RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D4.2.1

Collecting best practices for diversity-aware collaboration

Editor:	Fabian Flöck, Karlsruhe Institute of Technology (KIT)
Authors:	Ioan Toma, STI Innsbruck (STI); Francisco Javier Caminero, Telefónica Investigación y Desarrollo (TID); Angelika Adam, Wikimedia Deutschland (WIKI); Delia Rusu, Joséf Stephan Institute (JSI); Enrique Alfonseca, Google; Fabian Flöck, Karlsruhe Institute of Technology (KIT); Elena Simperl, Karlsruhe Institute of Technology (KIT).
Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	30/09/2012
Actual delivery date:	30/09/2012
Suggested readers:	Diversity aware tool developers
Version:	1
Total number of pages:	22
Keywords:	deliverable, evaluation, best practices, diversity, opinion mining, social media, Social Web, collaboration, Web 2.0, prototype, extensions, community

Disclaimer

Full project title:	RENDER – Reflecting Knowledge Diversity
Short project title:	RENDER
Number and title of work package:	WP4 – Diversity toolkit
Document title:	D4.2.1 - Collecting best practices for diversity-aware collaboration
Editor (Name, affiliation)	Fabian Flöck, Karlsruhe Institute of Technology (KIT)
Work package leader (Name, affiliation)	Fabian Flöck, Karlsruhe Institute of Technology (KIT)
Estimation of PM spent on the deliverable:	4

Executive summary

This deliverable covers the evaluation methodology for each of the diversity-enabling tools being developed in RENDER as a basis for the collection of the best practices that will be performed in the sequel of this deliverable, which is due in M36. For tools whose evaluation is covered in dedicated deliverables – in particular the ones directly related to the case studies – the evaluation is only briefly outlined, giving a pointer to the respective deliverable(s). Whereas the evaluation methodology and the results of its application essentially assess to which extent the corresponding prototypes fulfil their technical specification, the best-practice collection deals rather with the question of how to optimally fit these tools into their usage environment, and how they effect this environment in turn. In this context relevant issues include the integration of the tool into the work processes and daily practices of the end-users; the effect the tools (and their specific features) have on these processes and activities; and, as actual best practices, the design, development and distribution choices that can or could optimize these factors.

To achieve these aims, the deliverable outlines the evaluation methodology for each of the technical components of the project, on which the diversity-minded tools in WP4 and the case studies are built. The collections of best practices span the usefulness, usability, and acceptance of the tools by the end-users in a real-world scenario, the experiences they made, as well as a number of lessons learned by the developers of RENDER technology and case study prototypes during conceptual planning, design and software development.

Section 2 describes how the different tools extracting and processing diversity-relevant information in Wikipedia are evaluated for their functionality and their usefulness for Wikipedia readers and editors. This will be done by users testing the Task List Generator (mostly editors) and the Article Statistics and Quality Monitor (readers and editors). Their interfaces employ the results from the data mining performed by the components of the RENDER diversity toolkit (see D5.1.2) and will be evaluated for the usefulness of these displayed result scores and indicators as well as the usability of the interfaces themselves. Aspects include as well performance of the tools, felt gain in diversity and most important, the felt usefulness for everyday reading or editing tasks in Wikipedia. This data will be collected via structured questionnaires during a 3-4 weeks lasting test period. Developers will also be questioned for their lessons learned during the conceptualization and implementation of the tools.

The design, preparation and preliminary results of the usability evaluation of the Telefónica Opinion Mining Tool (T-OMT) have been collected in the deliverable D5.3.4, integrating the TwiDiViz tool as an interface. To cover the best practice aspect, there will be detailed question items regarding effectiveness (usefulness for operational daily business of the user), efficiency (realistic effort for use?) and general satisfaction (emotional aspects). Also it will be assessed how the tools fits into the existing infrastructure of the business units at Telefónica that will employ it.

The evaluation for the Google News Summarizer will be included in D5.2.4. In order to more accurately simulate actual use it improves on former evaluations of the techniques used (e.g. Delort & Alfonseca, 2012) that were run in very controlled settings. Therefore, some task-specific evaluation templates are implemented for the summarizer, reflecting realistic usage scenarios. The visualization tool will be evaluated via a user study, where users will be asked to rate it as a whole, its ease of use, as well as individual elements of the interface. Secondly, the participants will be asked to solve a news reading and exploring task using the tool. The best practices collection will be performed with the test users and will include the aspects of usefulness of the tool to online news readers, interface design in terms of intuitive use compared to current newsreader software and importance of the summarization result quality vs. interface quality for perceived usefulness and usability. Regarding the user interface, we will perform controlled experiments with a User Experience (UX) expert to finalize the first version. Next, we will identify a set of users interested in using the tool on a day-to-day basis will be gathered, and usage studies based on activity logs will be performed to identify the most widely used controls in the tool, followed by iterations building on these results.

List of authors

Organisation	Author
Ontotext	Mariana Damova
Telefonica	Javier Caminero
Google	Enrique Alfonseca
Wikimedia	Angelika Adam
STI	Ioan Toma
KIT	Elena Simperl, Fabian Flöck
JSI	Delia Rusu

Table of contents

List of figures

Abbreviations

CMF	Content management framework
CMS	Content management system
KDO	Knowledge Discovery Ontology
KIT	Karlsruhe Institute of Technology
LOD	Linked Open Data
RDF	Resource Description Framework
SDK	Software Development Kit
TID	Telefónica Investigación y Desarrollo
T-OMT	Telefónica Opinion Mining Tool
TwiDiViz	Twitter Diversity Visualization Tool
UI	User Interface
URI	Unique Resource Identifier
WP	Work Package
SMW	SemanticMediaWiki

1 Introduction

This deliverable covers the evaluation methodology for each of the diversity-enabling tools being developed in RENDER. This methodology forms the basis for the collection of the best practices that will be performed in the sequel of this deliverable, which is due in M36. For tools whose evaluation is covered in dedicated deliverables – in particular the ones directly related to the case studies – the evaluation is only briefly outlined, giving a pointer to the respective deliverable(s). Whereas the evaluation methodology and the results of its application essentially assess to which extent the corresponding prototypes fulfil their technical specification, the best-practice collection deals rather with the question of how to optimally fit these tools into their usage environment, and how they effect this environment in turn. In this context relevant issues include the integration of the tool into the work processes and daily practices of the end-users; the effect the tools (and their specific features) have on these processes and activities; and, as actual best practices, the design, development and distribution choices that can or could optimize these factors.

To achieve these aims, the deliverable outlines the evaluation methodology for each of the technical components of the project, on which the diversity-minded tools in WP4 and the case studies are built. The collections of best practices that follow span the usefulness, usability, and acceptance of the tools by the end-users in a real-world scenario, the experiences they made, as well as a number of lessons learned by the developers of RENDER technology and case study prototypes during conceptual planning, design and software development.

The remaining sections of the deliverable cover the evaluation and best-practice collection plans for the Wikipedia diversity toolkit, the Drupal extension, the Twitter Diversity Visualization Tool, including its integration into the Telefónica use case, and the news aggregation and diversification service developed in the Google use case.

2 Wikipedia tools

In this section we describe Wikimedia Deutschland's plans to evaluate the suite of tools, which will allows Wikipedia editors and other contributors to create a more diversity-minded Wikipedia.

2.1 Tools to manage diversity in Wikipedia

Within the Wikipedia use case we develop tools that support Wikipedia users (readers and editors) to find, understand and curate biased and other low-quality articles. We analyse different aspects of diversity, which are relevant for Wikipedia (as described in D5.1.1). Building upon the insights gained in this analysis we build analysis tools for each approach to enhance diversity. These tools are presented within a toolkit on the Wikimedia Toolserver[1] and represent show cases for each analysis approach.

The following five analysis tools as show cases are currently part of this toolkit:

- The LEA (Link ExtrActor)[2] is a tool to explore the thematic coverage of a Wikipedia article with the help of Wikilinks, which are internal language specific Wikipedia links.
- The Change Detector[3] is a tool to explore the timeliness of Wikipedia articles by observing the edits in different language versions.
- WikiGini[4] is a tool to visualise inequality in the distribution of ownership in a text by analysing and measuring the Gini coefficient of word authorship in an article over time.
- The Wikipedia Map[5] is a visualization of all geo-tagged articles of one language edition of Wikipedia at a time.
- Corpex[6] is a tool to explore corpora extracted from Wikipedia in several different languages. It provides visualizations for frequencies in a corpus, in particular the most frequent words, letters, and short two word terms.

The toolkit is subject to on-going development and we plan to extend it with further tools during the next months of the project. More details about its current and planned future scope are available in D5.1.2. The results of each individual analysis tool are used in two so-called *supporting tools* for enhancing diversity in Wikipedia:

- An Article Statistics and Quality Monitor (ASQM), which in particular enables readers to get a very quick overview of the status of an article.
- A Task List Generator (TLG), which enables established authors to create lists of articles belonging to a preferred category or the intersection of categories which need to be improved according to a certain flaw.

These supporting tools are described in detail in D5.1.2. There the functionality und the usage of the individual analysis approaches is explained elaborately. Both tools will be expanded if further analysis approaches are available or requested from the Wikipedia users.

During the evaluation process we are going to test if both tools are useful for the intended target groups quantitatively and qualitatively. As part of this evaluation we are going to evaluate the usability and performance of the single analysis approaches describes above.

[1] http://toolserver.org/~RENDER/toolkit
[2] http://toolserver.org/~RENDER/toolkit/LEA/
[3] http://toolserver.org/~RENDER/toolkit/ChangeDetector/
[4] http://toolserver.org/~RENDER/toolkit/WIKIGINI/
[5] http://toolserver.org/~RENDER/toolkit/WikiMap/
[6] http://toolserver.org/~RENDER/toolkit/Corpex/

2.1.1 Evaluation process

Each support tool has its own user group. Hence we will test these tools and assess the various diversity metrics separately.

2.1.1.1 Qualitative evaluation

ASQM

As mentioned above, the **Article Statistics and Quality Monitor** is aimed at supporting in particular readers to understand if an article has a bias in a certain direction and hence lacks quality. We are going to test this tool with a small group of 15 – 20 test users, incorporating specifically readers who never edited Wikipedia before. Wikimedia Deutschland provides several workshops to teach about writing and using Wikipedia. The audiences for those workshops are for example students[7], teachers[8], elderly people[9] or women[10], who are interested in Wikipedia, her sister projects or creation of free knowledge in general. We are going to work very close together with our colleagues from the department of education and knowledge[11] to get in touch with interested readers from different workshop target groups who want to become contributors.

The test groups will be supported by installing the ASQM gadget in Wikipedia. Each test user is requested to assess the tool with a number of articles. During this test period, test subjects will be requested to answer a questionnaire, which covers several aspects:

- Usability and performance of the ASQM tool
- Correctness, usability and performance of single statistic metrics and diversity analysis results
    - Currently involved are: LEA and the Change Detector
    - Soon will be involved: WikiGini and the results from JSI's NewsFeed
- Usage of analysis results to understand the status of an article and gain in diversity
- Usage of provided further information to start contribution
- Suggestions and further feedback

TLG

The **Task List Generator's** purpose is to support established authors to cure articles belonging to a certain category or intersection of categories, i.e. a preferred are of interest an editor is usually working on.

We are going to test the TLG with two user groups in Wikipedia. These groups are organised in portals and WikiProjects. In the German Wikipedia, these are called *Redaktionen* with some of them additionally following their own quality assurance procedures on top of the general Wikipedia quality protocol. The members of these groups cure and work on a specific field e.g. biology, chemistry, linguistics or medicines (see D5.1.1, section 2.1).

During the test period of 3 - 4 weeks, the users will be requested to answer a questionnaire, which covers the following aspects:

- Usability and performance of the TLG tool
- Correctness of single diversity and flaw analysis results
    - Currently are involved: the Change Detector analysis results

[7] http://wikimedia.de/wiki/Hochschulprogramm
[8] http://wikimedia.de/wiki/Schulprojekt
[9] http://wikimedia.de/wiki/Silberwissen
[10] http://blog.wikimedia.de/2012/06/14/wissen-teilen-frauen-starten-gemeinsam-in-die-wikipedia/
[11] http://wikimedia.de/wiki/Bildung

- Soon will be involved the results of: LEA, WikiGini and JSI'S NewsFeed

- Usage of analysis results and provided information to understand and cure biased articles
- Feeling concerning the gain in diversity
- Suggestions and further feedback

2.1.1.2 Quantitative evaluation

Additionally, we are going to analyze the acceptance and usage of the tools with these quantitative metrics:

ASQM

- How many people actually installed the gadget?
- How often is the ASQM requested?
- For what kind of articles is the ASQM predominantly requested?

TLG

- How many list generations have been requested?
- Which kinds of flaws are searched for?
- How many articles are called from a generated task list?

2.1.2 Collection of best practices

We are already in contact with single Wikipedia community members and members of WikiProjects who are interested in testing the tools. We are going to set up several questionnaire runs with different target groups as described above.

During the evaluation we will use these questionnaires to collect answers to the following questions:

- How do the readers and editors accept the tools? Why do they (not) accept/use them? What can be done in terms of design choices and development process, as well as guidance and support, to improve this?
- What could be done in general to improve the development process of such tools?
- How do the tools affect the work of the editors? Do they improve it? If not, what would have to be changed to do so?

In addition to the evaluation of the supporting tools, we are going to interview tool developers to gain best practices of their experiences. We already are in contact with developers and informed them about RENDER during several events like the Berlin Hackathon 2011/2012 and the RENDER/WikiData-Summit (see D6.1.4, sec 2.2.2; D6.2.4, sec. 2.1.2).

To contact these developers we will use the Toolserver mailing-list and ask to answer a questionnaire which collects answers to the following issues:

- Experiences during the development of their own tools
- Experiences during the integration process of their tool in the framework of the supporting tools ASQM and TLG

2.2 List of action items and milestones

The following steps during the evaluation and feedback process will be carried out:

1. Initialising processes until M27
    - Identification of test groups
    - Questionnaire creation
2. Testing period until M30
    - Instruction and guidance of usage
    - Supporting test process
    - Collection of questionnaires
    - Analysing feedback
    - Identifying the consequences of feedback to improve the supporting tools
3. Final feedback collection commencing at M32
4. Evaluation of final feedback M32 – M34
5. Report of evaluation results and best practices at M36

3 Drupal extension

The extension for Drupal, built by STI Innsbruck, is a module that provides a diversity-aware view on posts written within Drupal. It is able to show extracted topics and named entities and provides links for these. The extension uses core RENDER technologies, namely Enrycher for text processing of Drupal nodes and OWLIM[12] as a backend storage service for storing the diversity information extracted by Enrycher and formalized in RDF using the Knowledge Discovery Ontology (KDO).

3.1 Evaluation

To assess the diversity-aware extension for Drupal, a qualitative evaluation will be performed. We plan to recruit 15 Drupal developers and end-users to ask them to install, use and evaluate the extension. We will provide them with short, semi-structured questionnaires and collect their feedback on the correctness of the extension functionality as well as its usability and usefulness. The diversity aware extension displays inline annotations for the entities identified in the content of the articles. It also shows the diversity information related to the article in case, including related articles that are grouped in three categories based on the predominant sentiment, namely positive, neutral and negative. We will ask the recruited developers and end-users to test all these functionalities and to rate them.

3.2 Usability and usefulness of prototypes in real-world scenarios and collection of best practices

The diversity-aware extension for Drupal will be evaluated considering the following dimensions:

- How do the Drupal users/community accept the tool? Why do they (not) accept/use it or help develop it further? What can/could be done in terms of design choices and development process, as well as guidance, support and community approach, to improve this?
- What could be done in general to improve the development process of such tools?
- How does the tool affect the work of the Drupal editors (propose some metrics)? Does the tool improve it? If not, what would have to be changed for it to do so?

To answer the previous question we will put in place the following strategy.

- Present a first alpha with documentation to the online Drupal community at M25 and a stable beta at M30 (after implementing results of the evaluation)
- Entry in the Drupal project repository
- Prospected target groups/communities: “Semantic Web” Group and/or the “RDF in Drupal 7 initiative” and their respective forums/mailing lists
- Set up a forum/thread for feedback inside the respective group communication channels with some semi-structured questions/threads for issues and other feedback
- Choose contact person and establish somewhat intensive dialog at least for the first 2-3 weeks to get things started
- Collecting feedback:
    - Feedback form from developers will be collected by monitoring of forums, mailinglists, email-contact
    - End-user feedback will be received through developers

[12] http://www.ontotext.com/owlim

- Timeframe: Constant collection for bug fixing, commencing collection of final feedback by Drupal developers (including mediated end user feedback) at M32

3.3 List of action items and milestones

Based on the strategy described in the previous section we list in this section the list of action items and milestones that we will be implemented in the 3rd year of the project:

- The code and documentation of the first alpha release will be made available to the online Drupal community at M25
- The code and documentation of the stable beta release will be made available to the online Drupal community at M30. Once new updates are available we will update the code and documentation.
- The "Semantic Web" Group and/or the "RDF in Drupal 7 initiative" will be approached and the extension will be advertise on their forums/mailing lists beginning with M25.
- A feedback mechanism (e.g. forum/thread) inside the respective group communication channels will be setup by M26 and feedback will be collected

4 TwiDiViz: Analysis and visualization of diversity in Twitter data (including Telefónica internal tool)

4.1 Introduction

We describe case studies and success stories for using diversity-aware technology and offer best practices when implementing them, based on the experience gathered during the development of the software and the feedback received from RENDER early adopters.

4.2 Evaluation

The design, preparation and preliminary results of the usability evaluation of the Telefónica Opinion Mining Tool (T-OMT) have been collected in the deliverable D5.3.4 (due, as the current document, by September 2012). Below we summarize the most important issues from this evaluation but in case the reader wants to look further into the details, please see deliverable D5.3.4.

By M36 of the project (August 2013) an updated version of the evaluation and the final conclusions about the user experience will be presented.

4.3 Usability and usefulness of prototypes in real-world scenarios and collection of best practices

The development of the T-OMT is being implemented through several phases. In the first stage we have been collecting the requirements, describing them in the deliverables *D5.3.1-Initial requirements specification Telefónica case study* and *D5.3.2-Final requirements specification Telefónica case study*. By March 2012 an initial design of the application was released as well as a preliminary proposal for the integration of the RENDER capabilities in the Telefónica case study, following the before requirements. It was presented in the deliverable *D5.3.3-Initial version of diversity information extensions for Telefónica tools.* From that initial design we intensified the implementation work and an initial prototype of the T-OMT has been prepared and evaluated (*D5.3.4-Initial test and validation Telefónica case study).* Through this user testing we aim to obtain important feedback that helps to guide the following steps regarding the UI design as well as the integration of the RENDER diversity information skills. The final results will be presented in *D5.3.5 Final version of diversity information extensions for Telefónica tools* (June 2013) and *D5.3.6-Final test and validation Telefónica case study* (August 2013).

The initial evaluation process is based on the definition of usability proposed by the ISO 9241-Part 11 (ISO/IEC, 1998): "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." Thus, the evaluation covers the following dimensions:

- **Functionality/Effectiveness**: T-OMT is intended to provide information to business units about the impact and the sentiment which are associated to different products or campaigns. So, this point will be assessed asking to the participants of the evaluation if the tool works as they expect. It is worth to notice that the adequacy of these participants (i.e. their professional profile) will be very important in order to obtain significant results.
- **Efficiency**: to what extend the T-OMT may help the users in achieve a better performance is something could be measured in short term (e.g. the time taken to perform certain tasks) but also as a long term improvement if it would help to deal with the responsibilities of the staff involved in the business units (e.g. are the visualizations useful?, how may benefit from it?, etc.).

- **User satisfaction:** that is, how pleasant the tool is to be used. That is, how easily the users interact with the system or if the information is provided in clear and understandable format.

In D5.3.4-Initial test and validation Telefónica case study further details about the design of this first user evaluation (i.e. users' feedback, questionnaires used for collecting the subjective information, etc.) are accessible.

Additionally to the T-OMT evaluation we have conceived a planning to collect best practices in the use of this tool. Taking the same evaluators considered for test and validation of Telefónica case study (see D5.3.4), we will provide them, throughout the project time, with different releases of the tool and online forms in order to get feedback of several aspects. The releases and their dates are planned to be as follows:

- T-MOT v0.1 (end of August 2012)
- T-MOT v0.2 (end of December 2012)
- T-MOT v0.3 (end of March 2013)
- T-MOT v1.0 (end of June 2013)

Once a version has been released, an installable package (jointly with a manual of usage and the release notes) will be sent to the early adopters. During the time from a version to the next one the evaluators will be encouraged to use the T-OMT and then, at anytime, report the problems or comments of the current release. To this end, a preliminary online form 13 (see **Fehler! Verweisquelle konnte nicht gefunden werden.**) has been already designed, although there is room to modifications since different aspects from the tool could be highlighted to be analyzed.

Fehler! Verweisquelle konnte nicht gefunden werden. shows the distribution of the aforementioned ilestones.

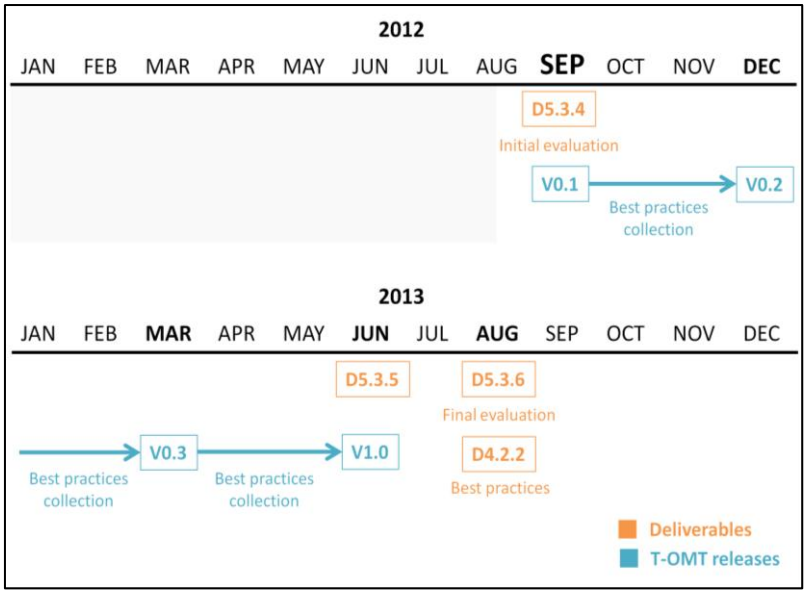

Figure 1 T-OMT roadmap for the evaluation and best practices collection

[13] https://docs.google.com/spreadsheet/viewform?formkey=dFVSMmNGZ3FKeGZCNmlJNDg5dlc5RVE6MQ

5 Google: News aggregation service

This section describes Google's plans to evaluate the news-based use case. It describes the plans to collect best practices in testing and developing the application prototype in the context of Google.

5.1 Introduction

Google collects news content from thousands of sources for Google News, including edited, professional articles and user-generated content such as technology blogs. This information is made available to users of web search, Google News and Google Alerts, amongst other products.

Although the Google News interface offers a diversity of sources describing the same event, in the context of the RENDER project we decided to explore a more explicit way to present diverse information to users, highlighting different points of view that can be found in the news. In most of these products, users specify their information needs or preferences by means of a short textual query. In the work described here, we provide additional controls to the user to automatically generate news summaries highlighting differences in the news. The core technology underlying the application is a combination of natural language processing tools and a text summarization library, with which to generate alternative human-readable, coherent descriptions of news stories.

5.2 Evaluation

5.2.1 Summarization

During development, to allow for a fast turnaround and a short development cycle, we have been using an automatic evaluation framework, consisting of previous test sets from the competitive summarization evaluations in the Text Analysis Conference (TAC). We participated in the 2011 competition and, although we were able to obtain state-of-the-art results on datasets from previous years, some of which were published (Delort & Alfonseca, 2012). After a few months iterating on that dataset we came to the conclusion that these evaluations are run in very controlled settings, which do not always reflect real usage scenarios. So, for example, in these competitions the news collections are always of the same size, and the documents are carefully selected to make sure that they are all about the same topic. These are two assumptions that hardly ever happen in real life, where automatic news clustering techniques are noisy and sometimes include off-topic news, and for some events there may be only a few articles available. Therefore, trying to optimize on this dataset would not necessarily translate into improvements in the real-world application.

In order to have a more realistic evaluation for our task, we have implemented some task-specific evaluation templates. Figure 2 shows a template example for evaluating, at the same time, the quality of the summary (informativeness), whether it reflects the expected polarity (positive or negative) and whether it shows information focused on the relevant entities selected by the user. This, applied to a random sample of news and possible user selections, allows us to track progress over time and ensure that the quality of the system is improving.

Now that we have built the first prototype system, aside from this template which allows us to track progress over time, a side-by-side template will also be developed, for comparing two alternative versions of the system, where the rater needs to choose one of the two possible implementation in blind settings, i.e. not knowing which side corresponds to which version for each evaluation item.

Deliverable D5.2.4 will describe the final results of this evaluation.

Please make yourself familiar with the news below:

Oil prices rise ahead of key European bank meeting
AP The price of oil rose above $96 a barrel on Thursday ahead of a meeting of the European Central Bank which is expected to announce a plan to help ease the eurozone's debt crisis. The ECB is expected to announce a bond-buying program to reduce high ...

Oil price rises ahead of key European bank meeting
2 hours ago • AP BANGKOK (AP) — Oil prices rose Thursday ahead of a meeting of European central bankers who are expected to announce a plan to help financially strapped countries. The European Central Bank is expected to announce a bond-buying program ...

Oil Rises a Second Day on U.S. Supply Drop, ECB Plan Optimism
Oil rose for a second day in New York amid signs of a reduction in US crude supplies and as European Central Bank President Mario Draghi prepared to outline his plan to stem the region's debt crisis. Futures gained as much as 1.3 percent after the ...

Oil Gains a Second Day as US Stockpiles Drop to Five-Month Low
By Ben Sharples on September 06, 2012 Oil rose for a second day in New York after an industry report showed stockpiles shrank to the lowest in more than five months in the US, the world's biggest crude consumer. Futures gained as much as 0.9 percent ...

Imagine that a user has specified that he is interested to know about the European Central Bank, and wants to read a summary of these news articles that highlights the most positive relevant information included in these news.

The user is presented with this summary:

ECB expectations boost crude oil prices

Crude oil futures settled slightly higher Wednesday, awaiting guidance from Thursday's US oil inventory data and signals from the European Central Bank's policy meeting.

US gasoline stockpiles probably fell 3 million barrels last week, according to the median estimate of nine analysts in the Bloomberg survey before the Energy Department report.

Does this summary accurately reflect the main news event reported in the original articles?

[ ] Strongly disagree [ ] Disagree [ ] No opinion [ ] Agree [ ] Strongly agree

Does this summary include a positive view on the event?

[ ] Strongly disagree [ ] Disagree [ ] No opinion [ ] Agree [ ] Strongly agree

Is this summary relevant about the European Central Bank?

[ ] Strongly disagree [ ] Disagree [ ] No opinion [ ] Agree [ ] Strongly agree

Figure 2 Evaluation template for summaries.

5.2.2 Enrichment and representation to end-user

The end-user will be provided with a diversified view on the information extracted and aggregated from Google news clusters. This diversified view will contain, along with the cluster summary, the entities identified within the clusters as well as sentiments expressed in the news articles. The end-user will interact with the visualization tool across these three diversity dimensions: aggregated summaries, extracted entities and sentiments, fix one of the dimensions e.g. select an entity of interest and observe the effects on the other two related dimensions.

The visualization tool will be evaluated via a user study. The participants to the study will be presented with the interface and asked to rate it as a whole, its ease of use, as well as individual elements of the interface. Secondly, the participants will be asked to solve a news reading and exploring task using the tool.

We are going to utilize information retrieval evaluation metrics (accuracy, precision, recall, F-measure) for assessing the interaction of the services behind the visualization tool. We are going to take into account several scenarios, by fixing one of the three diversity dimensions. One example would be: given an entity, were the aggregated summaries and identified sentiments relevant.

5.3 Usability and usefulness of prototypes in real-world scenarios and collection of best practices

The following are dimensions that we need to measure/observe:

How useful is this tool to online news readers? How does it compare to alternative tools, e.g. the current Google News interface? Which are the concrete use cases where its usefulness is maximized?

- What design choices can be done to improve usability and usefulness? Is the initial interface too complicated for the average user? Is it just a small evolution from current newsreader interfaces?
- How much depends on the quality of the output, and how much depends on the user interface? Is the interface easy enough but the summaries generated are either uninformative or ungrammatical? Do they faithfully convey the news story reported by the news agencies? Is diversity properly selected and highlighted?

Proposition:

- Regarding summarization quality, perform automatic evaluations independently of the usefulness of the tool, to make sure that quality is not a problem.
- Regarding the user interface, perform controlled experiments with a User Experience (UX) expert to finalize the first version. Next, identify a set of users interested in using the tool on a day-to-day basis, and perform usage studies based on activity logs to identify the most widely used controls in the tool, and iterate from there.

5.4 List of action items and milestones

Based on the strategy described in the previous section we list in this section the list of action items and milestones that we will be implemented in the 3rd year of the project:

- The initial use case prototype will be iterated upon, with the final, refined version available in M30.
- Based on user studies, a final user evaluation and news reader will be finalized in M36.

6 Conclusions

In this deliverable, we outlined for the diversity-enabling extensions to Web 2.0 platforms how we plan to evaluate them and, based on that evaluation, collect best practices concerning the development, implementation, introduction and usage of said tools. Most of the evaluation approaches employ end-user studies and focus on the end-user experience, not neglecting the developers' perspective in the process as well as that of the community liaison management in charge of establishing tools within their user community. These evaluation and best practice collection plans will be adhered to but will be adapted to the needs and circumstances dictated by the various tools and user communities they are connected to. As new features or components of specific tools emerge, they will be incorporated into the tool testing process. This could for example likely be the case for the Wikipedia tools that are not only rated and improved but also partly build by the community, as well as the Drupal tool. We are open to adapt and listen to these communities when it comes to not only adjusting the tools but also providing a realistic environment for best practice collection.

References

[1] Delort, Y., & Alfonseca, E. (2012). Description of the Google update summarizer at TAC-2011. Retrieved 09 19, 2012, from http://www.nist.gov: http://www.nist.gov/tac/publications/2011/participant.papers/GOOGLE.proceedings.pdf

[2] ISO/IEC. (1998). International Organization for Standardization 9241-11. Ergonomic requirements for office work with visual display terminals (VDT)s - Part 11 Guidance on usability.

Annex A

This annex contains the form, which will be used for getting feedback from potential adopters of RENDER technology at Telefónica. Besides an online version of this form is accessible at:

https://docs.google.com/spreadsheet/viewform?formkey=dFVSMmNGZ3FKeGZCNmlJNDg5dlc5RVE6MQ

The **FP7 RENDER project** is aimed to provide a comprehensive conceptual framework and technological infrastructure for enabling, supporting, managing and exploiting information diversity in Web-based environments (further details at http://render-project.eu/about-us/approach).

If you are reading this we assume you have installed a version of the Telefónica Opinion Mining Tool (T-OMT). If you would need support for the installation process or in order to solve any technical problem, please contact with Carlos Picazo at e.cpccp@tid.es.

The goal of T-OMT is to collect massive datasets from Twitter and use RENDER algorithms in order to give useful information about the current opinion of the product. Although this tool is using Twitter as a primary opinion source, comments about possible alternative source are always welcome. Currently T-OMT generates reports filtering the data by date, topics and language. These reports will show graphs about the mentions of the topics, sentiment overall, opinion geolocation, and related topics.

The evaluation process in which you are enrolled consists on collecting feedback from potential final users. So **we encourage you to use the tool as much as possible**. Next there is a form that you can use to report comments, proposals or any issue regarding the performance and the usefulness of the T-OMT software.

* Required fields

1. First name and last name *

2. How often do you use T-OMT?

Never	Less frequently than once per month	At least once per month	At least once per week	Daily
☐	☐	☐	☐	☐

3. When you use T-OMT, on average, how long do you spend using T-OMT?

Less than 5 minutes	5-10 minutes	10-30 minutes	About 1 hour	More than 1 hour
☐	☐	☐	☐	☐

4. When you use T-OMT, how often do you use the following features?

			Never	Rarely	Occasionally	Frequently
Mentions module			☐	☐	☐	☐
Sentiment analysis component			☐	☐	☐	☐
Cloud component			☐	☐	☐	☐
Map component			☐	☐	☐	☐

Detail view	☐	☐	☐	☐
Re-create old reports	☐	☐	☐	☐

5. How helpful do you find the following T-OMT features?

	Very unhelpful	Somewhat unhelpful	Indifferent	Somewhat helpful	Very helpful
Mentions module	☐	☐	☐	☐	☐
Sentiment analysis component	☐	☐	☐	☐	☐
Cloud component	☐	☐	☐	☐	☐
Map component	☐	☐	☐	☐	☐
Detail view	☐	☐	☐	☐	☐
Re-create old reports	☐	☐	☐	☐	☐

6. Mention situations where you have used T-OMT and for what purposes (e.g. to assess the impact of a product campaign)?

7. What extra features would you add to the T-OMT?

8. Do you have any additional feedback for us?

Thanks for your collaboration!

RENDER Team at TID