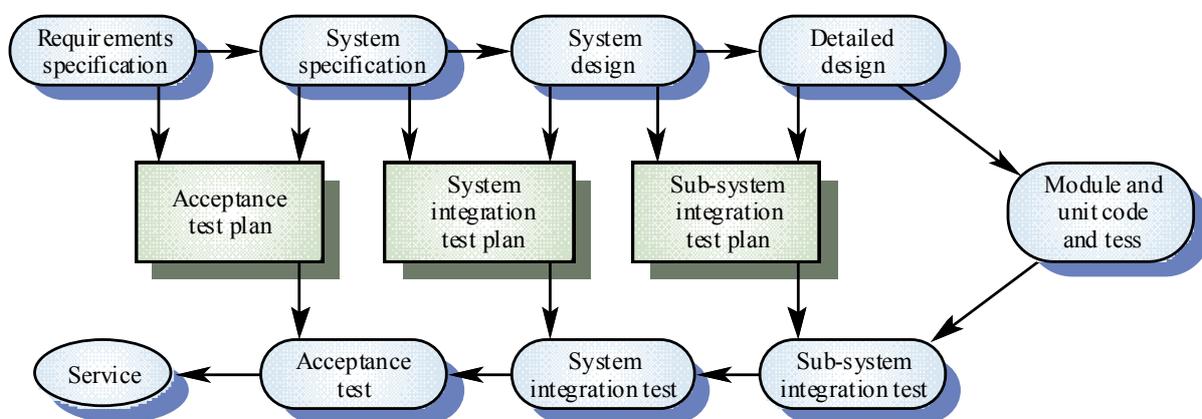


Figure 4: The V-model of development^{xix}

Critics of the V-model point out that the testing is often done too late and not often enough in the process. The main reason is the assumption that at the end of each step fully functional modules will be available. But once the integration starts, problems or bugs surface that lead to rework. Such late changes in a project are often both complex and time consuming. And they are difficult and harder to solve because subsystems must be changed and reconfigured as well.

“The V model fails because it divides system development into phases with firm boundaries between them. It discourages people from carrying testing information across those boundaries. Some tests are executed earlier than makes economic sense. Others are executed later than makes sense. Moreover, it discourages you from combining information from different levels of system description. For example, organizations sometimes develop a fixation on ‘signing off’ on test designs. The specification leads to the system test design. That’s reviewed and signed off. From that point on, it’s done. It’s not revised unless the specification is. If information relevant to those tests is uncovered later – if, for example, the architectural design reveals that some tests are redundant – well, that’s too bad. Or, if the detailed design reveals an internal boundary that could easily be incorporated into existing system tests, that’s tough: separate unit tests need to be written.”^{xx}

The suggested solution is to utilize an integrated evaluation approach, thus overcoming barriers between development processes as well as different modules. The earlier insights are gained on whether integration, functions and sub-systems successfully work together, the easier it will be to achieve good results. Even market research and new competition should be partly included in order to advance a project to the best level achievable.

3. SYNC3 user evaluation methodology

The SYNC3 user evaluation plan presented here is a systematic approach to ensure that the requirements of SYNC3 are met in the final product and that the modules released will have been validated.

Based on goals and best practices identified to be useful, this plan and report are designed to become a tool for the next phases of this project. Key points are a good understanding of the requirements as well as a detailed workflow. The outline used in this part is based on the IEEE 829 format.

3.1. SYNC3 objectives to be evaluated

As per the DoW, the goal of SYNC3

is to create a framework for **structuring, rendering more accessible and enabling collaborative creation** of the extensive user-provided content that is located in personal blogs and refers to running news issues. The main obstacle for blogosphere content exploitation has been the **difficulty to structure** the disparate and vast information found in blog posts.

Structuring that portion of the blogosphere that discusses the running news is the main objective of SYNC3. To achieve this goal SYNC3 follows the logical but unexplored path of deriving the thematology from the news articles and then applying it to the blogosphere domain. This method is based on the fact that news articles and comments on blog posts that refer to news articles share the same thematology. The novelty of SYNC3 lies in using the news event as the fundamental element around which the thematology is formulated. The news event in SYNC3 is defined as “something that takes place at a particular time and at a particular location”. To identify news events, SYNC3 is using an innovative algorithm combining statistical and language processing.

The SYNC3 system aims to also enable the creation of more user generated content either by authoring new material or re-organising the links structured by SYNC3 into user generated storylines.

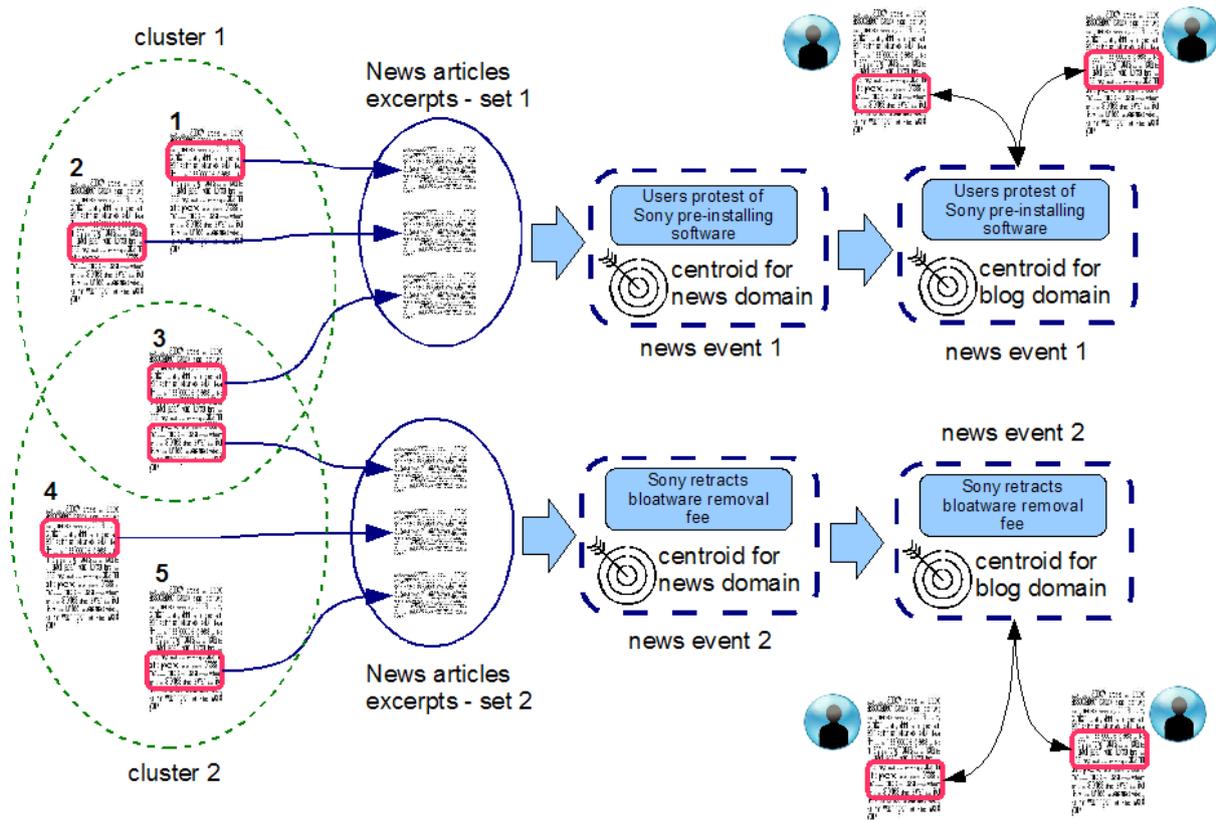


Figure 5: SYNC3 approach

Evaluation-related deliverables

This is the list of deliverables related to validation tasks in the SYNC3 project:

Table 5: Evaluation-related deliverables

Deliverable	Title	Responsible
D7.4.1	User evaluation report	EJC
D7.4.2	User evaluation report	EJC
D7.4.3	User evaluation report	EJC

Scope of testing

The entire SYNC3 user feedback activities are designed to enable a broad range of validation procedures, including tests of basic material such as layouts, rough sketches of user interfaces or functional modules, simulations as well as the prototypes of the complete system. This flexibility is needed to respond to the needs of the developers. The perspective is that any interaction with users can produce new insights how to build the final system.

However, in terms of validation the tests naturally become more meaningful as the system completion and especially integration progresses. Otherwise the risk of testing incomplete material might result in not detecting defects.

As discussed above, the validation aims to ensure that “the right product will be built” from a user perspective. In case that modules do not perform the desired tasks in a test situation for of technical reasons, the test result will be labelled as “failed”, and a note will be sent to developers asking for verification of the code.

3.2. Prototypes and evaluation scenarios

What will be evaluated in particular? This section provides a description of the suggested test and validation scenarios, based on the first prototype of SYNC3.

The 1st prototype, the delivery of which was originally planned for month 18 (September 2010) has been released in month 20 (November 2010). As planned, the 1st prototype

Provide[s] limited functionality compared to the next versions. Its aim is to prove the feasibility of the approach and pave the way for the next version.^{xxi}

The 2nd intermediate prototype is planned to be delivered in month 25 (April 2011):

All the main functionalities will be implemented. The performance of the overall system will be evaluated and changes towards the final prototype will be decided. Specific issues with each functionality will be tracked and scheduled for correction in the final prototype.^{xxii}

The 3rd prototype (final version) is planned to be delivered in month 31 (October 2011):

This will be a fully developed prototype with all the functionalities implemented and optimized as well as with integration of changes and additions from the evaluation of the intermediate version. Extensive technical and user evaluation will be performed after the delivery of the final version of the prototype.^{xxiii}

Obviously, the three main rounds of user evaluation need to take the different stages of completion into account. The 1st prototype is going to be checked primarily for basic functionality and general user understanding of the SYNC3 practical purpose and usability. In the second and third rounds, testing must dig as comprehensively as possible into the functional requirements and validate them with the different targeted user groups.

The following use cases were implemented in the 1st prototype and served as basis for formulating the tasks by means of which the test participants evaluated the SYNC3 prototype.

Table 6: Correlation of SYNC3 use cases with the first integrated prototype^{xxiv}

ID	Use Case Title	Supported in the 1 st Prototype
UC1	Perform keyword search	<input checked="" type="checkbox"/>
UC2	Perform advanced search	<input checked="" type="checkbox"/>
UC3	View search results	<input checked="" type="checkbox"/>
UC4	View today's news and blogs	<input type="checkbox"/>
UC5	Analyse event	<input checked="" type="checkbox"/>
UC6	Analyse blog	<input checked="" type="checkbox"/>
UC7	Analyse blog (advanced)	<input type="checkbox"/>
UC8	Monitor blogs	<input type="checkbox"/>
UC9	Create event relations	<input checked="" type="checkbox"/>
UC10	Associate events with blogs and news	<input checked="" type="checkbox"/>
UC11	Create groups	<input type="checkbox"/>
UC12	Export news stories	<input type="checkbox"/>
UC13	Manage workspace	<input type="checkbox"/>
UC14	Receive newsletter	<input type="checkbox"/>

3.3. User evaluation methodology

In the course of SYNC3 user evaluation, we employ tried-and-tested techniques to elicit useful feedback from the test persons. To this end, and different from the focus group brainstorming sessions during the user requirements analysis¹, user evaluation will usually be conducted in face-to-face sessions with a maximum of four persons present:

- The test leaders (from EJC), asking questions, giving instructions and assignments, debriefing, observing and taking minutes;
- The primary test user, performing the evaluation and subsequently teaching the secondary test user;
- The secondary test user, to be instructed about the system by the primary user.

Where two test leaders are available, one of them could stay in the room with the test user currently carrying out the evaluation while the other one asks the concluding questions to a previous test user (debriefing). Under time pressure, this will achieve the best possible time economy, while under more relaxed circumstances, one test leader is sufficient.

We use three main techniques to collect user feedback during practical testing, all complemented with audio documentation for backup and notes made by the test leader:

1. Thinking aloud and observation

This technique means that the test users will be given assignments they have to perform with SYNC3. The test leader encourages the users to permanently talk about his/her impressions and actions during the evaluation process. In such a way, the mental models by which users address a task or try to achieve a goal can be detected and analysed. All the while, the test leader observes carefully the subjects' behaviour in order to try and detect even semi-conscious interactions with the system or barriers which are not expressly addressed by the user.

The benefit of this approach is the fact that user behaviour and user satisfaction become immediately transparent. The need for modifications – if any – will become apparent, as will the possible need for specific training or introduction to the SYNC3 tool. At the same time, the professional users will express to what extent SYNC3 actually caters to their everyday work requirements.

2. Constructive interaction (teaching back)

This technique consists of two stages. In the first step, one test user gets the opportunity to try out and become familiar with SYNC3. In the second step, the same user explains the functionality of the system to the next user in line. The success rate of this direct user-to-user training is directly related to the mutual understanding of the system.

The particular benefit of this approach is that the first user is required to expressly verbalise his/her comprehension of how the system is working and how it is intended to be used. This task will therefore trigger a reflection process and prompt the first user to explain SYNC3 in a systematic fashion. This will reveal how deep the actual understanding has become at this point and highlight features which remain unclear or hard to grasp. In case this “Chinese whispers” test works well, the system has a very clear and easy usability; if not, the misapprehensions will highlight urgent action points.

However, this approach can only be used in situations when test users have some extra time on their hands and are not anxious to rush back to their regular tasks. It is also of limited use with the user interface simulation and the first prototype, as user interaction with the system requires concise stewardship through the available processes which users cannot know at this stage.

3. Collection of express feedback

Immediately after finishing their hands-on experience with SYNC3, the test users will be asked for their personal evaluation of the system. They will be asked to fill in a standardised questionnaire and will also be given the opportunity to independently express their opinion and possible suggestions.

The benefits of this technique are obvious, since it allows the collection of conscious cognitive reactions and recommendations. While such information alone, without the abovementioned first two steps would run the risk of misrepresenting the user experience – e.g., since people tend to rationalise or to respond according to pre-existing prejudices – in this case it will constitute a useful supplement to the observations made during the practical work with SYNC3.

However, all user evaluations must take into account that users frequently tend to react adversely and insecure to new, unaccustomed software. This is particularly true for those professional users who have long-term experience with other software solutions in the particular field of SYNC3. However, since the particular field of SYNC3 is rather recent, this reaction should not occur. Furthermore, the phenomenon that assessments made in surveys frequently turn out more negative than is warranted by the actual subject of enquiry will most probably come up during the SYNC3 user evaluation as well. The situation of being asked for opinion statements as such often leads to a particularly critical point of view.

3.4. Survey of exploitation opportunities

Having actual potential users try the SYNC3 system offers one of the best conceivable opportunities to find out about the practical application prospects in the media business sector for the eventual product. While participants of the user evaluation efforts will be primarily everyday users and not high-ranking decision makers, it is the assessment of exactly this kind of users that will influence the evaluation process in media companies to a great extent.

If the test users realize the gains in time and convenience and do not show adverse reactions to the SYNC3 system, they will become SYNC3 “ambassadors” in their respective organisations. Under the assumption that the participants in the user evaluation are representative of their peers in similar parts of the sector, their statements can also be used as a supporting measure in the eventual SYNC3 sales development.

In a separate effort, the final prototype will be presented to a number of executives at management level, who will be asked to assess the tool from the point of view of integration into existing work flows in their respective organisations and of the perceived balance between potential investment and efficiency gains.

4. SYNC3 user evaluation by use case

It should be noted that the following use cases were designed with the SYNC3 end product in mind. Quite obviously, they have different priority levels and pertain to very different situations and environments. Some will only be required by the time the system is released into commercial usage, while others speak to the core qualities of the envisioned tool.

Table 7: Summary of SYNC3 use cases and the status of their implementation in the 1st prototype^{xxv}

ID	User Group	Use Case Title	Use Case Description	
UC1	All	Perform keyword search	All users may perform queries to the system by using a free text search engine.	<input checked="" type="checkbox"/>
UC2	All	Perform advanced search	All users may perform queries using facets provided by the system like author name, source, time, etc.	<input checked="" type="checkbox"/>
UC3	All	View search results	All users may navigate within the search results, which may be presented in different ways, according to the needs of the user.	<input checked="" type="checkbox"/>
UC4	All	View today's news and blogs	All users will be presented with an overview of the most important news articles and blog posts.	<input type="checkbox"/>
UC5	All	Analyse event	Every user may view all metadata information and analysis of an event. This includes news sources, relevant blogs, sentiment analysis (for blog posts), and temporal/causal/location relations (among events).	<input checked="" type="checkbox"/>
UC6	All	Analyse blog	All related events and corresponding sentiment analysis of the blog are identified by the system.	<input checked="" type="checkbox"/> (partly)
UC7	All	Analyse blog (advanced)	The blogger's profile is displayed and a ranking system is available to the users.	<input type="checkbox"/>
UC8	Professional journalist	Monitor blogs	Particular blogs of interest are monitored for their sentiment and relations with traditional news sources regarding an event.	<input type="checkbox"/>

ID	User Group	Use Case Title	Use Case Description	
UC9	All	Create event relations	All users may use a personal space to create new relations between events of their interest and save this information.	<input checked="" type="checkbox"/>
UC10	All	Associate events with blogs and news	The users will suggest blogs or news items associated with an event, which were not identified by the system or were linked in a different way.	<input checked="" type="checkbox"/>
UC11	All	Create groups	All users may create groups which will share personally created news stories.	<input type="checkbox"/>
UC12	All	Export news stories	User created news stories may be exported to other social community platforms.	<input type="checkbox"/>
UC13	All	Manage workspace	The SYNC3 users define the set of news portals and blog sources from which they collect content	<input type="checkbox"/>
UC14	All	Receive newsletter	All users are offered a newsletter of events of their interest	<input type="checkbox"/>

A series of tasks to be performed by the test users during evaluation sessions has been produced to enable evaluation by means of the think aloud protocol. The tasks have been formulated to cover the use cases described in the table above, which correspond to the functions implemented in the 1st prototype. According to D8.2: First Integrated Prototype, which documents the first integrated prototype, the implemented functions are: news processing and analysis which enables the identification and analysis of news events, blog processing and analysis which enables the association of blog posts with news events, news events labelling and relation extraction which enables the creation of relations between news events, user interface which provides the environment through which the users interact with the system, search for news events and visualizes the generated results, and integration, consisting of integration of all these other functions.^{xxvi}

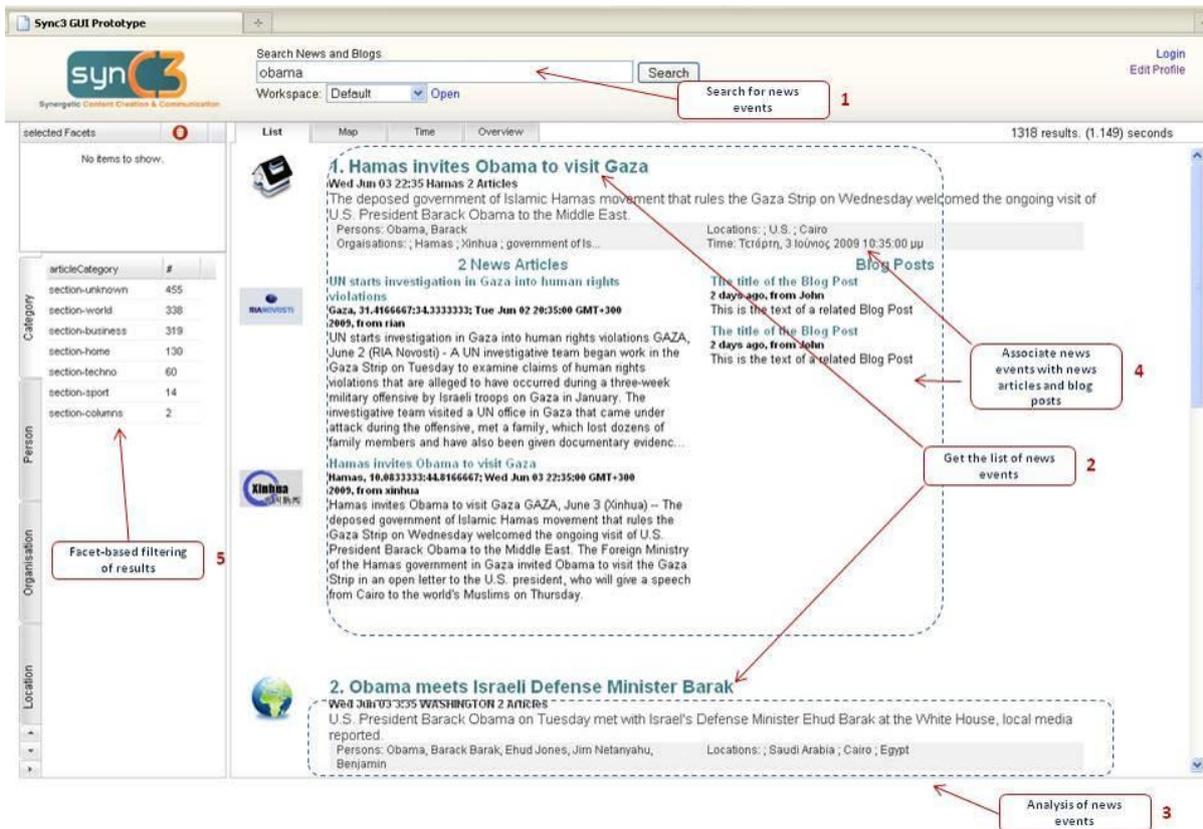


Figure 6: Visual representation of SYNC3 use cases related to the search function^{xxvii}

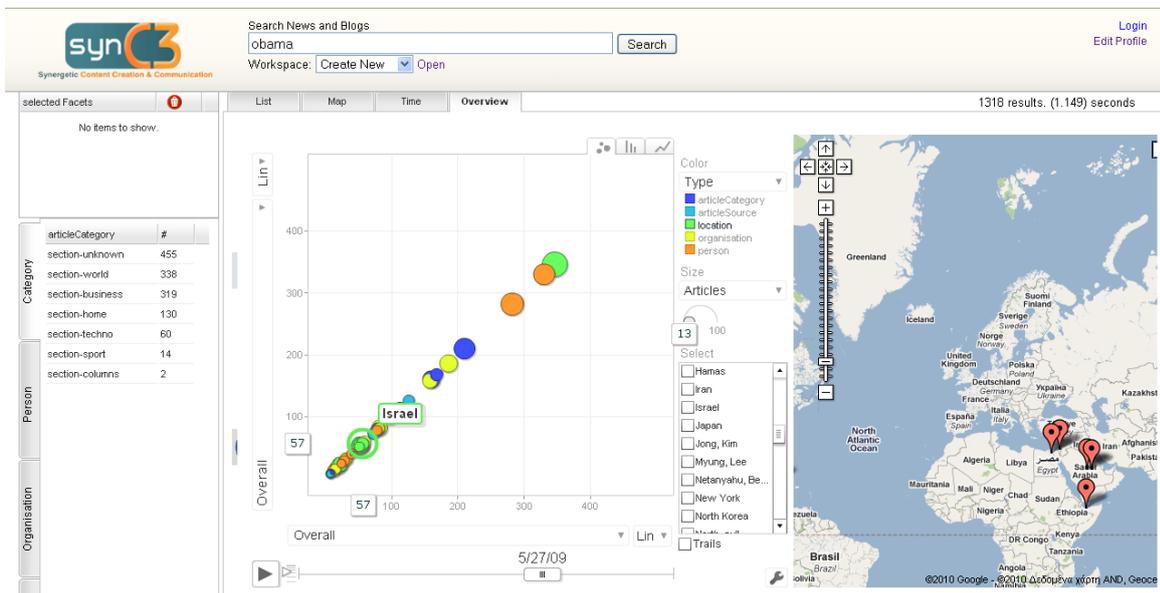


Figure 7: Visual representation of SYNC3 use cases related to search results visualization^{xxviii}

The following tasks have been formulated to evaluate the use cases that were implemented in the 1st SYNC3 prototype. Each task is accompanied by questions which the test leader can ask in order to stimulate the participant to make observations on aspects of the SYNC3 system which require user validation:

- [UC1 + UC5] Search for news about the BP oil spill by using the SYNC3 search function. Explore the generated news events. Are the results satisfactorily matching the query? Are the event titles meaningful/clear or not? Do you find it useful or not to have persons, organizations, locations involved in the event extracted and displayed?
- [UC3 + UC6] Select one of the generated news events and explore the news articles and blog posts that discuss it. How would you comment on the listed news sources? Are the listed blog posts relevant to the selected news event? Are you satisfied with the range of news sources listed? Are there sources that you would have expected to be listed and have not been?
- [UC10] Type the same query: “BP oil spill” in the Google News search engine and limit the results to the period January-September 2010 by using the “Custom Range” function. Look at the listed results and their sources. How would you compare them with the results you get from the same query in the SYNC3 prototype? Are there any sources that report on this topic that you consider essential for the SYNC3 prototype to list?
- [UC2] By using the SYNC3 advanced search function, filter the generated results by one or more of the following: ‘person’, ‘organization’, ‘location’, or ‘source’. Do you find this filtering function useful in your work? Why? Why not? Are the SYNC3 search capabilities satisfactory? What other kinds of search capabilities would be useful to have? Would you be interested for example to be able to search by keywords + positive/negative sentiment?
- [UC3] Explore the generated results with the Overview having the X axis set to ‘time’ and the ‘y’ axis set to ‘article’ to see how the number of articles corresponding to your query changes over time. Select an entity whose evolution to follow over time from the following categories: news source, person, organization or location. Set the Y axis to ‘overall’ to see the overall number of articles referring to a particular entity from the categories: news source, person, organization, location, until the date that you selected. Does the graph help improve your understanding of the generated results? Does it bring a different perspective on the generated news? Do you find such visualizations useful in your work or not? Which of the modes of visualizing results is the most meaningful to you? Which of the visualizations would be most useful in your work?
- [UC9] Select two or more news events from the search interface that you consider are related in some way and add them to the workspace.
- [UC9] Place them in a Group Canvas and name your Group Canvas. Do you find the function of grouping news items by means of the Group Canvas useful in your work?
- [UC9] Draw connections between the news events by using arrows and make notes on the relations between them. Pay attention to the type of relations the subject is creating. Do you want to freely add relations between news items or do you want to have standard options for making relations between news items? If so, what type of relations would like/need to establish between events?
- [UC6] Annotate one of the news events by using the rating scale. When blog posts will be transferable to the Workspace, sentiment can be annotated as well. What would you relate the sentiment to? Would you be interested more generally in the sentiment of the entire blog post or in sentiment towards certain mentioned entities?
- [UC9] Move some news items from the Group Canvas to the Group List. Are the offered possibilities to comment on the news items satisfactory? Do they enable the type of analysis that you usually perform on media content? What type of activity would you want to perform with the workspace?

Additionally, and depending on available time, some questions were formulated to gather user feedback on UC11-14, based on a User Profiling Infrastructure simulation provided by L3S and available at: <http://sync3.atc.gr/de.l3s.sync3.gui/groups/>.

- Would it be useful in your work to have the functionality of creating groups and joining groups?
- How would you like to get recommendations: based on keywords that you can fill into your profile or based on your queries?

At the request of the consortium partners the following questions were included in the evaluation sessions' script.

- Explain the participant that in the case of related news events a news article can be part of two news event clusters. What would you prefer: to have the same article listed twice in two news events or to have it just once in the most recent event? I.e.: To miss information or to have redundant information? Also, some articles listed under a news event are "near" duplicates in the sense that the only thing that distinguishes them is the URL. Would you be interested in having them all displayed or just one of them?

Each session is concluded by answering any remaining questions of the test participant and asking for an overall impression.

Integration issues (from the user perspective):

Do users perceive any flaws in the interplay between SYNC3 components?

How much time is consumed, and do participants feel that they increase efficiency with SYNC3?

Inherent Risks:

Overwhelming the user by an abundance of settings or difficult to understand options; scarcity of options. During actual evaluations, cutting the usage scenarios into smaller pieces so as to better accommodate the different prototypes might turn out useful, especially since the first two prototypes will most likely not cover the entire functionality of the SYNC3 final product.

5. The user evaluation process

The user evaluation looks at the integrated system only, not at individual modules. Therefore, the tests are best organised not by module or by function, but according to the use cases as described and listed above.

5.1. Recruiting user evaluators

EJC as a task leader is responsible for the recruitment of user evaluators. RIAN is responsible for the recruitment of user evaluators for the evaluation at RIAN headquarters in Moscow. The focus is on users coming from the target groups as described in Section 1.2. In order to receive a valid and balanced feedback from users pertinent to the SYNC3 domain, yet with different work specialisations as well as cultural backgrounds, the following test groups and test locations are chosen:

(1) Bucharest, Romania

Location: Bortun-Olteanu PR Agency

Date: 19 January 2011

Target group: communication professionals, PR professionals.

Recruiting method: EJC affiliates

(2) Moscow, Russia

Location: RIA Novosti headquarters (SYNC3 consortium member).

Date: 20-21 January 2011

Target group: News agency journalists and editors.

Recruiting method: Internally at RIAN (staff members and freelancers); selected participants of the original user requirements focus groups may be invited as well.

(3) Maastricht, the Netherlands

Location: EJC head office

Date: 25 November 2010

Target group: journalists, editors, bloggers

Recruiting method: Internally at EJC and EJC networks in the Netherlands and Belgium.

(4) London, UK

Location: Report International

Date: 16 December 2010

Target group: media monitoring specialists

Recruiting method: EJC affiliates

5.2. Expected number of tests

The expected total number of user evaluations depends on how well the integrated prototypes will pass the tests, because in the case of system failure or crashes, tests will need to be repeated in order to validate any fixes or changes. Also, the developing partners may ask for intermediate user evaluation to support their work before or between deliveries of the two set prototypes.

The expected number of users broadly ranges between:

- 20-25 individual test users for the 1st prototype
- 12-15 individual test users for the 2nd prototype
- 30-40 individual test users for the final prototype

A smaller number of participants suffices for the second cycle of user evaluations as the intermediate prototype will present relatively small developments. To enhance consistency of the user evaluations and to ensure that the progress between the three prototypes is better measurable, the test users from the first round will be asked to take part in the second evaluation as well. For the user evaluation of the final prototype, more participants will however be added since the full functionality requires more intensive testing. Furthermore, the SYNC3 product will be closer to marketability and the consortium should benefit from a broader base of user assessments to be integrated into the exploitation and business plan.

5.3. Qualification of the user evaluators

In order to obtain meaningful insights from the tests, basically any traditional and new media consumer or content creator would be qualified for a test of the SYNC3 system. However, to support the exploitation prospects of the SYNC3 system, it is business users who are able to validate the user requirements to the full extent and assess the eventual usefulness of the system. Moreover, if demanding professional users develop an aversion or resistance to the usability of the SYNC3 user interface or dismiss the eventual annotation results, this must be detected at the earliest possible stage so as to not jeopardise the system's exploitation prospects.

Experts in the field of testing advise to be wary of users who claim high levels of technical skills and programmers that claim to fully understand the business process. These types of individuals can cause more harm than good if they do not have the skills they believe they possess or may come with pre-defined concepts which SYNC3 is then likely to frustrate.

All test persons are invited cordially to participate. There is no particular incentive to take part in the tests other than a small present of minor value for external subjects. A single test is estimated to take between 30 minutes and a maximum of one hour.

Generally, any advance discussion of the system or its content should be avoided so as to guarantee a fresh and open approach of the evaluators. That however does not preclude a general introduction to SYNC3 and its user interface, or educating participants beforehand to the specific limitations of the prototype at hand so as not to frustrate their expectations for no reason.

5.4. Test environment and technical prerequisites

The user evaluations require any conference room or secluded office space, equipped with a computer with loudspeakers, a browser and a broadband Internet connection, as the system runs live on the technical partners' servers. With permission of the subjects, audio recordings are made of the testing sessions. The introduction and the questions asked follow a fixed script. This ensures that the test environment is basically the same for different users.

5.5. Managing user evaluation risks

The insights and information gathered through testing can be helpful to identify risks and contingencies that relate to the SYNC3 project. Risks include the delivery dates of modules and integrated prototypes, new and reworked items, and the ability of users to understand the tools provided by the software.

Overall risks to the project with an emphasis on the testing process are:

- Lack of qualified test participants;

- Lack of availability of required hardware, software, test data or tools;
- Late delivery of the software, hardware or tools;
- Delays in training on the application;
- Changes to the original requirements or designs;
- Complexities involved in testing the applications.

In more general terms, there are three more sets of risks relevant to SYNC3:

Inherent risks

There are some risk issues due to complexity and innovative functions that must be monitored as a part of testing. Examples are safety (e.g. user data, content alteration), multiple devices used, settings of preferences, server load, and software clients, but also the fact that users may not be acquainted with and thus irritated by novel functionality.

Requirements risks

Vague or unclear requirements or false assumptions result in additional risks. Based on testing insights these must be documented and communicated in order to decide whether they pose a threat to project success or can be shifted to later lifecycle versions of SYNC3.

6. Intermediate conclusion

This validation plan aims to assist the SYNC3 system development with a flexible as well as comprehensive and effective means to steer further development. The detailed set-up and the definition of criteria are designed to make sure that objectives and goals set upon inception of the project will be met.

The user evaluation plan and methodology is however a dynamic document, based on feedback from consortium partners and project needs. The number of tests, the scope and questions asked will thus differ from phase to phase. The basic process though will proceed as described. Therefore, close monitoring of the entire validation process will be one ongoing task for EJC.

Testing and validation can be a source of many enhancements as well as a means to control risks inherent with every software project. Based on results from the first test cycle, appropriate alterations of the plan must be developed for the follow-up phases such as bug fixing, fine-tuning and optimising the final product.

7. Results of the first user interface focus group

The focus group was organized at the EJC Maastricht office on Tuesday 30th March 2010. As it was a first focus group on the user interface, it was agreed that only close EJC contacts will participate (EJC employees and freelancers).

The focus group was moderated by the EJC (Mirek Hazer), while the user interface presentation and discussion was delivered by the L3S (Christian Kohlschutter and Dimitris Skoutas).

7.1. Set-up

7.1.1. Agenda

In order to allow quality discussion between participants based in Maastricht and L3S in Hannover, Skype videoconferencing and the screen sharing feature connected to the EJC's conference room were used. The focus group agenda was as follows:

1. Purpose of the focus group (EJC)
2. Overview of the SYNC3 project (EJC)
3. User interface presentation (L3S)
4. Feedback and discussion (L3S and EJC)
5. Wrap up (L3S and EJC)

The focus group was originally planned for one hour, but at the end it took almost 90 minutes because of good discussion and interest coming from the participants.

7.1.2. Participants

There was a set of 13 people associated with the EJC Maastricht office attending the focus group. It was of great benefit that these presented a real mix of media related potential users: professional journalists, bloggers and new media specialists coming from different background and countries, bringing in different expectations.

SYNC3 focus group organizers

- **EJC:** Miroslav Hazer (focus group moderator), Eric Karstens (WP7 Leader)
- **L3S:** Christian Kohlschutter (delivering UI presentation), Dimitris Skoutas (WP6 Leader) (both connected via Skype videoconference), Patrick Siehndel

EJC participants

1. Wilfried Ruetten – director
2. Ivan Picart – IT manager
3. Arne Grauls – web manager
4. Raymond Frenken – freelancer, professional business journalist
5. Sueli Brodin – freelancer, blogger, news editor
6. Rebecca Gehrler – project assistant
7. Liliana Bounegru – editor, new media specialist
8. Biba Klomp – project manager
9. Howard Hudson - editor
10. Adrianna Miara – associate editor, multimedia editor
11. Diana Lungu – associate editor

12. Daniel Paone – freelancer, new media specialist
13. Rina Tsubaki – project manager

7.2. Outcome of the Focus Group

The User Interface presentation was divided into three logical parts following an expected user scenario: 1) starting with search function, 2) going to the workspace and 3) adding the personalization and recommendation features.

There was a discussion followed by the presentation of each of the three parts. The key comments and recommendations made by the participants are included in the following sections.

7.2.1. Search Function

- How do you see if you're searching news events, news articles or blog posts? It should be visible which results come from news and which from the blogosphere. Suggestion made: divide the search results into more columns – separate search facets for news articles and blog posts.
- Does the information on the right side show the most referred to entities? Participants did not understand fully where the information on the right side comes from/how it is generated.
- How do you distinguish between news event sources and how is the order in which they are listed in the search results being decided upon? Are the most popular and read sources included as the first ones in the search result? Participants did not understand fully how the listed order is generated. “We need to understand why we see certain news event or news item as the first, or the second one in the search. It is also important to know where the results that we see are coming from and how they were generated.” *(It was explained that the order is based on the algorithm result.)*
- Where does the news event label and description come from? Is it taken from some of the news items or generated by the system? *(It was explained that both the news event label and description comes from the most representative news articles based on the statistical results.)*
- When a news item comes from a journalist at BBC, does it count as a news article or blog post? *(It was explained that in this case it will be listed among blog posts, because it is a news item bringing the opinion and sentiment of the author. E.g. <http://www.bbc.co.uk/blogs/>.)*
- Where do we see the sentiments/opinion? *(There will be a search facet for the sentiment analysis on the right side column, similar to the existing facets for geography, etc.)*
- **Observation:** The system is not unique. You could copy already some other existing systems and add only some features, such as sentiment analysis. **Suggestion:** the added value of the system is in answering the question: where and how do you find the sentiment? It would be of a great value to generate a cloud of keywords (together with sentiment analysis) that appeared in the news items (including comparison of keywords used in news articles vs. blog posts).
- There should be a possibility in the search field to go back to previous queries without having to delete added search terms.
- It should be possible to see a number of news items next to the news event label and description and see the list of news items divided into news articles and blog posts.
- The system should allow visualization of sentiment according to location.
- It is also important to know the location of the person who creates the workspace.

7.2.2. Workspace Function

- How do you determine sentiment? How do you decide what is negative and positive? You should also consider different levels of positive and negative sentiment. Sentiment analysis is the most interesting feature, but it needs to be explained to the users somewhere. *(SYNC3 provides sentiment analysis only on blog posts in relation to the news event. The sentiment analysis is based on lists of positive, neutral and negative words that are a standard in the sentiment analysis.)*
- How do you distinguish between a negative topic and the author's negative sentiment towards the news event? Which one does the system take into account? *(The authors' sentiment determines the sentiment analysis result.)*
- Can you imagine a situation when you will be using the workspace tool? (moderator question)
 - It can be useful for press officers working on media monitoring and analysis.
 - For journalists it is time consuming but if it's shared inside of a team then it's ok.
 - It can be useful in a newsroom, in case journalists want to see the difference in how the news event is covered in news articles and blog posts.
 - The tool could be of use to media research and academic groups who have more time for lengthy analyses and are interested in media content analysis and relations between types of media.
- Can the workspace graph be generated by the system itself? *(Not at this moment, this feature is not planned to be implemented.)*
- Is there a way to make more than arrows and arrange articles in the workspace, like make more lengthy notes and give titles? *(Yes. L3S presented all the possibilities which seemed to satisfy the participants.)*
- It would be useful if in the workspace you could have highlighted in the news items the elements which the system extracted in the search page (such as location, persons, organizations and maybe indicators of sentiment). This type of simple visualization of key elements in the news item according to the system would make analysis and connecting news items in the Workspace easier.

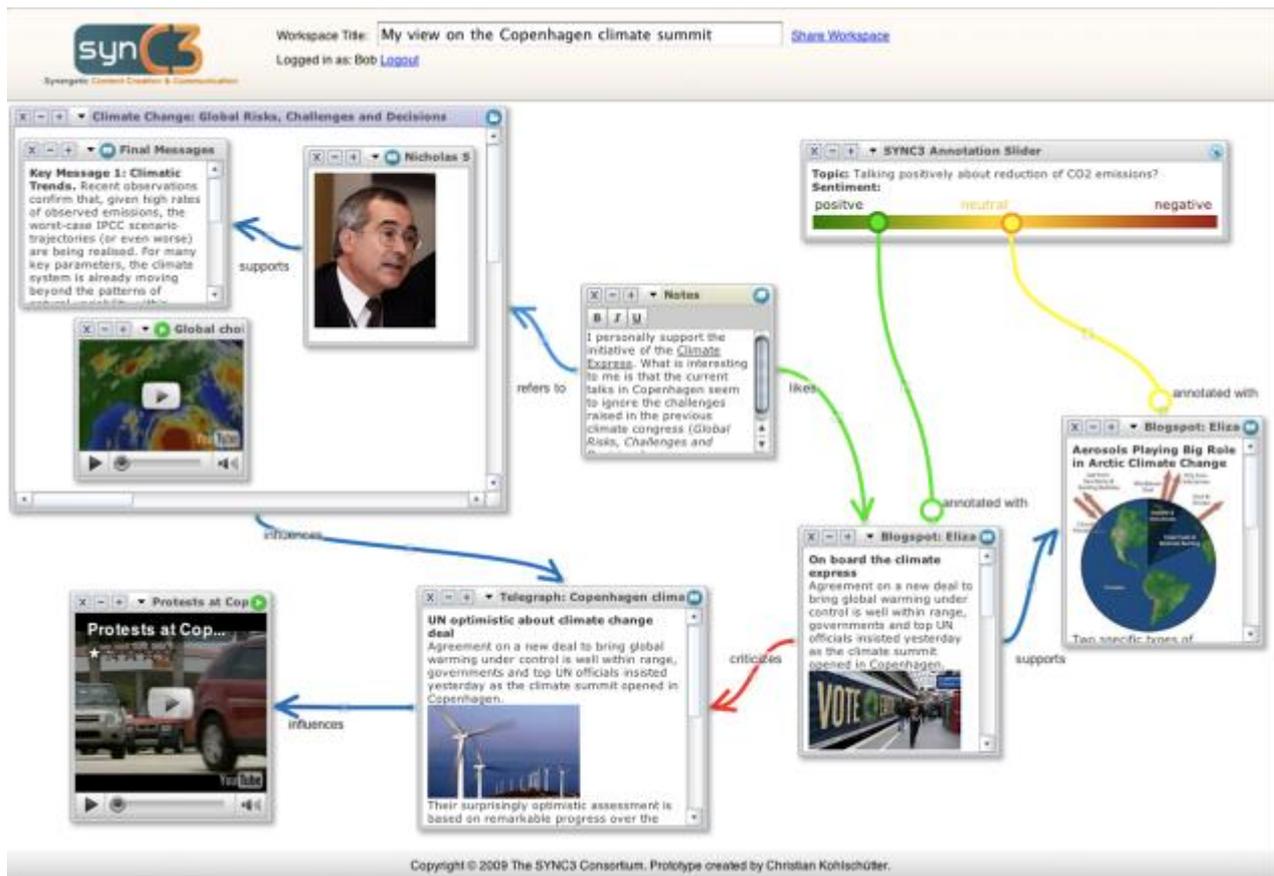


Figure 8: Screenshot of the SYNC3 Workspace

7.2.3. Workspace Personalization and Recommendations

- Would you like to use the workspace in closed or public opened groups? (asked by L3S)
 - It should allow both closed and public groups depending on the users' profile. Bloggers would be more into sharing it, but communication professionals would perhaps work more on company confidential analysis shared only within their team.
- How should the user receive recommendations? (asked by L3S)
 - The user should have to choose how often to receive recommendations – to have the possibility to individualize the periodicity and also determine keywords that users are interested in. Even the weekly updates as sent from LinkedIn groups are seen sometimes as a spam because there is too much of not relevant content in them.
- Comments
 - Is there a bookmarking system for both the workspace and the news events and items which are listed in search results?
 - There should be a word cloud in the workspace and also a widget for persons that are part of the news event.

7.2.4. General Comments

- Can the workspace be embedded in a blog post? People would like to take the workspace once created with them to use it for their blog posts, presentations, etc. *(Yes, it is possible, because each workspace has its own individual URL.)*

- Is it worth to provide the possibility to include comments on the whole workspace and how should be these comments displayed? (asked by L3S)
 - Is there a possibility to embed the workspace into a blog post or Facebook page?
- The language issue – using only English language is viewed as a restriction. *(The core lists of news items sources are picked up manually and they come from different geographies covering all continents.)*
- Maybe you could offer a premium version as a paid product to companies (media analysis and opinion mining service).
- What function lacks entirely from the system? (asked by L3S)
 - A follow up function: What do you do with the workspace? Is there a way to visualize workspaces on a certain topic? To search through them, etc? *(L3S reacted that this is a good suggestion.)*

7.3. Outcome from the individual testing

The participants of the focus group got 24 hours of access to the SYNC3 UI interface (under the non-disclosure agreement) hosted on the L3S server to be able to try it by themselves and provide more in-depth comments.

The following comments were gathered after the individual testing:

- The system works with AJAX requests on the background. This is a bad choice as:
 - searches are not bookmarkable,
 - the back button does not work (or you have to hack it to make it work).
- Ideally, the URL/SEARCH in the address-bar is user-friendly for power-users, like as in Delicious, for example:
 - <http://delicious.com/popular/coding>
 - <http://delicious.com/tag/earth>
- Ideally, every search result is immediately subscribable through RSS too (just like Delicious).
- “I don't like the "iframe" feeling of the scrollbars at all. Why not just adopt a fullpage layout?”
- The right column should have a title that better explains the function, such as "Further-Filter Results"
- The right-hand column items, could have a collapse/expand toggle function (especially when there will be more options available).
- When navigating a search result with 2300 items, the system loads new results while scrolling down. Most people are used to (Google) pagination and will prefer this over AJAX loading.
- Regarding workspaces, there is Yahoo Pipes (<http://pipes.yahoo.com>) which has some similar features. One element there that could be taken into consideration is the "union" item.

7.4. Conclusions of the focus group on user interface

These conclusions are viewed as a suggested list of TODOs, which the consortium (mainly L3S) will prioritize and try to address as much as possible, so in the future it is possible to revisit this document to see what actions have been taken and which issues have been addressed and how.

Based on the feedback from the participants, the following key fields should be further discussed and considered:

- Users do not understand fully how the search results are generated and what they stand for (how the system created the news event label and description, why is certain news event on the first place in the search result, etc.).
- To be really able to map and analyze the blogosphere, the users need to better understand and see what search results come from news articles and which ones come from blog posts (it was not clear enough that the news event is generated from news articles and sentiment analysis from blog posts – the potential users would need first quite detailed explanation of how the system works to be able to understand the displayed results).
- Users view sentiment analysis as one of the key features but are not fully familiar on how to really distinguish between positive and negative sentiment and what the sentiment really refers to.
- The workspace is viewed as a useful tool but it will be used differently by different audiences. Because it is seen as something which can be very time consuming, it needs to be really user friendly and fast to understand on what one can do with the user interface (some gallery of examples can be helpful). It is needed to also provide the possibility to search through already created workspaces and enhance the bookmarking and sharing possibilities.

The conclusions of the first focus group on the user interface were communicated to the consortium and particularly to L3S. An improved version of the user interface was tested during the first prototype user evaluations, which are the object of the next chapter.

8. Results of the 1st prototype user evaluations

The results of SYNC3 user evaluations will be consolidated into three main parts, one for the evaluation round of the first prototype, one for the evaluation round of the second prototype and another one summarizing results related to the final prototype. Additionally, supplementary user evaluations will be summed up as well, such as several informal ad-hoc test runs with the user interface.

This first user evaluation report presents the results of the first prototype evaluation resulting from the four rounds of evaluation which took place at the EJC headquarters in Maastricht, the Report International headquarters in London, the Bortun-Olteanu PR Agency headquarters in Bucharest, and RIA Novosti headquarters in Moscow. Two individual tests were done with an EJC staff member with journalistic background and a media researcher at the University of Amsterdam, the results of which are included below.

The following table provides an overview of user evaluation dates and locations related to the 1st prototype.

Table 8: Overview of user evaluation dates and locations

Location	Date(s)	Number of subjects
EJC headquarters, Maastricht	25 November 2010	5
Report International headquarters, London	16 December 2010	4
Bortun-Olteanu PR Agency headquarters, Bucharest	19 January 2011	4
RIA Novosti headquarters, Moscow	20-21 January 2011	8
Total	Four sessions	21

8.1. Evaluation framework

Demographics

A total of five news media professionals and media researchers took part in the user evaluations at the European Journalism Centre (EJC) in Maastricht on 25 November 2010. The diversity of the EJC environment was reflected by a variety of language proficiencies: French, Italian, Spanish, Moldovan and Dutch. All participants were fluent in English. In terms of computer languages and technologies, they all felt comfortable to use in their daily work Microsoft Office applications. Two of them were experienced in HTML and CMS coding. In terms of distribution of subjects, all participants were junior or mid-level staffers aged between 25 and 35, with one below 25. Three women and two men evaluated the system, all of them actively working in the news media field either as journalists, editors, project assistants or media researchers. Their daily jobs involve production, monitoring and analysis of news media content (including multimedia content) as well as following developments in media and journalism and organizing training, research and communication projects to cover of EU developments, foster media pluralism and media innovation, and improve journalism skills. All participants from the EJC had prior knowledge of SYNC3, as they had taken part in the earlier focus group on the user interface, worked on the project or received internal updates on the development of the project.

At the Report International headquarters in London (a media monitoring and analysis agency which tracks and collects media coverage, analyzes it qualitatively and quantitatively and reports on it to their clients, with offices in around 45 countries and covering 39 languages), a total of four senior professionals (the managing director, the client service director, the solutions strategy manager and an account director) participated in user evaluations on 16 December 2010. All were proficient in English. With several years of experience in media research, in their daily work they are involved in media content monitoring and analysis at macro-level, focusing on outlining, scrutinizing, assessing and monitoring various strategies,

policies, methods and instruments for media analysis. Their expertise was visible in the valuable comments they made on the SYNC3 prototype which touched not only on functionality and usability aspects but also on product and business strategies. One out of the four participants was a woman. The dominant age group was 35-50, with one participant aged between 25 and 35.² None of the members of the group had been in touch with earlier versions of the SYNC3 system, nor had they heard about the project before having been approached by the SYNC3 consortium.

At Bortun-Olteanu in Bucharest, a PR agency belonging to the Millenium Communications group, a total of four junior and mid-level PR and communication professionals participated in the SYNC3 user evaluations on 19 January 2011. All test participants were proficient in English and had Romanian as mother tongue. They were all accustomed to using the Internet in their daily work, particularly social media. Their daily activities ranged from management of PR accounts in terms of budgets, strategies and communication plans, to implementation of online communication strategies by means of monitoring online mentions and managing online reputation for several brands, events coverage and competition monitoring. All participants of this round were women, two of which were aged between 18 and 25 and the other two between 25 and 35. None of the test participants had any prior knowledge of the SYNC3 system.

At the RIA Novosti news agency in Moscow, a total of eight mid-level and senior staffers (editors, journalists, translators, project managers and media monitoring and analytics professionals) participated in the SYNC3 user evaluations over a period of two days, on 20 and 21 January 2011. All test participants were proficient in English and typically had Russian as their mother tongue, as well as knowledge of several other foreign languages such as Spanish, Portuguese, German, Japanese and French. They were accustomed to using the Internet in their daily work and a small number of them were proficient in Google Analytics tools as well. All RIA Novosti questionnaire respondents were male. The dominant group age was 25-35 with the exception of one participant aged 35-50.³ Their daily activities involved coordinating the monitoring of foreign media, coordinating foreign-language news production, media research, analysis and reporting (including international news and opinion trends).

Half of the overall number of respondents held a MA/MSc degree and 40% a BA/College degree.

Evaluation set-up and sequence

Typically, the evaluation sessions were conducted one subject at a time, with a time limit of one hour per session. Exceptions were one session in which two subjects tested the prototype at the same time, and the RIA Novosti evaluation round. Due to the large number of test participants, at RIA Novosti user evaluations were held in groups over a period of two days. A few of the sessions, the ones conducted with senior professionals ran a little over time and one had to be cut short due to the expiration of time over which the testing space was available. Participants were first briefly introduced to the SYNC3 project and to the structure of the session ahead. The limitations of the current prototype compared to the end SYNC3 product were further briefly listed and the participants were invited to ask any questions they may have before starting the actual test.

² Two out of the four Report International participants filled in the evaluation questionnaire to date, so the current analysis is based on data gathered from these two questionnaires and from the evaluation sessions.

³ Six out of the eight RIA Novosti participants filled in the evaluation questionnaire to date, so the current analysis is based on data gathered from these six questionnaires and from the evaluation sessions.

The live prototype was made available by ATC online at: <http://sync3.atc.gr/de.l3s.sync3.gui/> and covered the functions: Exploration and Browsing Interface and the News Story Creator. In order to gather feedback on the User Profiling Infrastructure, a simulation was provided by L3S at: <http://sync3.atc.gr/de.l3s.sync3.gui/groups/>. Overall, the system functioned without flaws with the exception of several cases in the user evaluations round in Bucharest as well as in the two individual tests, when the Overview visualization as well as the News Story Creator were unavailable for testing. In these cases the test participants were presented with a screenshot of the News Story Creator and explained the concept behind it to collect user feedback. Since the focus of the evaluation were the use cases implemented in the live prototype, the User Profiling Infrastructure simulation had a lower priority and was presented to the test participants only when time was left after testing the live prototype. Less than half of the total number of participants was presented with the User Profiling Infrastructure simulation. Participants from all evaluation rounds were presented with a version of the prototype with nearly the same functionalities. One functionality that was not available in the first evaluation round at the EJC was made available for testing to the participants of the rest of the evaluation rounds, namely the Map View.^{xxix}

Live prototype and simulation testing

The testing of the live prototype and the User Profiling Infrastructure simulation followed a script detailing the tasks to be performed by the test users, in accordance to the use cases implemented in the early prototype. The test was divided into three parts according to the three main modules of the UI: Exploration and Browsing Interface, the News Story Creator and, whenever time permitted it, the User Profiling Infrastructure simulation. Each task was accompanied by questions which the test leader could ask in order to stimulate the participant to make observations.

In testing the Exploration and Browsing Interface, for the sake of consistency, all participants with the same professional profile were suggested to run the same query. Journalists and media researchers were suggested to query “BP oil spill.” Whenever the participants had particular information interests they were allowed to run any other query of interest, to enable comparative assessment of the information they got on a topic of interest from the SYNC3 prototype to information they got from other news and blog search engines. In the case of media monitoring and analysis professionals they were asked to search for information about a brand or product they are monitoring in their daily work. Some of the keywords used for search were: “Nokia CEO”, “Gillette” and “Unicef”. The participants were asked to explore the generated results with the list view, map view and the overview. Further, they were asked to filter the results by using the advanced faceted search. During the entire task performance process participants were invited to make observations on search functionality and usability: simple and advanced search facilities, search/retrieval performance, proportion and quality of sources coverage, meaningfulness/relevance and precision of retrieved records, response time, usefulness of the alternative map view and overview in understanding the generated results, etc. To facilitate commentary the test leader prepared specific questions for the participants: *Are the results satisfactorily matching the query? Are the event titles meaningful/clear or not? Do you find it useful or not to have persons, organizations, locations involved in the event extracted and displayed? Do you find this filtering function useful in your work? Why? Why not? Are the SYNC3 search capabilities satisfactory? What other kinds of search capabilities would be useful to have? Would you be interested for example to be able to search by keywords + positive/negative sentiment?*, etc. Typically this part of the session lasted around 20 minutes, also fuelled by the interest of the participants in the search interface and the more general questions about how the system works which this function prompted.

In testing the News Story Creator usability and functionality the participants were asked to perform a series of tasks such as: to select two or more news events from the search interface that they consider related in some way and add them to the workspace; to place them in a group

canvas and name the group canvas; to draw connections between the news events and annotate one of the news events by means of rating and to make observations on functionality and usability guided by questions such as: *Do you find the function of grouping news items by means of the 'group canvas' useful in your work? Are the offered possibilities to comment on the news items satisfactory? Do they enable the type of analysis that you usually perform on media content? What type of activity would you want to perform with the workspace?* Typically this part of the session lasted around 10-15 minutes.

For the User Profiling Infrastructure simulation the participants were explained what it is planned to do and were demonstrated some of its function with the provided simulation. Next they were asked for their opinion on the functionalities that are planned to be implemented with particular emphasis on receiving recommendations at the request of L3S. According to the remaining time, this part of the session lasted between 5 to 10 minutes.

At the request of the consortium partners to collect some user feedback on specific issues, the participants were asked the following questions: *What would the user prefer: to have the same article listed twice in two news events or to have it just once in the most recent event? I.e.: To miss information or to have redundant information? Some articles listed under a news event are "near" duplicates in the sense that the only thing that distinguishes them is the URL. Would you be interested in having them all displayed or just one of them?*

Besides the qualitative input provided by the users, behavioural data on the interaction between user and system was also collected, such as user effort in learning and using the interface. The complete series of tasks and accompanying questions was presented in Section 4 of this report.

Supplementary exercises

The test participants were asked to also fill in on their own a summarizing questionnaire that was emailed in the course of the day. The questionnaire aimed to gather qualitative data about user attitudes towards various components of the SYNC3 system in a format that can be quantitatively processed. The questionnaire contained four sections and referred only to the live prototype. Section one enquired demographic details of the participants (age, gender, education), as well as their job title and a list of programming languages and technologies that he/she feels comfortable working with on a daily basis. Sections two and three aimed to measure user appreciation of system usability and functionality. Section four enquired a general rating of the prototype. Section five enquired basic business exploitation opportunities and was adapted to the participants' professional profile as well as the current state of the prototype. The questionnaire is available at: <http://www.surveymonkey.com/s/sync3questionnaire>. A total of eighteen test participants filled in and returned the questionnaire to date, although participants did not necessarily respond to all questions.

Restrictions and limitations

Before discussing the results, two caveats must be raised. First, as a little over half of subjects were recruited from the ranks of SYNC3 partners, the system was met by a friendly crowd. User evaluations carried out with completely independent third-party subjects might result in slightly more critical or sceptical reactions. Second, the prototype tested at the EJC and Report International had known functional limitations as follows:

- Sentiment analysis was absent.
- The prototype was integrated with a limited portion of the repository (month of July for the EJC evaluation and January to mid September 2010 for the Report International evaluation).
- The user could log into the workspace with his Gmail account but could not save his work, could not create groups or export news stories created in the workspace or register to receive newsletters.

- News articles and blog posts could not be transferred to the Workspace.
- Causal relations between news events were not identified by the system.
- The map view was not implemented by the time of the first evaluation session; it was implemented by the time of the second evaluation session.

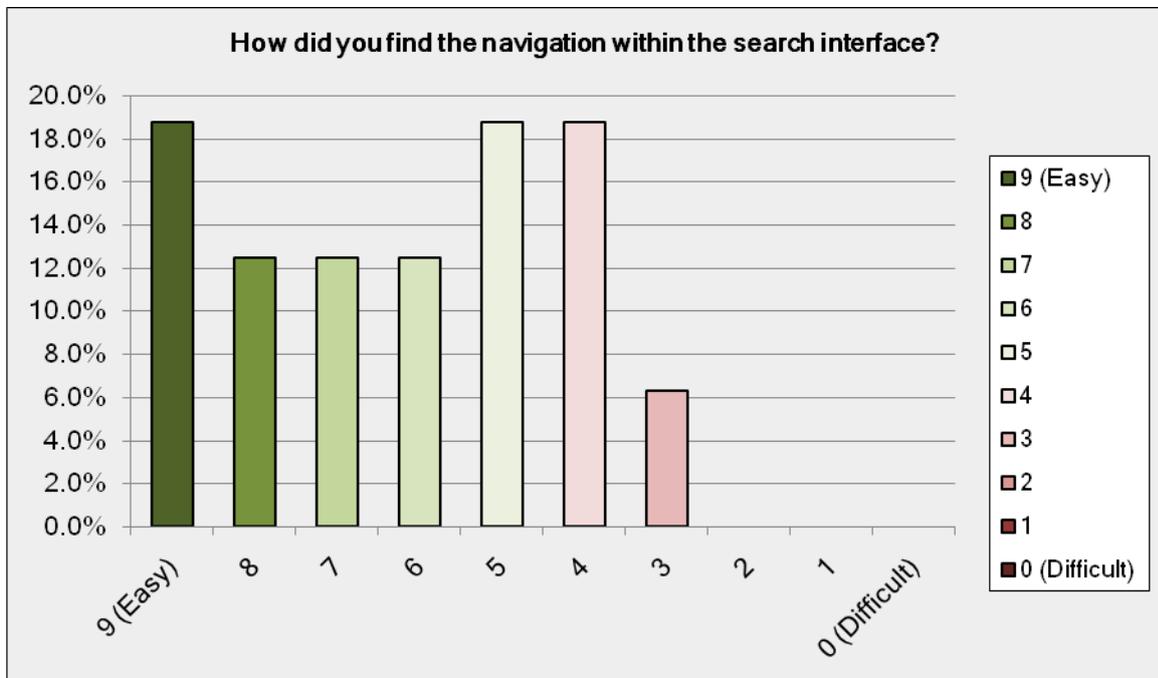
8.2. Questionnaire results

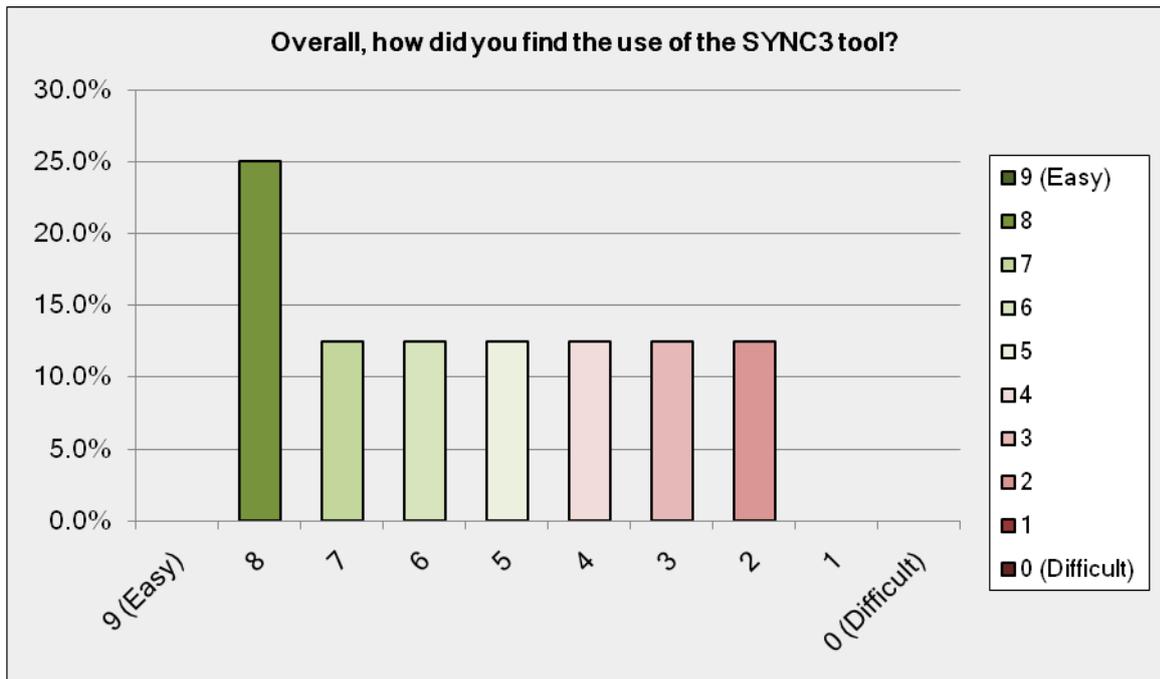
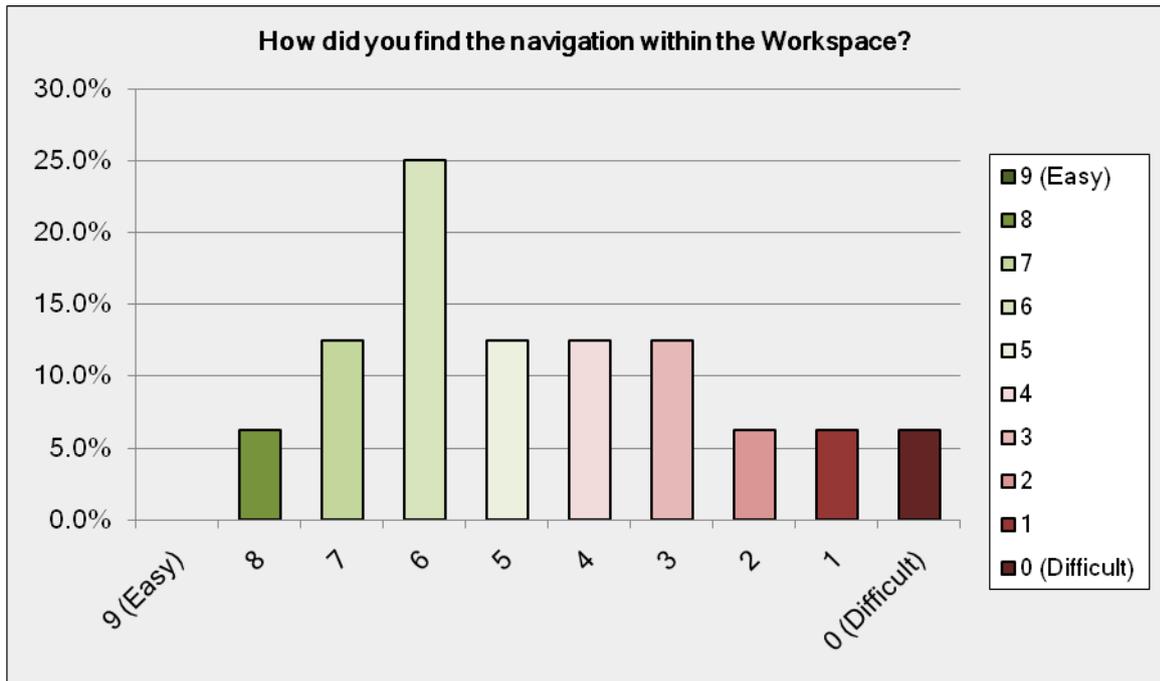
The questionnaire contained a combination of multiple choice questions and open questions requiring free-text input aimed to measure the user’s subjective satisfaction with the system and to gather business and market information. The multiple choice questions allowed subjects to rate on two scales:

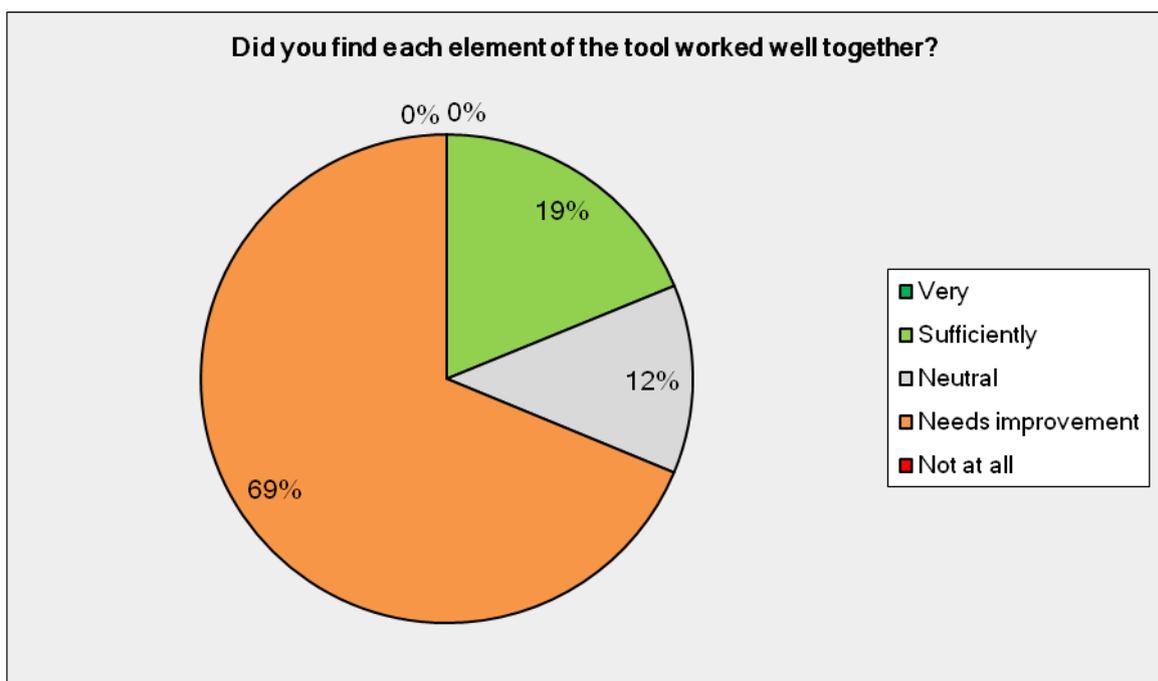
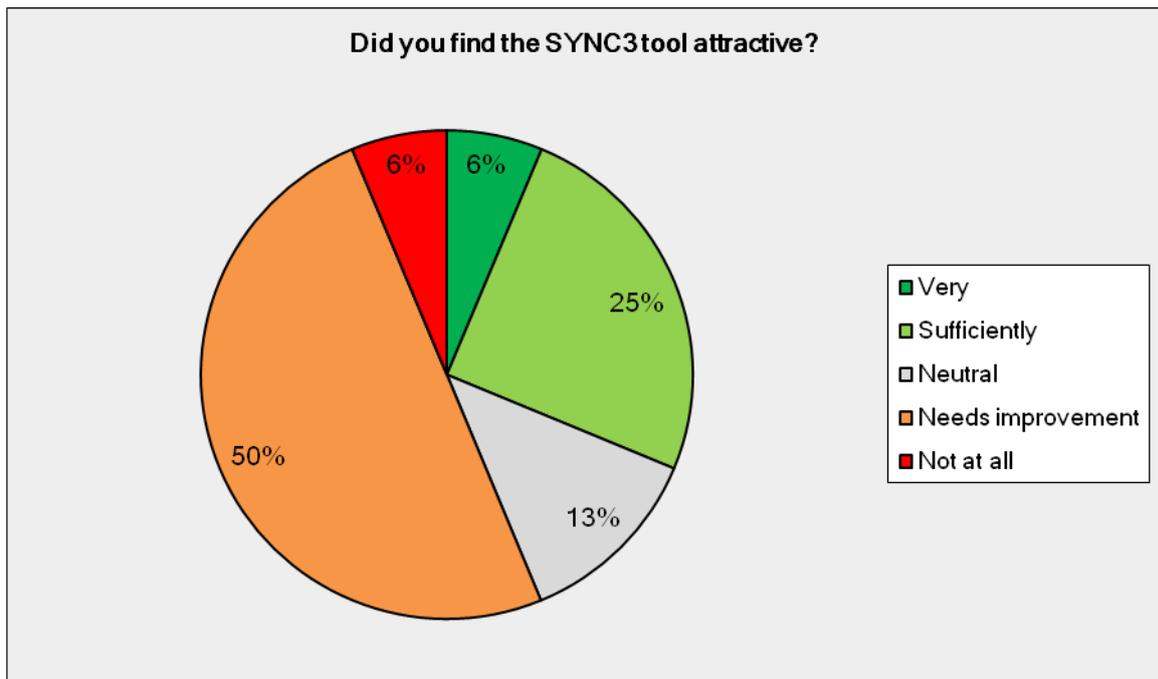
- a comparative rating scale, in which the participants were asked to rate various usability and functionality aspects of the SYNC3 prototype along a well-defined, evenly spaced continuum (a scale of five from very positive over neutral to very negative). This scale is aimed to measure the intensity of their attitude towards the prototype based on the experience of using it during the test session.
- a semantic differential scale, whereby the participants were asked to rate various usability and functionality aspects of the SYNC3 prototype based on a ten-point rating scale that had two bi-polar adjectives at each end, such as “easy-difficult”, “very clear-confusing.” The advantage of using this scale is that compels the participant to choose to a certain extent one or the other qualifying attribute.

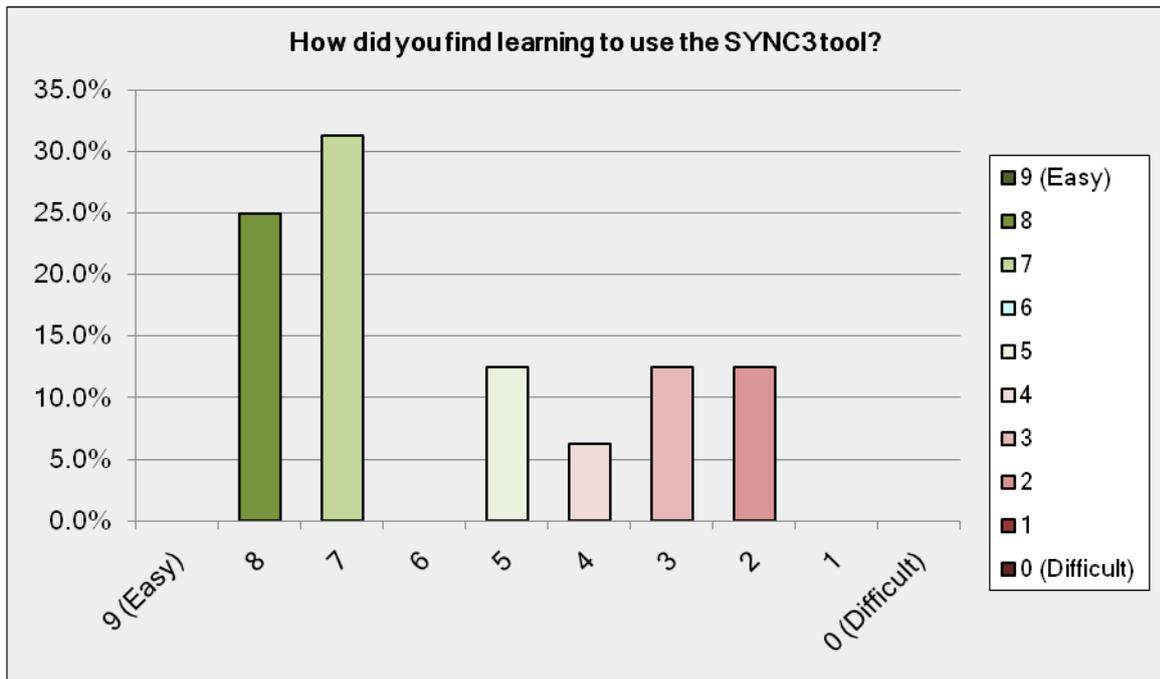
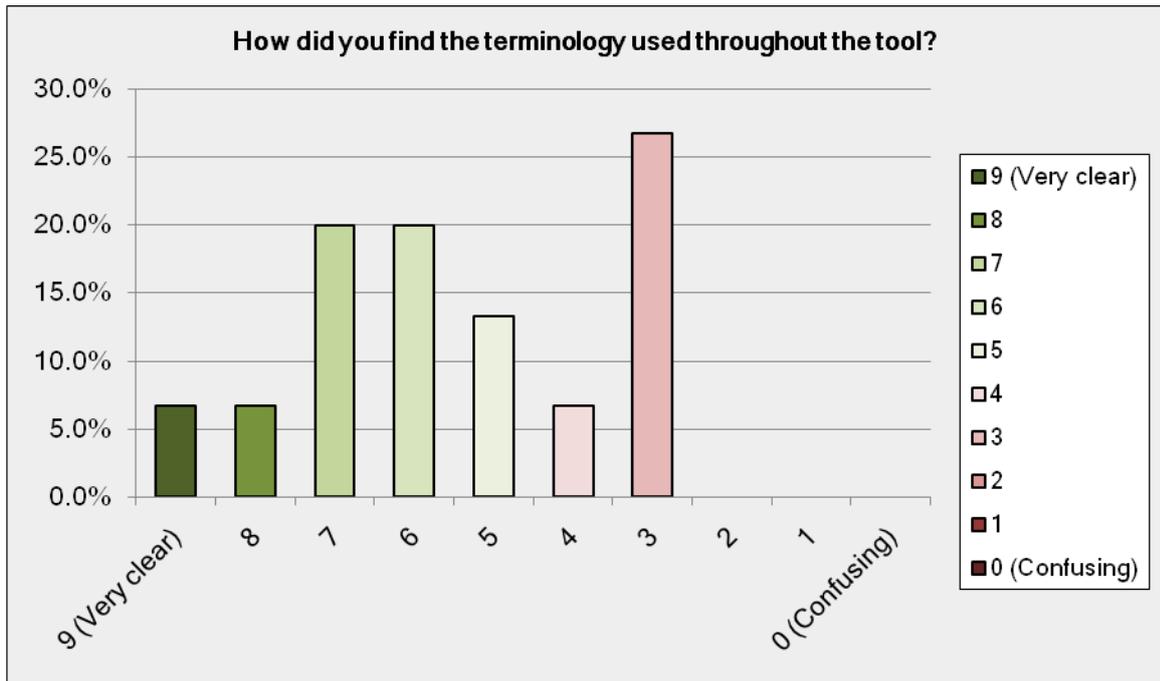
Below are the results of the multiple choice questions for usability and functionality information.

Usability information

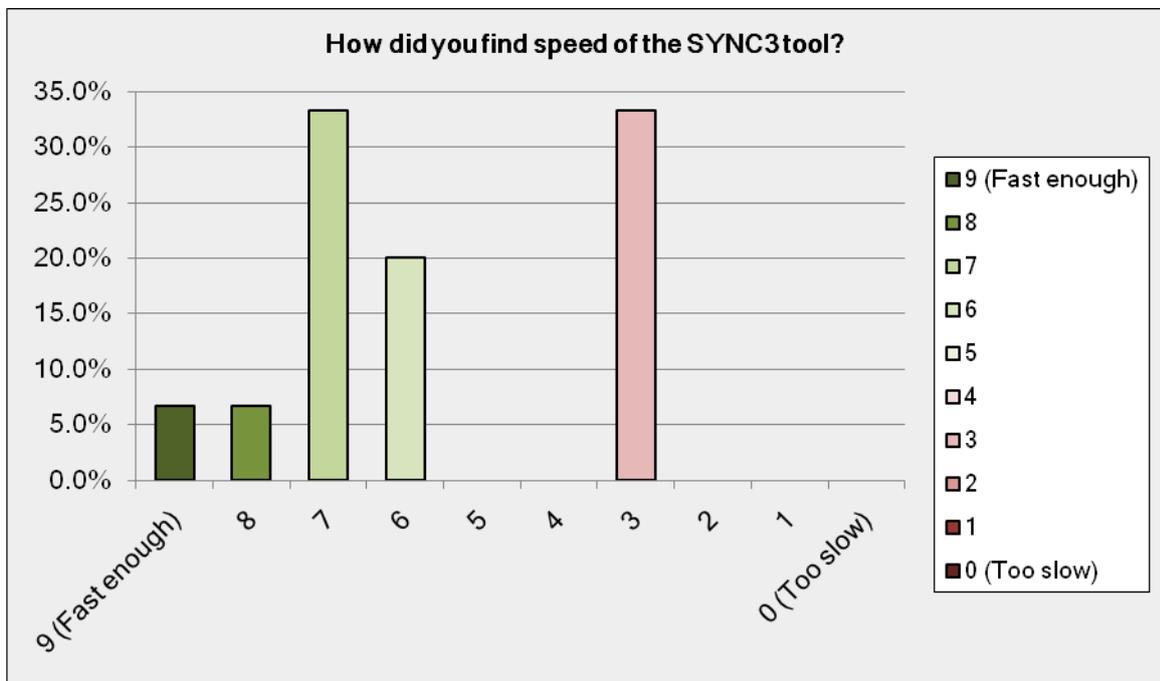
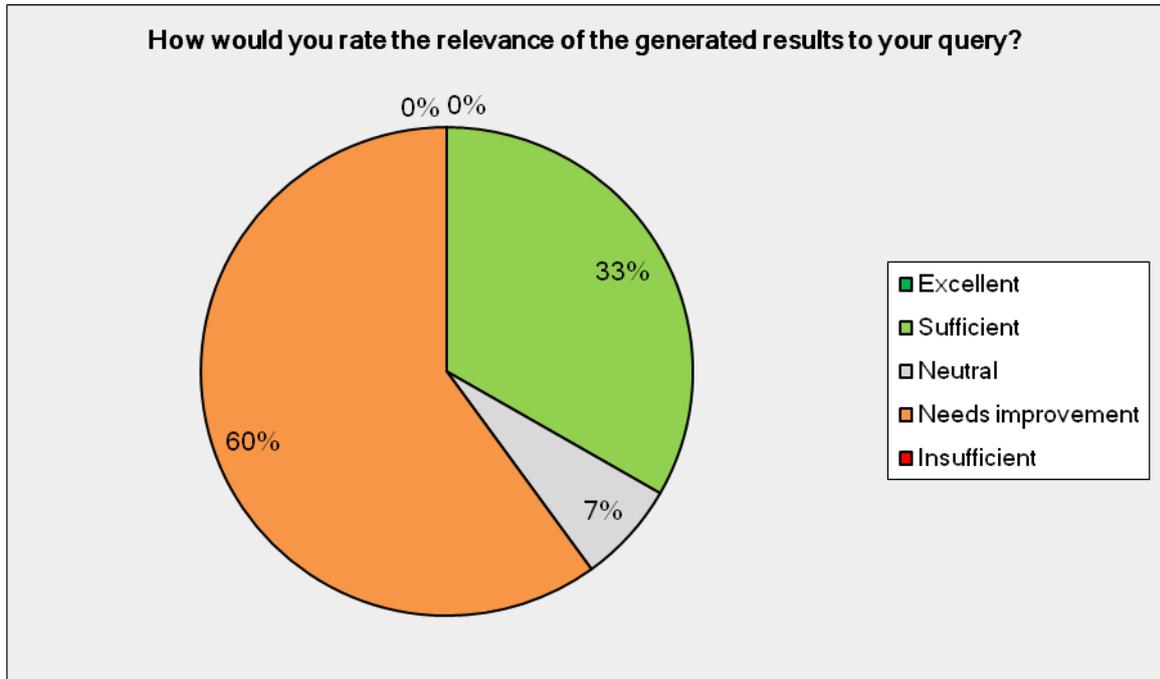


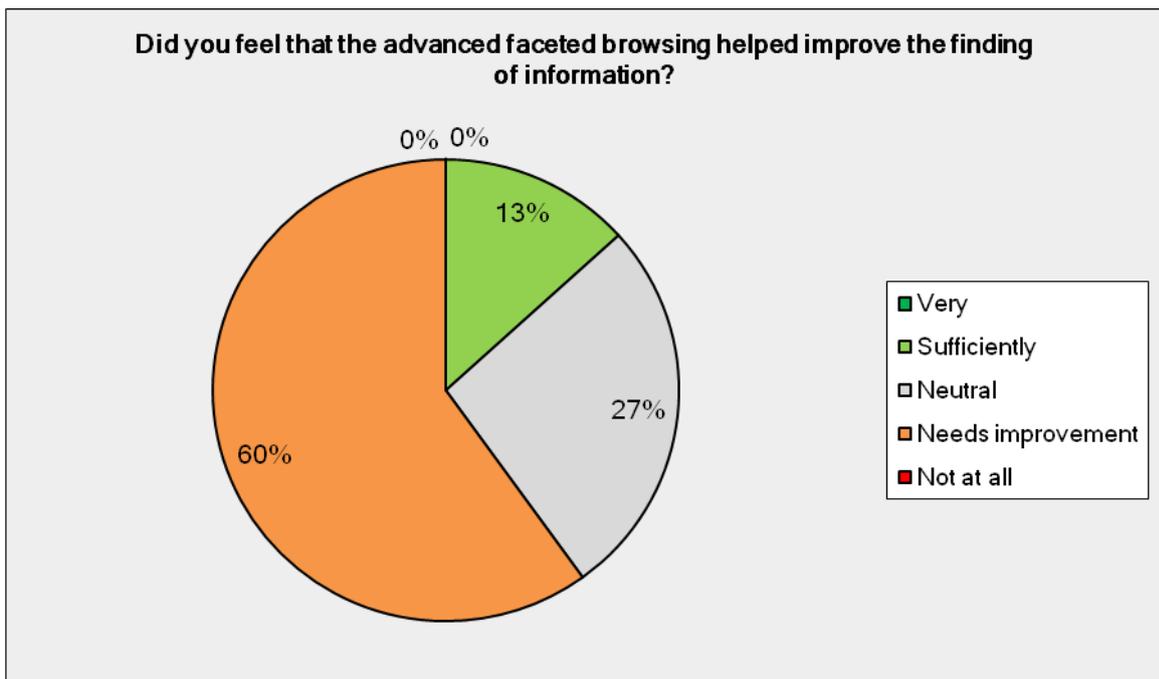
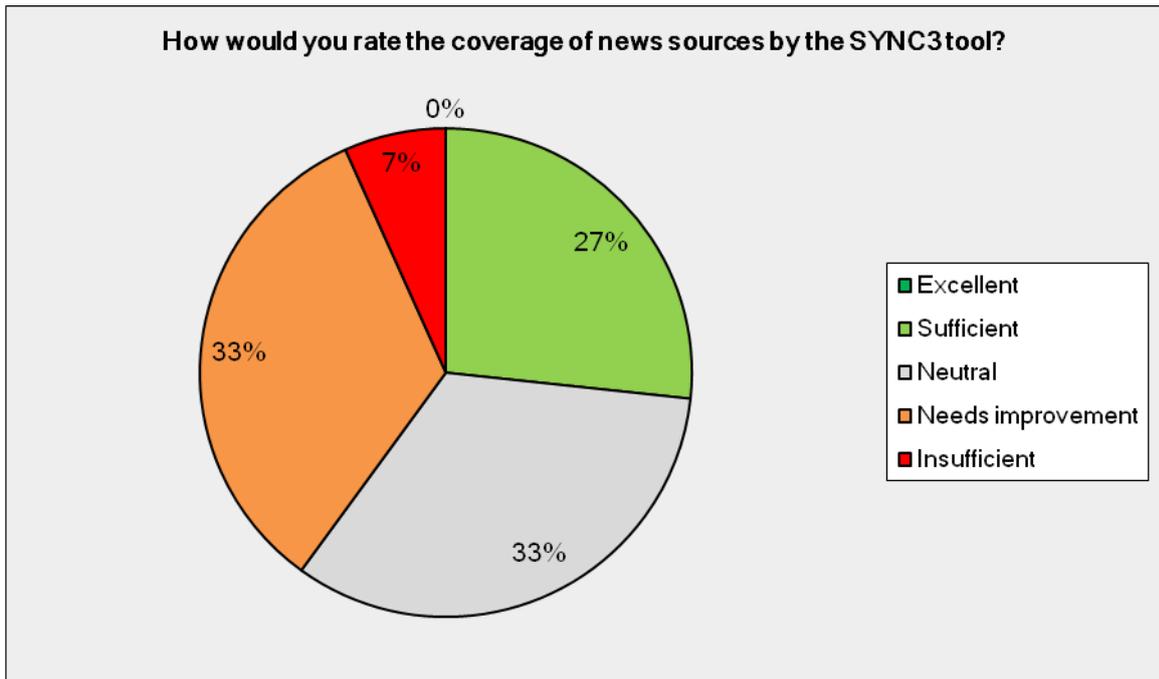




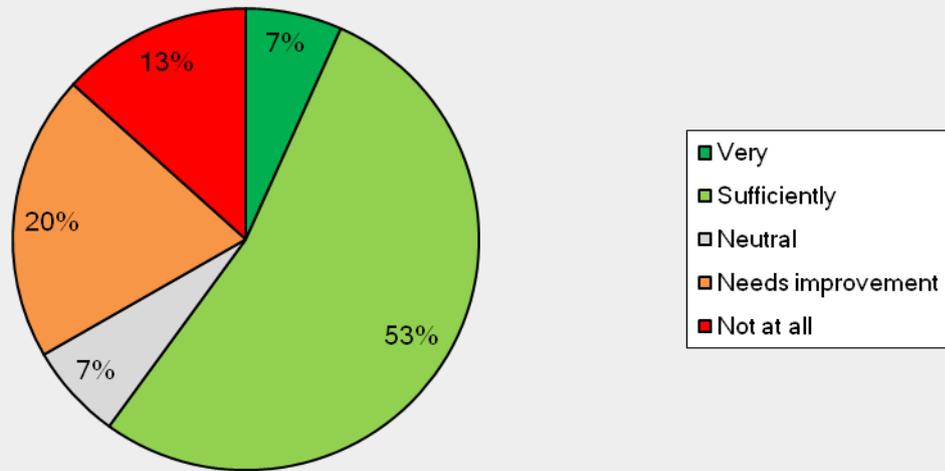


Functionality information

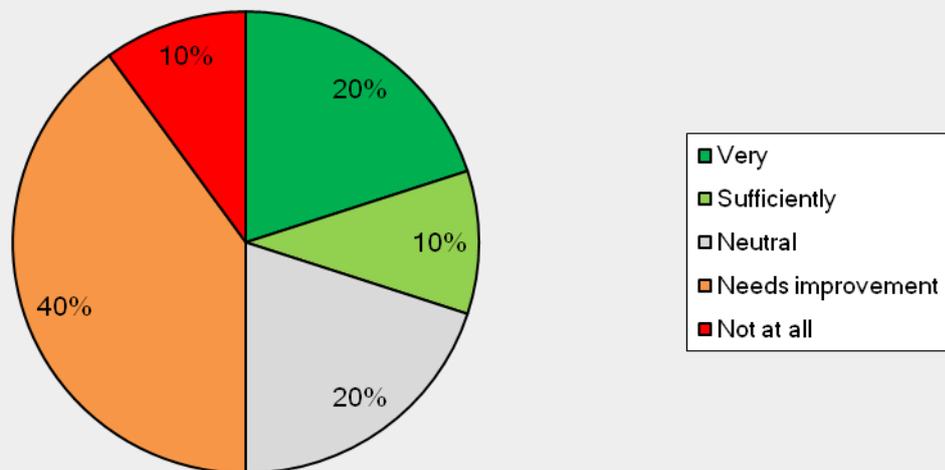


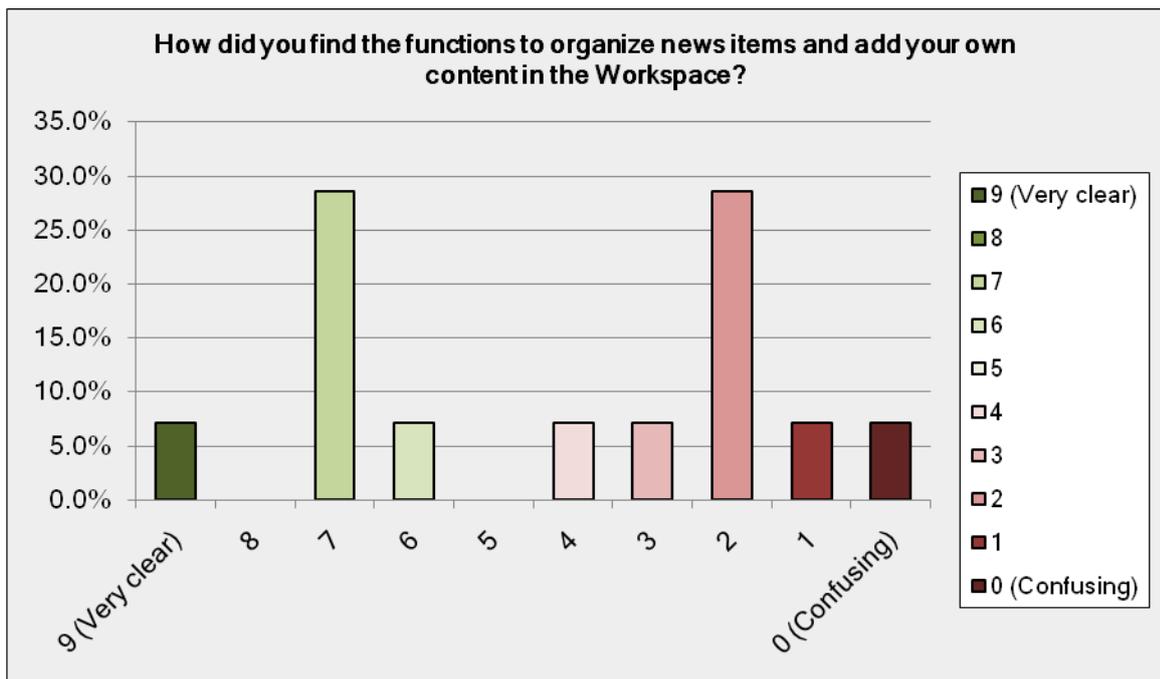
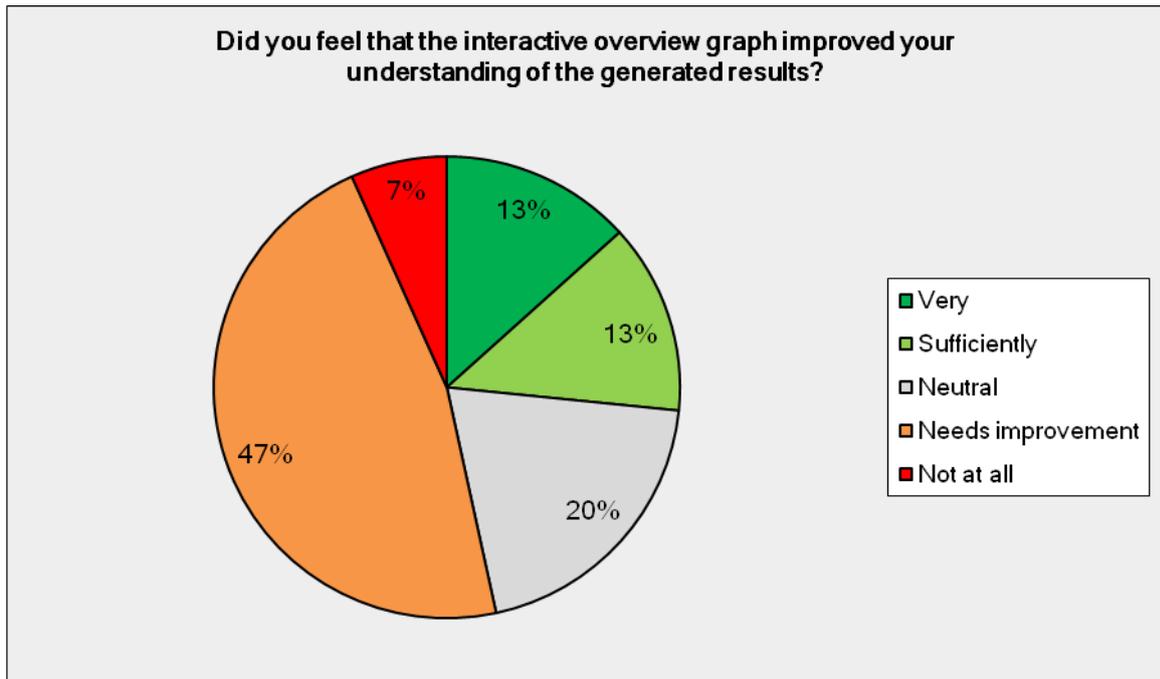


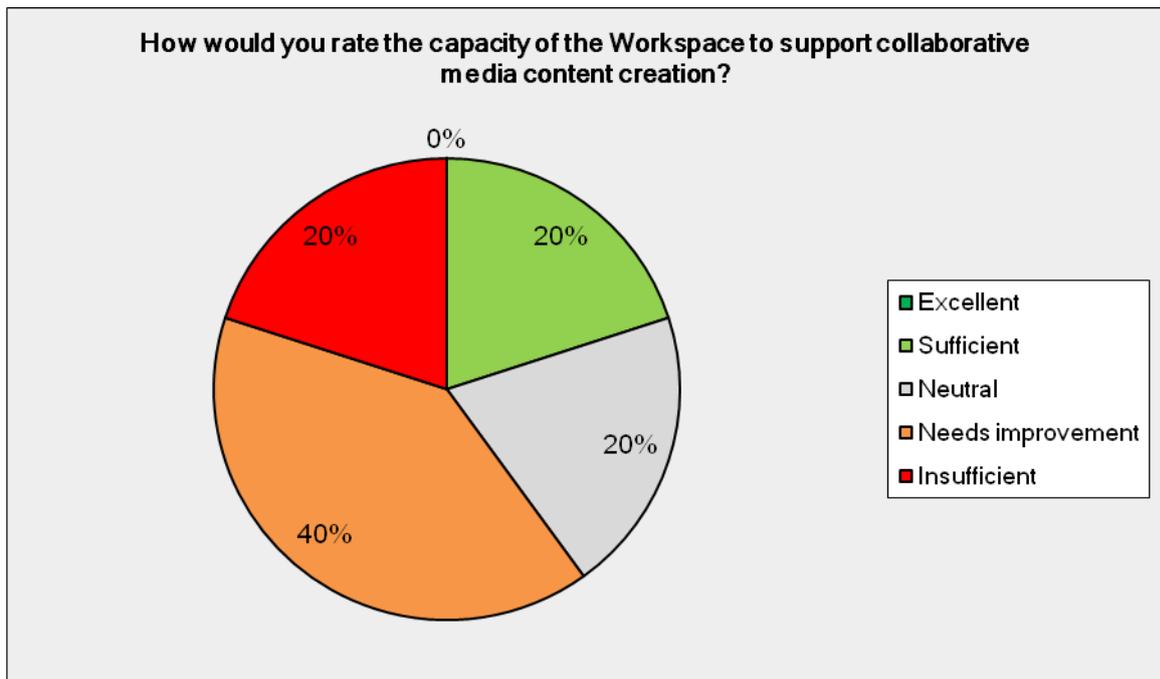
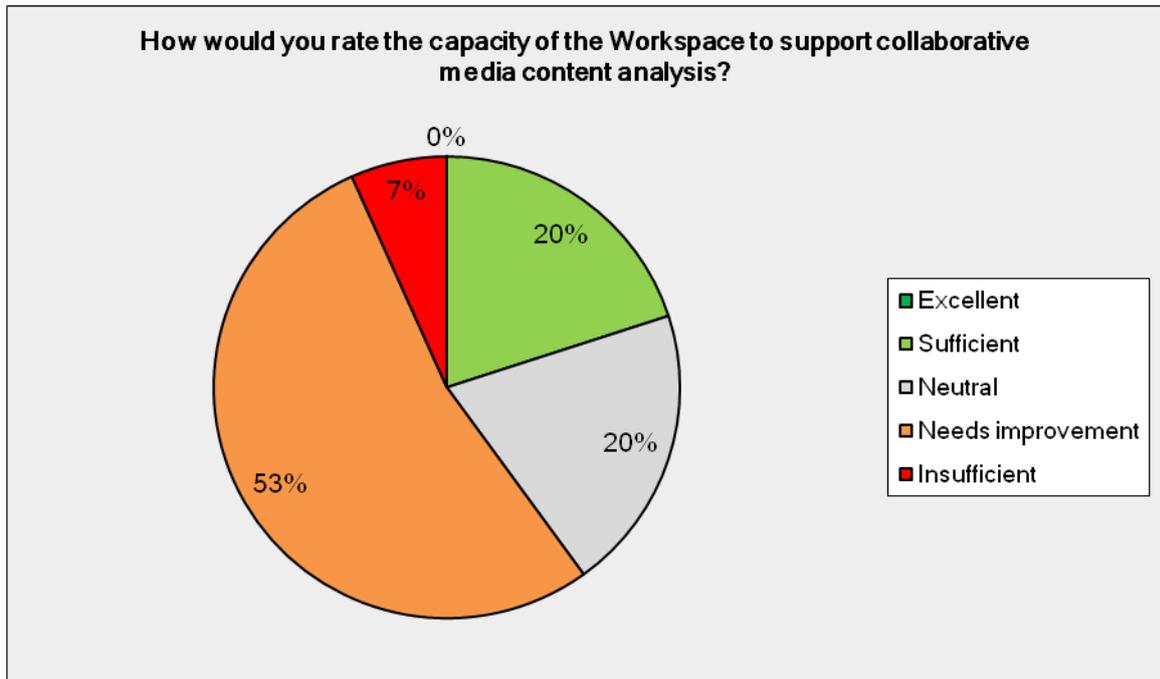
Were the event labels appropriate/clear enough in describing the news events?



Did you feel that the map view improved your understanding of the generated results?







From these results collected through the dedicated questionnaires, the following conclusions can be drawn:

- (1) Questions pertaining to usability have generally received positive ratings as they have been answered by using the points 9 to 5 pertaining to the positive side of choices on the scale, meaning that there were no major frustrations regarding interaction with the system and interface layout. Exceptions were questions referring to navigation within

the Workspace. About 40% of the respondents rated the navigation within the Workspace with values ranging from 0 to 4, with 0 representing “difficult.”

- (2) In terms of functionality, the results of the questionnaire were more critical than the evaluation of the same aspects during testing. About half of the respondents considered that the tool needs improvement in terms of relevance and ranking of generated results to the query, news sources coverage, advanced faceted browsing, map view, interactive overview graph, and capacity of the workspace to support collaborative media content creation and analysis. The rating of the system as “needing improvement” suggests that the concept behind the SYNC3 system has been met with the approval of the test participants, which was one of the main objectives of the first prototype user evaluations, and reflects the naturally “raw” status of a system in its first prototype stage.
- (3) Two thirds of the respondents favourably rated the accuracy of the event labels in describing the news events and over half of the respondents favourably rated the speed of the SYNC3 tool.

As could be expected, free-text responses given were rather varied in depth, and several respondents answered to selected ones only or to none at all. The feedback can be summarized as follows:

Usability

The SYNC3 concept and intention was met with the approval of the respondents who envisaged that its usability, applicability and usefulness will increase with the size of the data set available. To exploit the potential of SYNC3 to the fullest, it was also suggested that effort should be directed towards connecting SYNC3 with a variety of social media platforms. Respondents praised SYNC3 for its capacity to enable better understanding of the dynamics between traditional and social media by linking together news articles with blogs that relate to them and thus syncing together two conversations and meshing their information into one. SYNC3 was deemed a useful tool for journalists and media researchers. Particularly interesting and perhaps unique was considered the breakdown of big events into smaller events that is at the core of the SYNC3 concept, by means of its definition of a news event at a very granular level.

Although well-received at conceptual level, the area most often indicated by the test users as requiring further development and improvement in order to become a usable tool was the Workspace. Besides the envisioned uses, the Workspace was seen as particularly useful in facilitating media archiving. Half of the respondents found that the Workspace needs to be made easier to use and improved to make its commercial utility more obvious. To increase its usefulness, the Workspace must be made possible to integrate with existing user workflows and materials belonging to those work flows by for example opening up the Workspace and allowing materials from outside the SYNC3 tool such as videos, photos, links etc. to be imported and analyzed. One responded indicated that the Overview graph, a useful analysis instrument should also be made possible to transfer to the Workspace.

Another element that was indicated as needing further improvement was the Overview whose presentation needs to be made clearer. Respondents indicated that tutorials are a way to make clearer the way tools such as the Workspace and Overview work and what they do. In terms of search capabilities, respondents indicated that advanced search options by publication date and numeric range would increase the usefulness of the tool. In addition to the planned functionalities, the respondents suggested that it would be useful if SYNC3 also included statistical overviews and visualizations to show connections and relationships between extracted entities. In terms of sentiment analysis, it was suggested that the influence of the bloggers needs to be taken into account when measuring sentiment. *[Test leader's note: a way to further exploit the list of blog source credibility criteria is to somehow include in the listing of blog results a value corresponding to the credibility of the blog or rank them according to credibility.]*

Functionality

This section of the questionnaire received far lesser and less extensive responses than in the usability section. Some observations, such as the necessity to make the workspace simpler and easier to use have been reiterated while other respondents invoked the comments made during the testing on these issues (see Section 8.3).

Business and market information

Comments in this section were connected to the professional profile of the respondents. Four respondents with a communication/PR professional or media analyst profile, one professional journalist, one media researcher and one blogger provided their opinions on business and market opportunities of SYNC3. While acknowledging the limited functionality of the current prototype, the four respondents with a communication/PR professional or media analyst profile positively appreciated the capacity of the SYNC3 to become a useful tool for the media monitoring market provided the tool is integrated with a much larger dataset and more user testing is conducted to gather documentation in order to make it more functional and easy to operate. To become a competitive tool, it was suggested that the Workspace should allow import of different media such as videos and photos, and provide more statistical means of analyzing them along with an archiving option.

In terms of comparison with existing competitors, respondents indicated that there are more advanced commercial solutions on the market for web crawling and suggested that a niche focus would give a competitive advantage to SYNC3: focusing perhaps on political/public sector/NGO themes? The respondent with a professional journalist profile indicated as competitors: Google News, Icerocket, Manymoon and Yahoo Pipes, all of which are being monitored by the SYNC3 consortium. Additional competitors mentioned during the evaluation sessions were Radient 6, moreover.com, analytics.com, and Factiva. The concept behind SYNC3 met with the approval of the respondents with a blogger profile as well, who indicated that SYNC3 would be a useful tool for the blogging community as it enables bloggers to have their voices heard in a more timely fashion.

General conclusions

Far lesser respondents provided responses to this section of the questionnaire and their answers were brief. Besides emphasizing the above-mentioned perceived advantages and drawbacks of SYNC3 (enabling better understanding of the dynamics between traditional and social media, breaking down big events into smaller granularity events vs. Workspace and Overview have to become more user friendly and enable import of materials from outside the SYNC3 system), respondents indicated that while understanding that handpicked sources help keeping the data clean - with regard to comprehensiveness of sources coverage, this is a drawback. This observation meets the agreement of the consortium, as the limited set of sources is designed for the developing stages of the system, while the end product will freely roam the Internet. In addition, the search and browsing functions were positively appreciated. Limiting coverage to English language sources and excluding other types of social media such as micro-blogging services were also indicated as drawbacks of the system. An indicator of blog post opinion in the list of generated results was also deemed useful. Some of the main overall drawbacks indicated by respondents were: lack of sentiment [*Test leader's note: the absence of sentiment analysis is a planned fact given the early stage of development of the system*], and difficulty to use the Workspace. As overall improvements it was suggested that the set of news and blog sources be expanded [*Test leader's note: the expansion of the set of sources is planned in the development of the project*]. A business strategy suggestion that was made is that the system specializes on a particular theme or sector and uses sector-specific intelligent libraries for NLP and automated sentiment).

8.3. Evaluation session results

Exploration and Browsing Interface

Observation of uninitiated users interacting with the SYNC3 Exploration and Browsing Interface yielded the overall impression that they quickly grasped its purpose and main functions. Overall, the responses to the Exploration and Browsing Interface were very favourable. Users appreciated that the tool was clean and clear visually as well as its speed. One user indicated that the search bar should be bigger and have a more central position on the page (at eye level).

Clustering of articles around news events and having results from news and blog sources listed on the same page in separate columns was unanimously appreciated. In terms of volume of results it was remarked a couple of times that more results from news as well as blog sources were expected to be listed, particularly for news events which include emotional words in the case of blog posts. Expanding the sources dataset towards having the system freely roam the Internet in its final version is on the agenda of the consortium. As solution to this issue one senior participant suggested that a way to automatically pull blogs should be examined by potentially checking which blogs link to news sources most often. While junior professionals were sometimes disoriented as to which results represent news articles, blog posts and news events, senior professionals recognized the three types of news items immediately. Interest was shown towards indicating in a more telling way where news coverage triggers blog activity.

In terms of search capabilities, it was remarked that misspellings in queries are not recognized and corrected. It was indicated that searching by date or date range should be enabled as well. Most of the users indicated that the searched keywords should be highlighted in the generated results as visual aid for quick finding of the searched information. The relationship between the searched words and the listed news event labels was not immediately clear to all participants. To make the relationship more clear perhaps the searched keywords should be highlighted in the news event label. Media monitoring professionals were particularly interested in selecting the pool of sources in which the search is being done. An archive of searches was suggested to be posted on the search page along with a list of news categories (politics, culture, sports, etc.).

In terms of generated results, about half of the test participants indicated that not all generated results were relevant to their query. For example, in the case of the query “oil spill Gulf of Mexico,” one of the highly ranked listed results was an event referring to an oil spill in China, where Gulf of Mexico was only incidentally mentioned at the end of one of the articles. For the query “Haiti earthquake” the generated results refer to Haiti elections and not to the earthquake. Participants found it disorienting that in the top listed news events the searched words were absent from the news event label. The participants expected that in the (top) listed results the searched terms would be part of the main reported event and not part of a secondary event. In some cases not all of the searched terms could be identified in the text of the news article or blog post but in other sections on the article page. For the query “Gillette” most of the top listed results do not refer to the razors’ brand but to an individual whose name is George Gillett. *[Test leader’s note: This is due perhaps to the fact that the current ranking algorithm ranks based on a formula which calculates the frequency of the queried words in the retrieved results. In ranking the results the algorithm needs to be tweaked to take into account whether the queried words are part of the main event reported on in the news article or if they are incidentally mentioned in the article.]* It was remarked that not all generated blog posts were related to the events they were linked to either. Local events, such as events related to Dutch politicians, proved not to be covered. The latter issue is expected given the limited set of major news sources that the system is working with in the early stages of development. The connection between the number of results displayed in the top right corner and the number of listed news events was not immediately clear to some of the participants. The pleasing appearance of the news agency logos which accompanies each news article was appreciated. One user indicated that in ranking results fresh results should be ranked higher when they have the same level of relevance to the query. In comparing the results generated by the SYNC3 prototype with the results generated by

Google News Search for the same query, the participants remarked that Google News Search returns a much larger number of relevant articles. *[Test leader's note: Whereas the quantitative difference was anticipated and can be explained by the limited set of news sources indexed in this phase of development of the SYNC3 system, the limited number of relevant results for rather general queries indicates an issue that requests further work.]* It was indicated that the number of items (news articles and blog posts) clustered for which news event should be indicated at the news event label before the user clicks on it to see the news articles and blog posts. One user indicated that there should be an option to allow the user to select what they want to see: results from blog posts or results from news articles.

On a macro-level, one senior participant wondered whether such refined event granularity is necessary at a practical level. To increase the utility of the tool the very refined granularity of events was suggested to be complemented by grouping events into a big cluster. Perhaps the navigation of information should start with a visualization of large amounts of information in big patterns and then break into granularity by means of the list view. The user wants granularity but also wants to see the whole picture. This can be achieved by adding a meta-level where the fine-grained events are sorted under overarching events defined by selected parameters.

The test participants welcomed the idea of having metadata (named entities and dates) extracted and displayed but indicated that more work needs to be done on the way they are being displayed to make the difference between the categories of extracted entities more visible maybe by means of colours. It was however remarked that the metadata extracted from news articles, namely persons, organizations, locations, dates, is not always accurate. For example, the name of a hurricane is listed under persons and the name of a person under organizations. Names are not always displayed in full. Near duplicates (for example: name + surname vs. surname) are not always recognized and are displayed twice as separate entities instead of a unique entity. The various categories of metadata were not of equal importance to all participants. *[Test leader's note: give the user the option to have displayed the categories of metadata that he/she is interested in to avoid information overload].* On the other hand, other categories have been indicated as potentially even more useful to be extracted than persons for example (the user already has an idea about what persons to expect to be involved in an event when searching for information on it), namely topics, products, issues. This observation belonged to a media analyst working in a for-profit media monitoring company. The participant gave the example of commercial applications which extract agenda topics or issues that are setting the debate on a particular theme. To bring the analysis enabled by the system a step further, it was suggested by one of the participants that extracted entities be aggregated by means of drawing connections between them.

The issue of having an indicator of influence of bloggers displayed along with the generated results from blog sources came up in the evaluation session as well. While SYNC3 took credibility as a criterion for selecting the blog sources that are crawled, the test participants with a media monitoring profile indicated that reach or readership of the blog is an important criterion for identifying influencers and should be included in the list of SYNC3 criteria for selecting blog sources to crawl. *[Test leader's note: a way to further exploit the list of blog source credibility criteria is to somehow include in the listing of blog results a value corresponding to the credibility of the blog or rank them according to credibility.]*

It was not immediately clear to the participants who tested the Map View which locations are being displayed on the map. The title of the tab "Map" was also indicated as unclear. The positioning of the "Map" as second tab in the search interface does not match its relevance. The Overview tab could be listed before it. They remarked with surprise that a smaller number of locations than the ones identified in the list view are being represented on the map. Participants whose work takes place within national borders remarked that it would be useful to be able to select and zoom into a region on a map, i.e. a country, and have more specific locations, i.e. cities, displayed for the particular selection. Participants appreciated positively the fact that a snapshot of the article was displayed when mousing over a location on the map.

Sentiment analysis was an issue towards which participants generally showed great interest. As reaction to the description of the SYNC3 mode of sentiment analysis, it was suggested that sentiment analysis be expanded from blog post level to blog source level. The blog post must be considered in its environment, the blog as a whole, when identifying sentiment. A larger context than a blog post should be taken into account when attempting to identify sentiment, namely the type of publication as well. Not all opinion expressed in the blogosphere weighs equally so the degree of influence of the blog in the blogosphere should be taken into account when identifying sentiment.

Advanced keyword search

The idea behind the advanced search function was welcomed by all users and its necessity was even anticipated by some. Some suggestions for its improvement were made as follows:

- Instead of having a separate menu for selected facets, selected entities should be highlighted in the list of facets.
- To simplify the interface, only categories of entities (i.e.: location, organization, person, source) should be initially displayed and the entities themselves should be displayed by clicking on corresponding categories.
- There was a sense of disorientation when, after selecting more entities in the advanced search, the page went blank for a second.
- The trash icon used for deselection creates confusion as it connotes deletion not deselection.
- Most users expected that selected facets would automatically get deselected when doing another search, although some preferred that selections are maintained for future queries as well.
- Bug: the top trash icon/ "deselect all" button does not work.
- The vertical listing of facets is difficult to read and makes it difficult for users to differentiate between the extracted entities. It was not immediately clear to some participants that clicking on the arrows in the vertical list of facets would display more facets. Listing the facets horizontally was suggested instead.

Overview

This dynamic way of visualization of generated news and blog results was met with great interest although it produced some confusion as to what it represents. Generally participants did not grasp quickly what the overview visualization shows and remarked that it was not the most straightforward way of showing information. To maximize its utility it was remarked that it should be made to show the rhythm of a news story, to enable the pulse check of a story and its volume, as well as to enable comparisons in coverage between blogs and news. Below is a list of user and test leader comments on the Overview and the interaction between the user and the Overview:

- To be made more user-friendly, a tutorial was suggested to be made available to understand the Overview. The tutorial should contain more detailed information on menus and a general explanation on how it works (maybe even a video).
- More attention needs to be paid to the presentation and spelling. Some users were disoriented as to why some entities are capitalized and others are not.
- To make the visualization clearer it was suggested that when one entity is selected to be explored with the Overview, the bubbles representing the unselected entities should be more transparent or even better not represented at all.
- When clicking on a bubble the user expected to be sent to the text of the corresponding article(s).
- There was a sense of disorientation as to why the animation is flowing backwards sometimes when the X axis is set to time.

- An option to manually edit the visualization should be made available. Some users, following the model of more commercial products such as Analytica, expected more types of visualizations and statistical analysis to be available.
- The entities displayed for selection in the menu should also be categorized to simplify display.
- Some users wondered whether for a larger time span the Overview would still be clear enough to bring meaningful information.
- Bug: timeline works with only limited time range (a week in July).
- A “select all” button would be useful.
- It was not immediately obvious for some participants that the tag describing a bubble can be dragged outside the visualization when wanting to have a better view of the visualization.
- For comparative media analysis or competition analysis it would be useful to be able to dynamically visualize the evolution of coverage of two search terms in parallel.
- It was noticed that when pressing the pause button you are sometimes being brought to start and suggested that the visualized narration should always start from the point where it was paused.
- Generally the “bubbles” visualization was considered a more clear way of visually representing information than the other two available types of visualizations although some media monitoring professionals accustomed to line charts preferred the alternative lines visualization.
- It was suggested that the Overview should include the option to visualize topics in relation to sentiment and the blog activity plotted around them.
- One user drew attention to one downside of animations: it cannot be included in printed reports. An option for exporting the overview visualization to be included in reports should be made available.
- Test participants with PR and media monitoring background were mostly interested in visualizing results per news source or location.
- The explanation of colour coding is not as important for the user as to be placed so close to the top of the page. The list of categories by means of which to filter the results displayed in the dynamic graph was considered to need a more prominent place on the page than the explanation of colour coding.

News Story Creator

Generally, the test participants remarked that working with the News Story Creator is time-consuming and fiddly. Test users had difficulties in moving items around and making connections between them. To be accepted by the users, it has to be made easier and less time-consuming to use. A tutorial should be made available for this function. A suggestion made to make it easier to use was to enable making connections between items by clicking on the connection end-point instead of drawing arrows. Few found it difficult to understand what it does on a practical level, which is natural given that this function was not envisioned to be used by all the targeted user groups that tested the prototype.

The strong point of the New Story Creator was indicated as the capacity of easily transferring news items in a space where they can be analyzed and archived. While, as expected, not very useful for journalists, it was indicated that it could be useful for teaching journalism because people can edit others’ work. For the media monitoring and analysis market, the Workspace might be used for brainstorming as a starting point in preparing reports if not cumbersome and time consuming. With further development, it could also be used as dashboard to be shared with clients to draw attention to issues. Participants with PR and media monitoring background indicated that the Workspace would be useful in illustrating case studies of media coverage and cross-media influence in the case of crisis communication for example, or in following the results of a communication campaign in the media.

Below is a list of user and test leader comments on the Workspace and the interaction between the user and the Workspace:

- Bug: right clicking to transfer items to the Workspace did not always work.
- Users expected to have news articles exported to the Workspace as well when adding news events to the Workspace.
- Most users indicated that the widgets should be displayed at all times in the top left corner.
- In order to be used for content creation, all Word application capabilities should be made available for Notes.
- Icons such as the folder icon created confusion as they made the users want to click on them to have something opened.
- Potential bug: when the arrow gets off screen it cannot be manipulated anymore.
- Users showed interest towards importing items from outside the tool (videos or pictures) into the Workspace.

User profiling infrastructure:

In terms of receiving recommendations, the participants indicated that they prefer to be asked if they want more information before being recommended items. One user remarked that the recommendations page has the MySpace logo.

Specific issues requesting user feedback:

In the case of near duplicates, would you prefer to have all duplicates listed or just one?

All users responded just one, but one responded near duplicates should be listed if they are updates. Another user indicated that the response to this question depends on the type of use that the tool is being put to. In content analysis frequency does not matter but in coverage analysis any instance is relevant.

[Test leader's note: From the point of view of news distribution research it is important to see all the instances in which a story was published – Google News did so until Sept. 2007. Now it offers showing duplicates as an option. One possible option is to give the user the option to choose between having duplicates listed or not.]

Is it helpful to have the same articles listed in different events?

The responses to this question varied. One user confirmed and suggested to have them prioritized and list at the bottom the ones that have been listed under other events as well. Another user however remarked that even in this situation you get the impression that they are duplicates.

What other search capabilities would you want to have? Would search by keywords and sentiment be useful or not?

Users unanimously welcomed the idea of having the possibility to search by keywords and sentiment.

Business and market information

The idea behind the SYNC3 project was met with great interest and enthusiasm by most of the senior Report International test participants. The managing director particularly showed great confidence in the potential of SYNC3 and expressed interest towards being kept updated on the developments of the project and potentially collaborating at the level of research or of the end product exploitation. The expertise of these senior professionals materialized in suggestions at the level of business strategy as well. It was suggested that SYNC3 would gain a competitive

advantage if it focused on a topical niche such as the European public sector or European politics.

8.4. Table of issues raised by users

In addition to the descriptive summary above, this table provides an overview of the issues raised by users during evaluations and of the wishes expressed. The list will be discussed by the consortium and serve to prioritize actions to be taken during further development. In bold are the top ten issues proposed to be addressed by the consortium from a user’s perspective. It should be noted that the table does not include favourable aspects expressed by users or observed by the test leaders, as these are no cause for further action. Therefore, the following list must not be taken as an overall assessment of the quality of the first SYNC3 prototype.

No.	Area	Issue	Comment
1	Functionality	Improve relevance of generated results to queries.	In ranking the results the algorithm needs to be tweaked to take into account whether the queried words are part of the main event or are incidentally mentioned in the article.
2	UI	Make Overview chart more clear, simple and user-friendly.	
3	Functionality/UI	Make influence of blogger visible in the UI in the list of results and take it into account when calculating sentiment.	
4	Functionality/UI	Set up tutorial to explain the Overview and the Workspace.	
5	Functionality	Open up Workspace to materials from outside the SYNC3 tool.	Allow import of pictures, videos and also the Overview visualization into the Workspace.
6	Business strategy	As there are already on the market advanced solutions for web crawling the best way to go for SYNC3 might be niche focus such as: political/public sector/NGO themes.	
7	Functionality	Have the system recognize and correct misspellings in queries in order to enable it to generate results for queries with misspellings.	
8	UI	Highlight searched keywords in the generated results.	
9	UI	More accurately display the extracted metadata on persons, locations, organizations, etc.	Sometimes a word is not displayed in full.
10	Functionality	Improve accuracy and eliminate duplicates from extracted entities.	Duplicates: name + surname vs. surname, person sometimes thought of as

			organization.
11	Functionality	Map connections between extracted entities.	
12	Functionality	Extract additional categories such as topics, issues.	
13	Functionality	Make representation of location on the Map View comprehensive.	Not all extracted locations are represented on the map.
14	Functionality	Cluster fine-grained events under an overarching event (story).	
15	Functionality	Move sentiment analysis from blog post level to blog level.	Consider the blog post in its environment, in context, to decide whether sentiment is positive or negative – you need to take into account the context, the type of publication to decide on sentiment; sentiment should be defined by the influence of the blog.
16	UI	Highlight selected entities in the list of facets instead of having a separate menu for selected facets.	
17	UI	Eliminate the page going blank for a second after selecting more entities in the advanced search.	
18	UI	Replace the trash icon with an icon that connotes deselection instead of deletion.	
19	Bug	The selected facets are not automatically deselected when doing another query.	
20	Bug	The top deselection icon meant to undo all selections at once does not work.	
21	UI	When one entity is selected to be explored with the Overview, the bubbles representing the unselected entities should be even more transparent or not represented at all.	
22	UI	Presentation in the UI and spelling require more attention.	For example, when are some entities capitalized and when are they not.
23	UI	Enable navigation to articles from Overview.	
24	UI	Add option to manually edit the Overview visualization.	
25	Bug	Animation with time view is sometimes going backwards.	
26	UI	Categorize entities available for selection in the Overview in locations, persons, organizations, etc. as well.	
27	Bug	Timeline works with only limited time range	

		(a week in July).	
28	UI	Add “select all” option to Overview.	
29	UI	Design the Overview to enable comparisons in coverage between blogs and news or between different search terms.	
30	UI	When pausing the overview, the animation should not be brought to start but continued from the point where it was paused.	
31	UI	Add sentiment visualization to Overview.	
32	UI	Make moving widgets around and making connections between items in the Workspace less cumbersome and time-consuming.	By making connection between items by clicking on the connection end-point instead of drawing arrows perhaps.
33	Bug	Right-clicking to transfer items to Workspace does not always work.	
34	UI	Have widgets displayed at all times in the top left corner in the Workspace.	Prezi is an example of neat display of widgets.
36	Functionality/UI	Make more Word application capabilities available for Notes.	
37	UI	Have icons visually represent functionalities or widgets in a non-ambiguous way.	i.e.: the trash icon connotes deletion not deselection, the folder icon for the event widget makes the user click on it to open something, the User Profiling Infrastructure uses the MySpace logo.
38	Bug	When arrows get off-screen they cannot be manipulated anymore.	
39	UI	Near duplicates being essential information or not depends on the use of the tool: for content analysis or distribution analysis. Perhaps add option for the user to choose whether he/she wants to have near duplicates displayed or not.	
40	UI	In the case of the same article being listed under different events have the unique ones prioritized over the repeated ones.	
41	UI/Functionality	Enable search by keywords and sentiment.	
42	Functionality	Expand news sources coverage to include local sources as well.	
43	Functionality	Expand blog sources coverage.	
44	Functionality	Enable search by date/date range.	

45	UI/Functionality	Provide context by sending user to article(s) when mousing over a bubble in the Overview.	
46	UI	Make more immediately obvious that the tag describing a bubble can be dragged outside the visualization when wanting to have a better view of the visualization.	
47	Functionality/UI	Provide more types of statistical overviews and visualizations.	
48	Functionality	Eliminate listing of duplicate blog posts which link to blog post comments.	
49	Functionality/UI	Offer the possibility to select the visualization date range for the Overview.	

9. Conclusions from 1st prototype user evaluations

To date, the objectives of user evaluation of the 1st prototype of the SYNC3 system were successfully achieved. No technical issues showed up in three of the four planned rounds of user evaluations and user interaction was not disturbed by system crashes. As a result, test subjects could concentrate on functionality as well as usability of the system.

The scope of participating subjects to date was good, as all relevant target groups were represented (though with media analysts as a centre of gravity), and several nationalities and mother tongues were present. A mix of early-career, mid-level and senior test participants took part. The testing of the first SYNC3 prototype benefitted from the feedback of several high-level test participants belonging to the executive/decision-making level, although most of the test participants were entry and mid-level professionals, which enabled feedback to span both system functionality and usability aspects as well as business strategy for the final application.

User feedback was very rich in detail. As anticipated in the methodology, the vast majority of subjects commented on a limited number of similar issues; unique issues or observations were the executive-level strategic remarks that only indirectly referred to the qualities of the actual prototype.

As per the DoW, the 1st prototype was supposed to “prove the feasibility of the approach and pave the way for the first version,”^{xxx} by enabling the gathering of user feedback on functionality and usability to help guide the remaining development process. Both these objectives were fully achieved. The SYNC3 concept and intention was met with the approval of the test participants. Particularly well-received was the capacity to enable better understanding of the dynamics between traditional and social media by linking together news articles with blogs that relate to them. The alternative ways of visualization of generated news and blog results, the Overview and the Map View, were met with great interest although the Overview particularly produced some confusion as to what it represents. A series of suggestions to increase their user-friendliness has been collected during the evaluations. Although well-received at conceptual level, the areas most often indicated by the test users as requiring further development and improvement in order for SYNC3 to become a usable tool were: the relevance and ranking of the generated results to the user query, the Overview and the Workspace.

In the coming months, it is paramount that the required further improvements particularly with regard to relevance and ranking of the generated results, the Overview and the News Story Creator, be implemented and that the scope of the SYNC3 prototype be extended to include sentiment analysis which is one of most user awaited functionalities of the system.

10. References

- ⁱ https://hestia.atc.gr/sync3/doku.php?id=d7.4.1-3:user_evaluation_report_ejc_m20_m27_m33
- ⁱⁱ <http://en.wikipedia.org/wiki/Validation>
- ⁱⁱⁱ <http://en.wikipedia.org/wiki/Validation>
- ^{iv} <http://www.softwaremag.com/L.cfm?Doc=archive/2001feb/CollaborativeMgt.html>
- ^v <http://www.softwaremag.com/L.cfm?Doc=archive/2001feb/CollaborativeMgt.html>
- ^{vi} <http://www.softwaremag.com/L.cfm?Doc=archive/2001feb/CollaborativeMgt.html>
- ^{vii} <http://www.softwaremag.com/L.cfm?Doc=archive/2001feb/CollaborativeMgt.html>
- ^{viii} New Models for Test Development, Version 1.0, see: <http://www.testing.com/writings/new-models.pdf>
- ^{ix} Krug, Steve: *Don't make me think*, 2nd Edition, p. 12.
- ^x A newer version of this paper (published in 1994) can be found here: http://www.useit.com/papers/guerrilla_hci.html. Additionally an extensive list of sources discussing the number of users needed to test can be found at: <http://portal.acm.org/citation.cfm?id=92499>
- ^{xi} Steve Krug, *Don't make me think*, p.144
- ^{xii} Steve Krug, *Don't make me think*, p. 146
- ^{xiii} See also: Jacob Nielsen's Alertbox column „Why You Only Need to Test with 5 Users“, (www.useit.com)
- ^{xiv} Hwang, Wonil and Gavriel Salvendy. “Number of People Required for Usability Evaluation: the 10±2 Rule.” *Communications of the ACM* 53.5 (May 2010). p. 131
- ^{xv} Nielsen and Molich in Hwang, Wonil and Gavriel Salvendy. “Number of People Required for Usability Evaluation: the 10±2 Rule.” *Communications of the ACM* 53.5 (May 2010). p. 131
- ^{xvi} Virzi, R.A. “Refining the test phase of usability evaluation: How many subjects is enough?” *Human Factors* 34 (1992): 457-468. Quoted in Hwang, Wonil and Gavriel Salvendy. p. 131
- ^{xvii} Law, L. -C. and Hvannberg, E.T. “Analysis of combinatorial user effect in international usability tests.” *CHI Conference on Human Factors in Computing Systems*, ACM (2004): 9–16; Slavkovic, A. and Cross, K. “Novice HEs of a complex interface.” *CHI '99 extended abstracts on Human Factors in Computing Systems*, ACM (1999): 304–305
- ^{xviii} <http://en.wikipedia.org/wiki/V-Model>
- ^{xix} New Models for Test Development, Version 1.0, see: <http://www.testing.com/writings/new-models.pdf>
- ^{xx} New Models for Test Development, Version 1.0 of 03/28/00 <http://www.testing.com/writings/new-models.pdf>
- ^{xxi} DoW, p. 23
- ^{xxii} Ibidem.
- ^{xxiii} Ibidem.
- ^{xxiv} Kallipolitis, Leonidas, Vasilis Tountopoulos and Nikos Sarris. *D8.2: First Integrated Prototype*. 2010. p. 15. https://hestia.atc.gr/sync3/doku.php?id=d8.2:first_integrated_prototype. Accessed 28 Dec. 2010. p. 14
- ^{xxv} Based on “Table 1: Summary of SYNC3 use cases.” *D8.2: First integrated prototype*. pp. 10-11
- ^{xxvi} Kallipolitis, Leonidas, Vasilis Tountopoulos and Nikos Sarris. *D8.2: First Integrated Prototype*. 2010. p. 15. https://hestia.atc.gr/sync3/doku.php?id=d8.2:first_integrated_prototype. Accessed 28 Dec. 2010

xxvii Ibid. p. 12

xxviii Ibidem.

xxix D 6.1.1. Map View. p. 17

xxx DoW, p. 23