

MULTISENSOR

Mining and Understanding of multilingual content for Intelligent Sentiment
Enriched context and Social Oriented interpretation

FP7-610411

D7.7

Final System

Dissemination level:	Public
Contractual date of delivery:	Month 35, September 30 th , 2016
Actual date of delivery:	Month 35, September 30 th , 2016
Workpackage:	WP7 System Development and Integration
Task:	T7.3 Technical infrastructure T7.4 System development
Type:	Prototype
Approval Status:	Final Draft
Version:	1.0
Number of pages:	83
Filename:	D7.7_FinalSystem_2016-09-30_v1.0.pdf

Abstract

This document describes technical components and infrastructure for the Final System (FS) of the MULTISENSOR platform. It provides an overview of the system, organisation and composition of the components (modules), as well as demonstrates improvements with respect to the Second Prototype (SP) (D7.6). The FS combines improved features of the SP with final versions of the new services.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information

at its sole risk and liability.



co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	26/08/2016	Document initiation	A. Bilous (everis)
0.2	06/09/2016	EVERIS contributions	A. Bilous, E. Jamin (everis)
0.3	21/09/2016	Contributions from partners	A. Bilous, (everis) + ALL
0.5	23/09/2016	Integrated document	A. Bilous (everis)
0.6	27/09/2016	Internal review	D.Liparas (CERTH)
1.0	30/09/2016	Final version	A. Bilous (everis)

Author list

Organization	Name	Contact Information
EVERIS	Emmanuel Jamin	ejacques@everis.com
EVERIS	Andriy Bilous	andriy.bilous@everis.com
EVERIS	Víctor Sauri	victor.sauri.santacreu@everis.com
CERTH	Stefanos Vrochidis	stefanos@iti.gr
CERTH	Dimitris Liparas	dliparas@iti.gr
CERTH	Ilias Gialampoukidis	heliasgi@iti.gr
CERTH	Anastasia Moumtzidou	moumtzid@iti.gr
PRESSRELATIONS	Leszek Blacha	leszek.blacha@pressrelations.de
ONTO	Boyan Simeonov	boyan.simeonov@ontotext.com
ONTO	Vladimir Alexiev	vladimir.alexiev@ontotext.com
LinguaTec	Reinhard Busch	r.busch@linguatec.de
LinguaTec	Boris Vaisman	b.vaisman@linguatec.de
EURECAT	Ioannis Arapakis	arapakis@eurecat.com
UPF	Gerard Casamayor	gerard.casamayor@upf.edu

Executive Summary

The D7.7 of the MULTISENSOR platform provides a technical overview of the development and integration of the Final System (FS), explains the final version improvements with respect to the second prototype and the new services features, which took place during the development and the integration.

The objectives for the FS were:

- The integration of newly delivered and final version of existing MS services,
- The full data migration into the semantic repository (semantic integration),
- The deployment of the advanced multimodal indexing and retrieval engine, and
- The optimisation of the three Use Cases (UCs) applications to be aligned with new functionalities and user's needs.

The deliverable summarises the status of all the services that are integrated in the final system for 3 demonstrators (one for each use case). The following URLs can be used to access MULTISENSOR Final System:

UC1 Application: <http://grinder1.multisensorproject.eu/uc1/>

UC2 Application: <http://grinder1.multisensorproject.eu/uc2/>

UC3 Application: <http://grinder1.multisensorproject.eu/uc3/>

Abbreviations and Acronyms

CI	Continuous Integration
CMR	Central Multimedia Repository
CNR	Central News Repository
DB	DataBase
EBS	Elastic Block Storage
EC2	Elastic Compute Cloud
ECU	Elastic Compute Unit
FP	First Prototype
FS	Final System
FTP	File Transfer Protocol
FTS	Full-Text Search
HTTP	HyperText Transfer Protocol
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
MPEG	Moving Picture Experts Group
NER	Named Entity Recognition
OPS	OPERationS repository
PPA	Personal Package Archive
OWL	Ontology Web Language
RDF	Resource Definition Framework
REST	Representational State Transfer
SIMMO	Socially Interconnected and MultiMedia-enriched Object
SOA	Service Oriented Architecture
SP	Second Prototype
SPARQL	SPARQL Protocol And RDF Query Language
UC	Use Case
UC(x)	Use Cases (1, 2 or 3)
UCS	Universal Character Set
UI	User Interface
UTF	UCS Transformation Format
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

Table of Contents

1	INTRODUCTION	8
2	FINAL SYSTEM ARCHITECTURE	10
2.1	Global view	10
2.2	Status of the previous Prototype (Second Prototype)	11
2.3	Objectives of the Final System	11
2.4	Status of the Final System	11
3	INTEGRATION FRAMEWORK	16
3.1	Approach.....	16
3.2	High-level view	16
3.3	Offline modality	17
3.3.1	Crawlers	17
3.3.1.1	Media collector (PR crawler)	17
3.3.1.2	Twitter collector	17
3.3.2	Repositories	18
3.3.2.1	RDF Repository	18
3.3.2.2	Central News Repository (CNR/CMR).....	18
3.3.2.3	OPS Repository	18
3.3.2.4	MongoDB repository for CERTH services	19
3.3.2.5	SOLR UPF services.....	19
3.3.3	Content extraction pipeline (CEP).....	19
3.3.3.1	Language detection	19
3.3.3.2	Translation.....	20
3.3.3.3	Named Entities recognition	21
3.3.3.4	Entity Linking service	22
3.3.3.5	Concept extraction	22
3.3.3.6	Dependency parsing	23
3.3.3.7	Relation extraction	24
3.3.3.8	Polarity and sentiment extraction.....	24
3.3.3.9	Extractive summary and query-based extractive summarisation	26
3.3.3.10	Classification.....	26
3.3.3.11	Context extraction.....	27
3.3.3.12	Audio extraction and ASR.....	29
3.3.3.13	Concept and Event detection	29
3.3.3.14	Indexing (Simmo Mongo Storing).....	31
3.3.3.15	RDF Validation	32
3.3.3.16	Storing RDF	33
3.3.4	Content Alignment Pipeline (CAP)	33
3.3.5	Social Media Analysis Pipeline (SMAP)	35

3.3.6	Platform Security	37
3.3.7	Platform monitoring	38
3.3.8	Platform testing services.....	39
3.4	Online modality.....	40
3.4.1	Business Shared Services	40
3.4.1.1	Content delivery	40
3.4.1.2	Semantic search	42
3.4.1.3	Topic-Event detection	44
3.4.1.4	Similarity search	45
3.4.1.5	Machine Translation.....	46
3.4.1.6	Abstractive summary.....	46
3.4.1.7	Hybrid Search	47
3.4.1.8	Contributor analysis	48
3.4.2	Other Online Services	49
3.4.2.1	User profile.....	49
3.4.2.2	Reference Data.....	50
3.4.2.3	Decision support.....	52
4	PROTOTYPE APPLICATIONS	54
4.1	UC1: Journalism Use Case	54
4.2	UC2: Media Monitoring Use Case.....	60
4.3	UC3: SME internationalisation Use Case.....	69
5	CODE ORGANISATION	77
5.1	Source tree layout (D7.4 updates).....	77
5.2	Continuous integration environment	78
5.3	Packaging	78
5.3.1	Java modules.....	78
5.3.2	Node.js modules	79
6	INFRASTRUCTURE.....	80
6.1	Current farm (D7.6 updates)	80
7	DEMONSTRATOR URLs AND INFORMATION	82
8	SUMMARY AND CONCLUSIONS	83

1 INTRODUCTION

In D7.1 (Technical roadmap), a general roadmap and technical vision for the implementation of the MULTISENSOR platform was established. The user and non-functional requirements in D8.2 and the technical vision were combined in D7.2 (Technical requirements and architecture design) to define the global architecture of the system and its subsystems, workflows and interfaces. In D7.3 (Operational prototype), the user interface and most of the services were implemented as a dummy version. In D7.4 (First prototype), most of the services were implemented and integrated as a basic version. In D7.6 (Second prototype), most of the services were provided and integrated as an advanced version and new functionalities were integrated in their basic versions. Finally, in D7.7 (Final system), all the services are integrated in their advanced version.

The “walking skeleton” for the technical roadmap laid out in D7.1 is presented below:

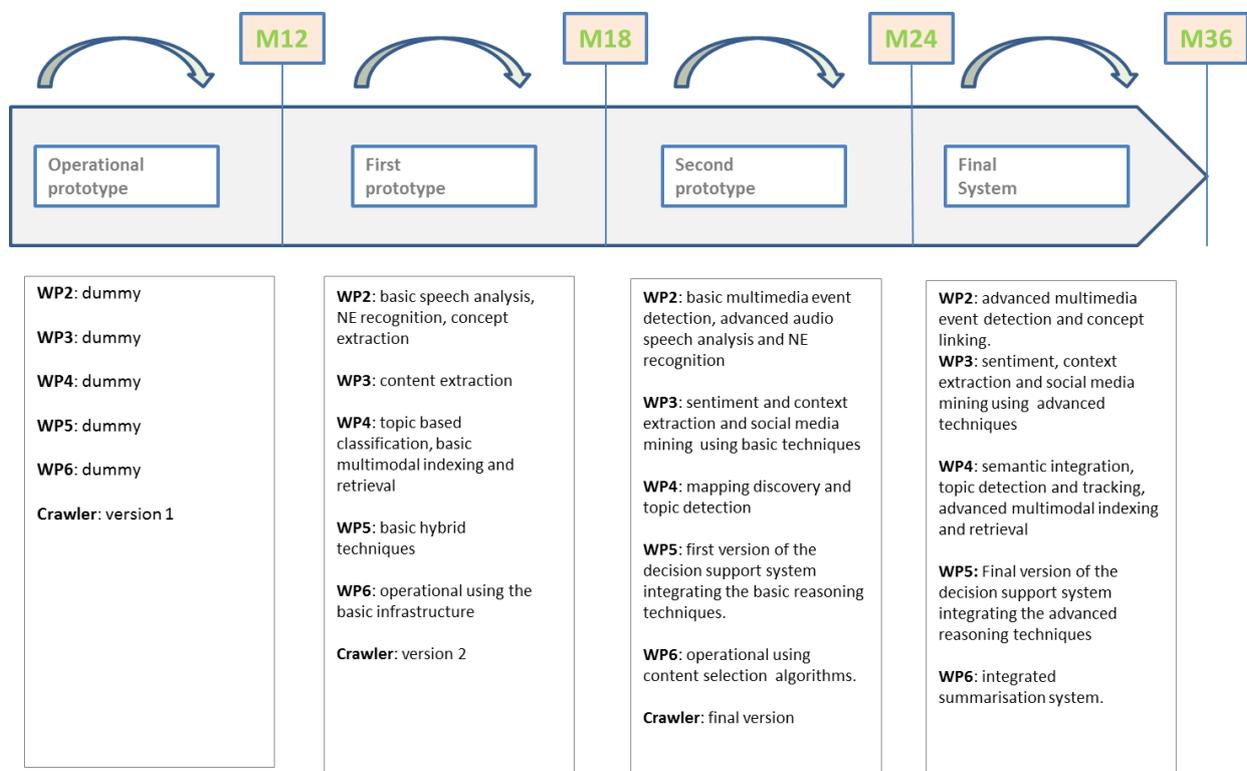


Figure 1: Technical roadmap

The purpose of this document is to provide a technical reference for the D7.7, which is the fifth technical milestone (MS5) of the project (M36). D7.7 contains the following sections:

- **Section 2** contains a high-level technical overview of the Final System (FS).
- **Section 3** contains a description of the integration status of the framework.
- **Section 4** contains a description of the online applications UC(x).
- **Section 5** contains the code organisation of the MULTISENSOR project.
- **Section 6** details the technical infrastructure hosting the FS.
- **Section 7** contains links and details for accessing the demonstrator application for reviewers.

- **Section 8** presents a brief summary and conclusions.

2 FINAL SYSTEM ARCHITECTURE

2.1 Global view

The global architecture for the MULTISENSOR platform has been discussed in details in D7.1, D7.2, D7.4 and D7.6.

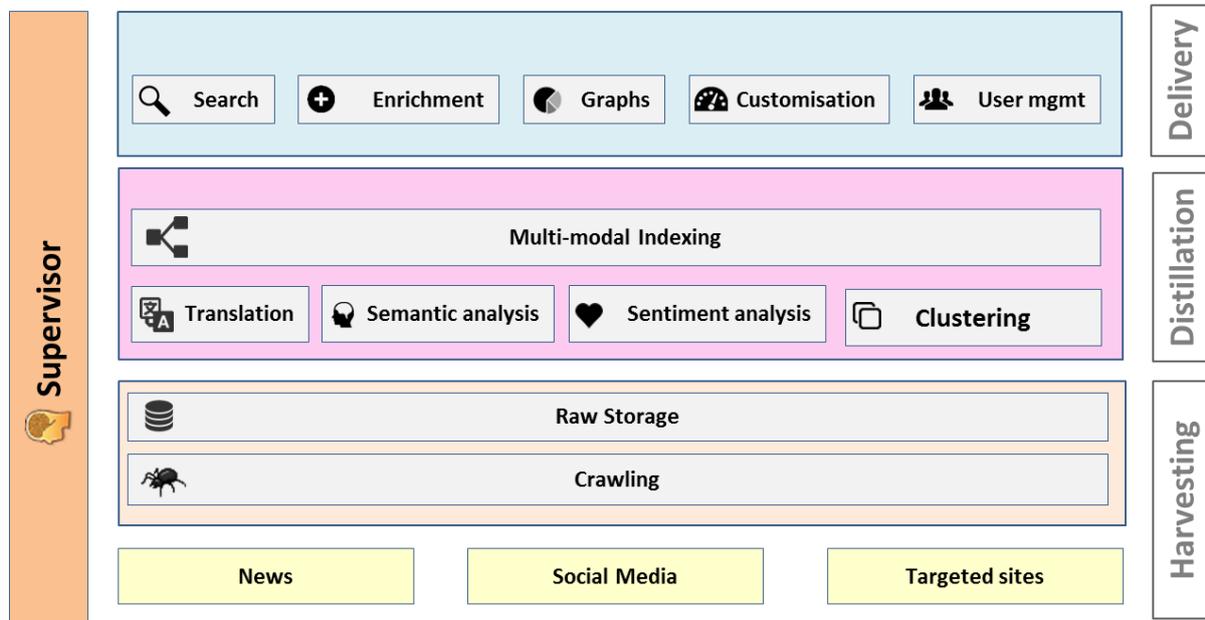


Figure 2: MULTISENSOR logical architecture

The MULTISENSOR architecture (see Figure 2) is based on a SOA approach and encompasses two discrete modalities: offline - asynchronous processing of harvested data (see D7.2, section 4.2.2), and online - synchronous retrieval, delivery and exploitation of the analytical data (see D7.2, section 4.2.3).

D7.7 explains the current status of the system in terms of repositories, services, processes and workflows of the Final System.

The most important efforts were focused on the improvements of the offline modality and the refactoring of the UCx applications. At first, on the offline modality, the CEP was consolidated with the integration of all the services into an advanced version and with the deployment of the multilingual CEP to process the articles in the five considered languages. For the UC applications, the user interfaces were largely improved to integrate the advanced functionalities (such as the hybrid search and the display of the most relevant information such named entities and the concepts). In addition, the third UC application for the SME internationalisation was redesigned to display the final version of the decision support.

2.2 Status of the previous Prototype (Second Prototype)

In D7.6, the status of the Second Prototype was explained. A quick summary of the elements that were part of it is provided below:

- The system was hosted on a powerful cloud infrastructure in order to process with the CEP the maximum quantity of articles. The server also hosts two of the three UC applications.
- The four repositories of the Data Layer were in place.
- The crawlers were implemented and deployed.
- All the main services were implemented. Only a few of them were provided in a baseline version.
- The different services were integrated into the platform, i.e. they can interact between themselves, store and collect data from the different repositories.
- The RDF repository was populated with the extracted knowledge produced by the CEP.
- The three Use Cases applications were improved with the interaction of the available online services and the re-design of the user interfaces (mostly in UC3).

2.3 Objectives of the Final System

The Final System considers the delivery of all services (Online and Offline) and the three UC applications in their final versions. In addition, the knowledge base needs to be populated with a reasonable amount of multilingual, textual, social and multimedia SIMMOs.

The objectives of the FS are:

- The implementation of the secure operational architecture, including individual modules running on partner premises;
- The integration of the advanced multimedia event detection and concept linking for the content extraction;
- The integration of the sentiment, context extraction and social media mining using advanced techniques for user-centric content extraction;
- The integration of the semantic integration, topic detection and tracking, advanced multimodal indexing and retrieval for content integration and retrieval;
- The integration of the advanced reasoning techniques for reasoner and decision support;
- The integration of the information production: integrated advanced summarisation system, and
- The delivery of the UC applications with an optimised UX.

2.4 Status of the Final System

The Final System (FS) represents complete integration of all services and specific components like multilingual, multimedia, repositories and use cases.

According to the T7.3 technical infrastructure progress status:

- The cloud infrastructure was optimised. For this, three important aspects have been improved: the security framework, the monitoring tool and the CEP testing tool.
- The server for the Knowledge base (GraphDB) was populated according to the plan. It currently includes multilingual textual, social and multimedia SIMMOs. Moreover,

Ontotext provided components for the Bulgarian pipeline processing (rule-based module for morphological tagging, mate tools for POS tagging and dependency parsing, lemmatisation and ontology mapping) and flexible restful API for services that have replaced SPARQL queries and the DB querying.

- The SMAP server was provided by CERTH and it is functioning as a remote endpoint for social services, like Influential user detection and Community detection.

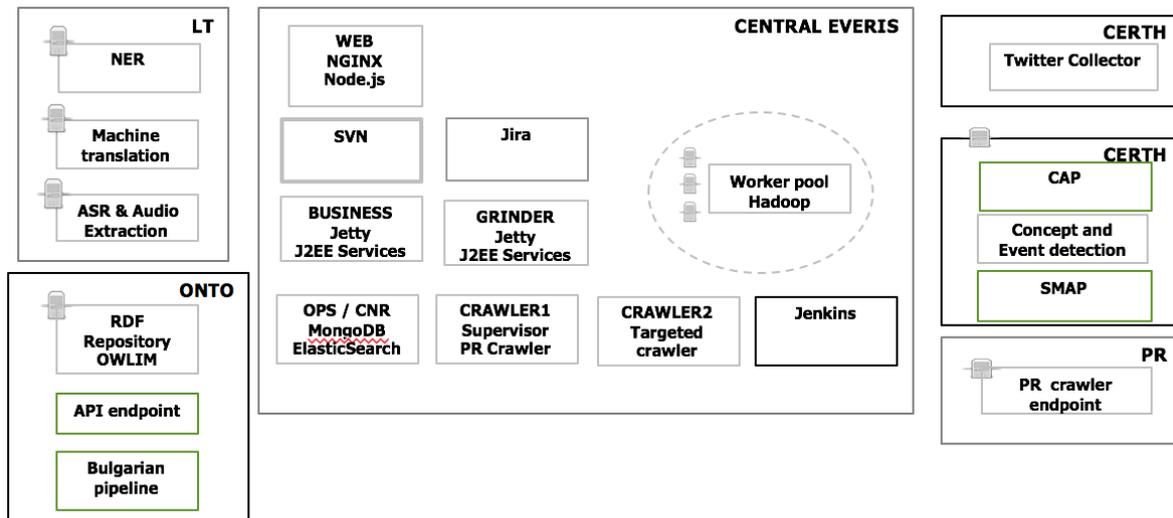


Figure 3: FS – T7.3 Technical infrastructure

We can underline that the FS has multilingual CEP with several services that were not available in SP and advanced services in comparison to SP.

Among them, NER for Bulgarian language, entity linking, entity alignment, Simmo mongo storing and multimedia services like ASR, concept and event detection.

Below, there is a list of the offline services with a description of the modifications that were applied during the Final system phase. First, the modifications related to the CEP services are explained:

- **NER:** The support for the Bulgarian language was developed and integrated;
- **Dependency Parsing:** The instance per language was successfully deployed;
- **Relation Extraction:** The performance of this service was improved, and some multilingual improvements were applied as well;
- **Concept Extraction:** The final version on specific and generic concepts was delivered and was able to target all the UCs topics;
- **Sentiment Analysis:** The score and performance have been improved;
- **Context Extraction:** Advanced contextual features (fluency, richness, technicality) are being generated (numeric score);
- **Extractive summarisation:** The quality was improved and customised according to the text length;
- **Categorisation:** The performance was improved and Word2vec was added as a new modality for the service.

In addition, some new services have been added to the CEP to improve the quality of the results:

- **Entity Linking:** Babelfy was requested in a multilingual mode to obtain the recognised concepts and entities;
- **Entity Alignment:** The service permits to remove the redundant entities annotated between NER and Entity Linking;
- **Simmo Mongo Storing:** The processed SIMMOs (only in English version) are also stored in mongoDB instance, which is hosted on CERTH server.

Finally, the multimedia services of the CEP have also been improved:

- **Concept and event detection:** RDF support has been added;
- **ASR:** The service supports German and English video transcription;
- **Node js** modules were developed and improved to periodically retrieve multimedia articles and their assets (video, images), to be processed in the multimedia pipeline, multimedia representation (concepts – keyword cloud, ASR – video subtitles) in UC1.

All these services are described in more details in Section 3.3.3.

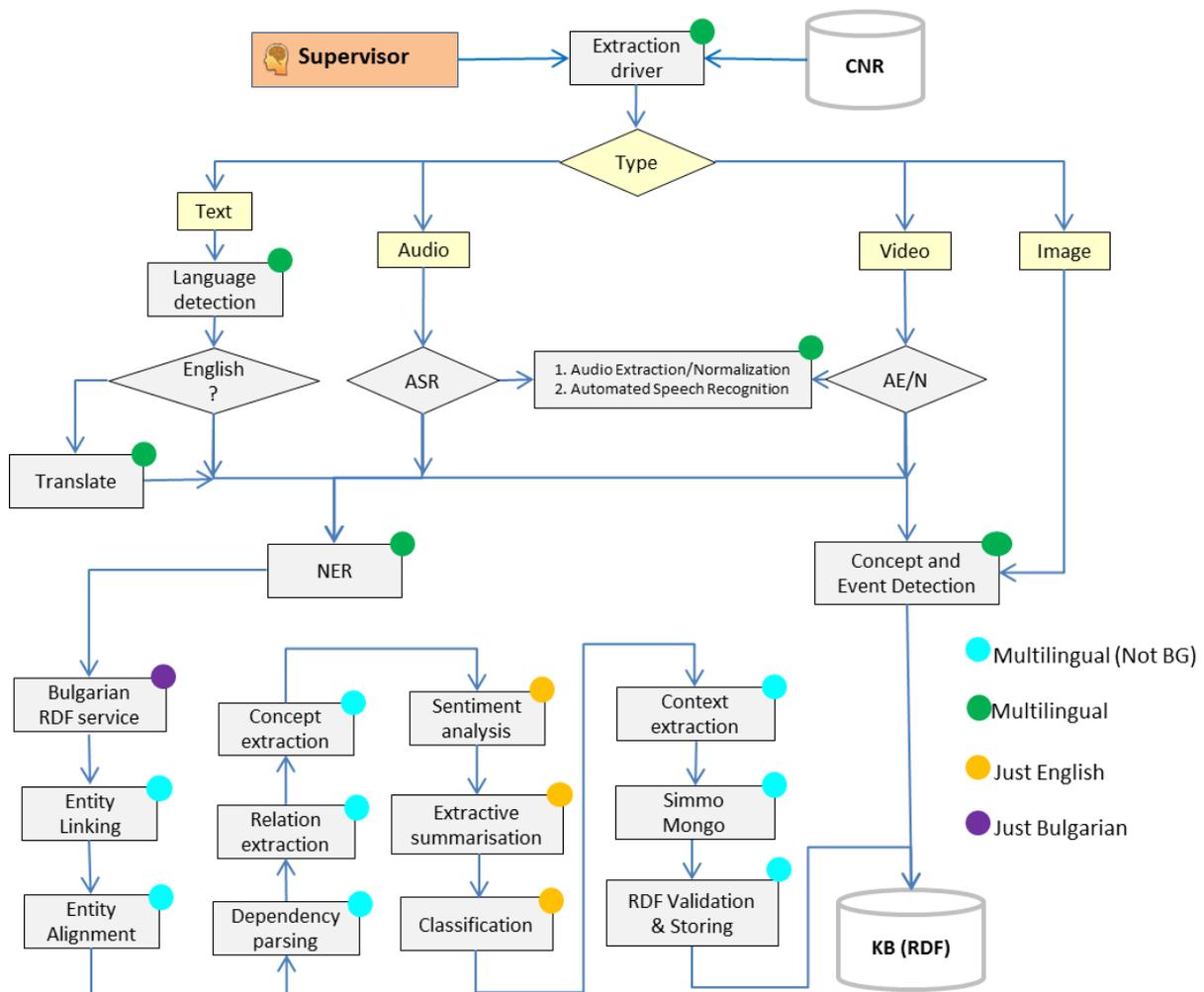


Figure 4: FS, CEP – T7.4 System development

For the Online modality, the modifications applied on the services are the following ones:

- **Content delivery:**
 - Extra methods have been inserted to obtain RDF information (specifically for UC2);
 - The welcome view has been created;
 - A new method to retrieve the specific and generic concepts per list of articles was developed.
- **Community Detection:** The service was improved, deployed and integrated into UC2 and UC3.
- **Most influential users:** The service was improved, deployed and integrated into UC2 and UC3.
- **Topic an Event detection:** The service was improved, deployed and integrated into UC1 analyst view.
- **Similarity Service:** The service was improved, deployed and integrated into UC1.
- **Decision Support:** More indicators were added and the user interface was partly re-designed.
- **Summarisation Extractive:** The following methods were implemented:
 - Single document
 - Multi document
 - Keyword base summary
- **Hybrid Search:** The new service was developed and integrated into UC1.
- **Semantic Search:** The service was improved, deployed and integrated into UC1 and UC3.

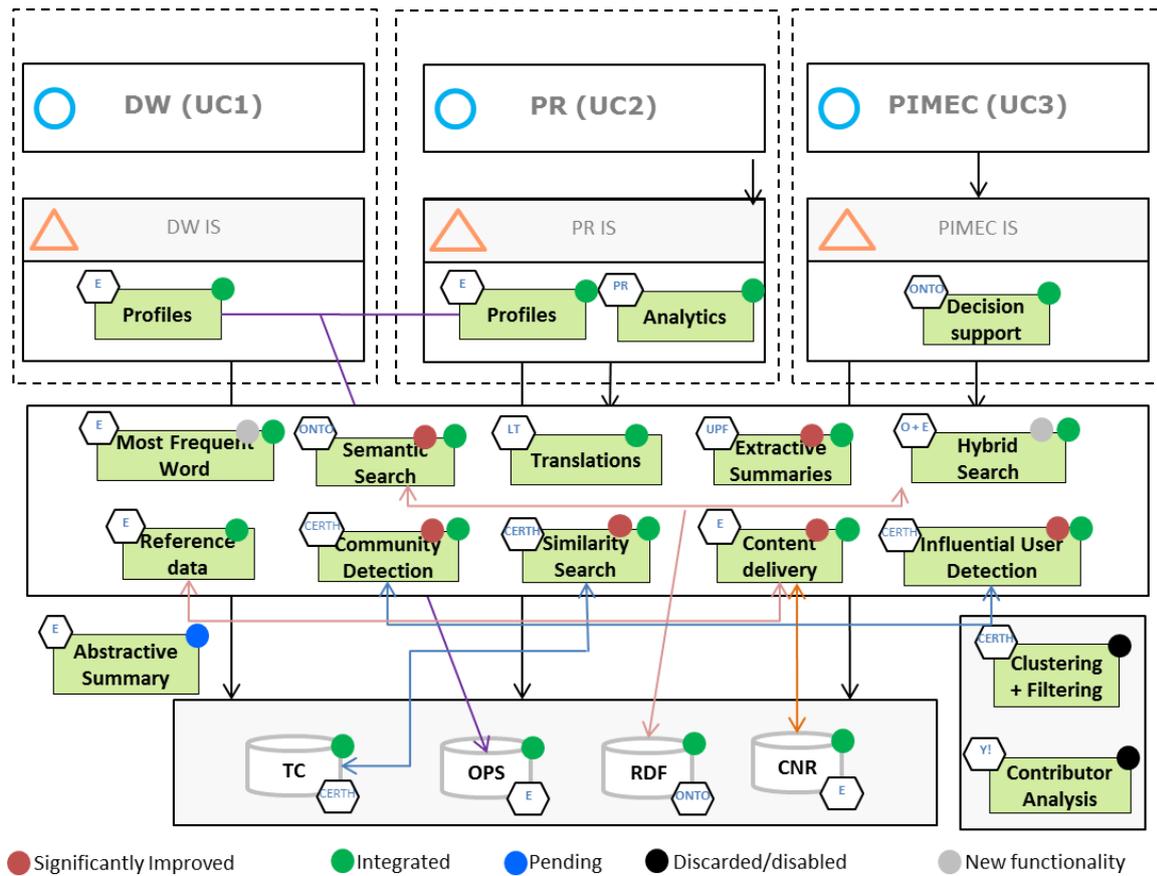


Figure 5: FS, Online modality, T7.4 System development

All these services are described in more details in Section 3.4.

3 INTEGRATION FRAMEWORK

3.1 Approach

For the Final System, all the services and the repositories are deployed, integrated and fully functioning. In this version, all the services are provided in their advanced versions.

In the Second Prototype, the development infrastructure was significantly improved by automating processes such as testing, compilation, execution and deployment of the services. For the Final System, the development infrastructure did not require any improvements. It was used as it was provided during the Second Prototype.

3.2 High-level view

The logical layers of the MS architecture are represented in Figure 6. In the architecture, only the Twitter Collector (TC) was inserted in order to cover the necessity to process a huge quantity of tweets without the dependency of the Twitter API.

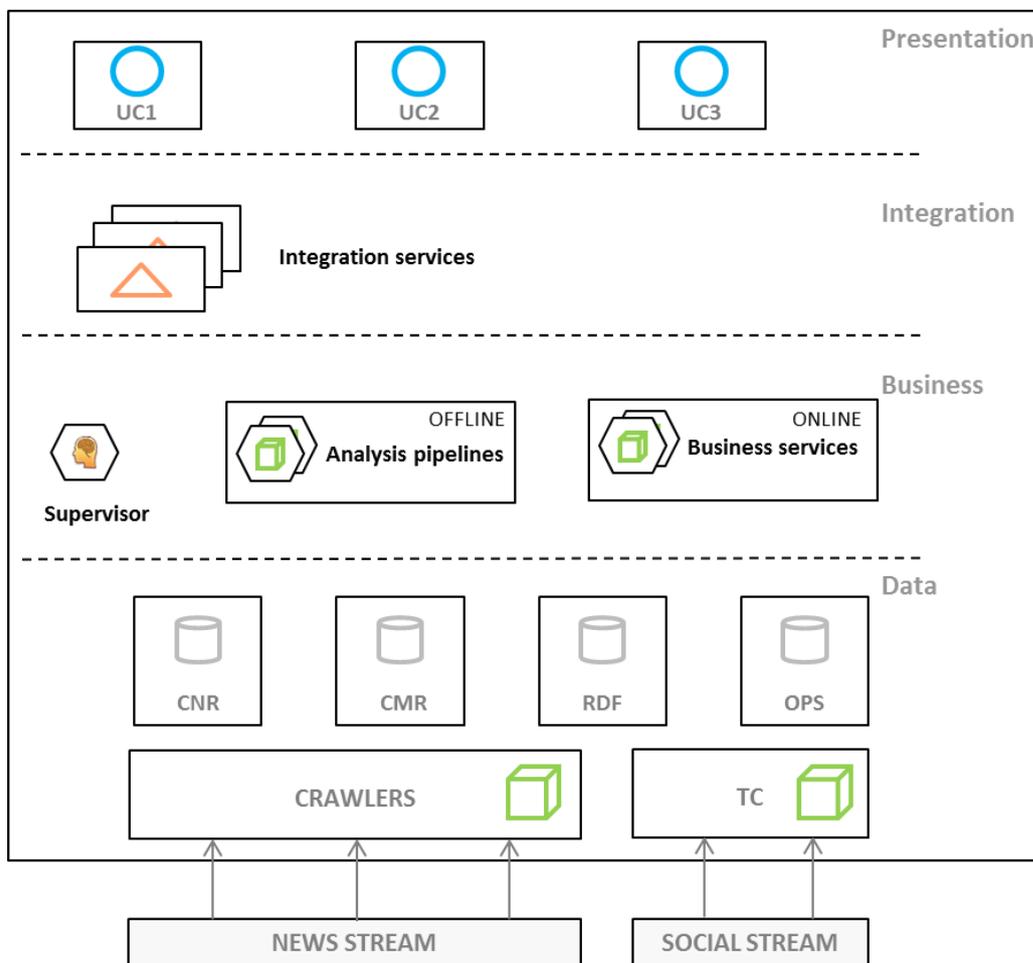


Figure 6: Final System MULTISENSOR high-level view

The Final System developments were focused on improving and integrating the Offline and Online services, improving the crawling content quality, extending the crawling process on the multimedia dimension and continuing the intensive population of the KB repository with analytic data.

3.3 Offline modality

3.3.1 Crawlers

The previous Prototype contained crawlers provided by PR and EURECAT, implemented and integrated. Crawled data provided by PR, CERTH and EURECAT crawlers as a JSON endpoint, have been cleaned (for example, to remove duplicated data), indexed and stored into RDF knowledge base provided by Ontotext.

During the Final system, the situation of the crawlers did not change. Only more work was done to ensure a reasonable distribution between the multilingual, multimedia and social media that was crawled. Data from the abovementioned crawlers are being stored into the CNR.

3.3.1.1 Media collector (PR crawler)

As described in D7.4 and D7.6, PR's proprietary crawling technologies aggregate, index and extract content from international news websites. Within the MULTISENSOR project, PR provides news articles for each use case based on specific keywords that have been defined by the user partners.

All information is accessible via an API, which has been described in D7.4 in detail.

During the project's lifetime, crawling functionalities have been extended to also extract links to images, videos and audio files. These links are provided via the PR API, in addition to the textual content from crawled websites.

No updates since the SP were made on PR's crawler.

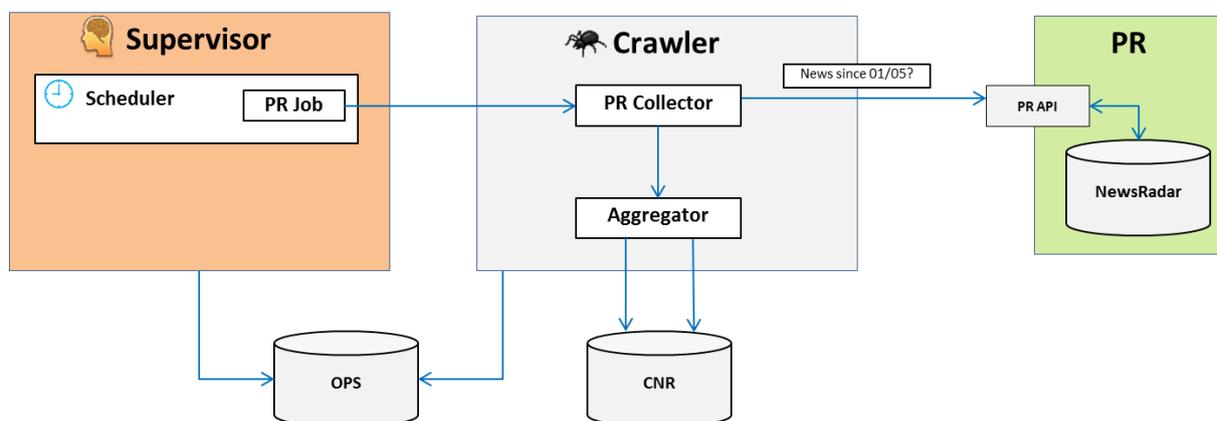


Figure 7: Final System MULTISENSOR crawler flow

3.3.1.2 Twitter collector

The Twitter crawler comes from the previous SocialSensor (FP7-287975) project, in which Stream Manager was developed. Stream Manager contains a number of APIs that collect incoming content relevant to a keyword, a user or a location from a set of social streams (Twitter, Facebook, Instagram, etc.). The Twitter crawler specifically gathers Twitter posts for a set of hashtags, which are pre-specified for each Use Case separately. These posts, as well as information regarding the author and the associations found within the posts are then stored into a MongoDB database. The Twitter crawler runs every 30 minutes for the hashtags of each Use Case and if new posts are found, they are stored inside the database.

The current total number of posts stored into MongoDB is approximately 43K posts for UC2 and 1,7M posts for UC3 hashtags (there are separate sets of hashtags for the different sectors and products considered in UC3, but only one set of hashtags has been specified for the whole of UC2). Finally, the Twitter posts gathered by the Twitter crawler and stored into MongoDB are fed as input to the services (Influential User Detection and Community Detection) of the Social Media Analysis Pipeline (SMAP).

3.3.2 Repositories

The FS uses the following repositories:

- The RDF repository to store all the KB produced by the offline modality and some relevant Linked Open Data datasets, provided by Ontotext;
- The CNR to store all the crawled documents;
- The OPS to store the Operations information like the User Profile information;
- Mongo repository for CERTH services, and
- SOLR for some UPF services (Concept Detection and Abstractive Summary).

3.3.2.1 RDF Repository

The RDF repository holds all MULTISENSOR data in semantic format (RDF, RDFS and OWL). This includes the ontologies, external datasets like DBpedia, Geonames and many statistical indicators from World Bank and Eurostat. The main function of the triplestore is to store the SIMMO objects. This data forms the MULTISENSOR knowledge about the world. It is based on GraphDB-Enterprise which is a high-performance and clustered semantic repository created by Ontotext.

For the First Prototype, GraphDB-SE was just used as a knowledge repository, but after that decision was made, we wanted to take advantage of the search engine functionality too. This implied the migration to GraphDB-Enterprise, which is an advanced version of GraphDB-SE distribution.

In the semantic repository, more than 125,000 SIMMOs are stored. This represents a total of 4,000,000 RDF triples.

3.3.2.2 Central News Repository (CNR/CMR)

The Central News Repository (see D7.2, section 4.2.4.1) is the raw storage dump for the Crawlers (Site and Media collectors). The CNR is implemented as an Elasticsearch instance, allowing storing “big data” without degradations in performance.

No updates since the SP were made on CNR. The CNR repository contains more than 12,000,000 multimedia, multilingual or social media items.

3.3.2.3 OPS Repository

The Operations Repository (OPS) provides fast, read/write structured data storage for any systems in the platform that require it (see D7.2, section 4.2.4.4).

The OPS is implemented as an instance of Mongo DB only for user related management in all UCx. It permits to store the information like user credentials to login the web application, the user session, profiles preferences and other information related to user management.

No updates since the SP were made on OPS.

3.3.2.4 MongoDB repository for CERTH services

MongoDB repository was introduced in the FS in order to optimise performance in terms of execution time on CERTH services like Similarity search, Category classification, etc. It contains information of all articles stored in GraphDB. Information from multimedia (text, image, video) existing in these articles is also included in MongoDB. It is hosted in the CERTH server and it is accessible only through CERTH web services.

3.3.2.5 SOLR UPF services

A Solr instance was introduced in FS in order to optimise performance in terms of UPF services, like Concept Extraction Service. The Solr instance was integrated into the MS main server (grinder) within the production environment. So the Solr indexes are enriched every time a pipeline execution is performed.

Another separated Solr instance is used for Abstractive Summarisation, however it is hosted in UPF premises.

3.3.3 Content extraction pipeline (CEP)

The improvements of the CEP were already introduced in Section 2.4 “Status of the Final System” and Figure 4. In the following sub-sections, the description and improvements for each service are provided.

3.3.3.1 Language detection

The language detection component is used to identify the language of each crawled article for the future processing in the full system. The identified language is very important for further processing, because all modules are language dependent. Therefore, this module has very high expectation to the accuracy of its results.

The language detection is deployed on the LT server called Language Identifier Server (LIS) and is remotely called by the MS platform.

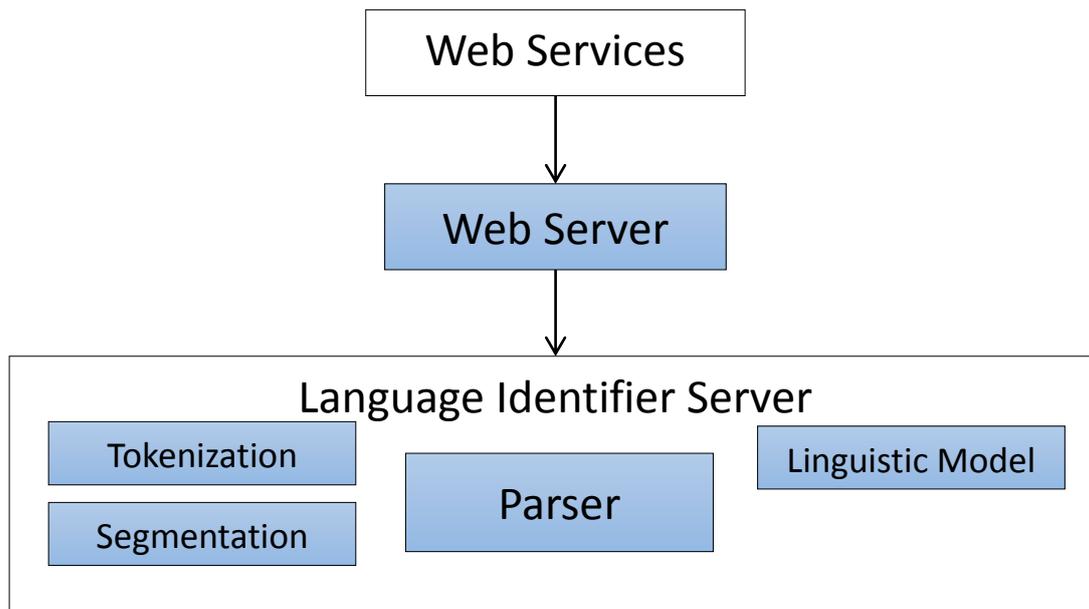


Figure 8: LIS simplified architecture

During the project, the Language Identifier module was developed and improved to the real time server based and high performance architecture.

Service name	Language Identifier
SVN artefact / remote endpoint	http://services.linguec.org/rest/lang/identify
Functional Description	Recognises language of a text
Deployment status	Remotely deployed
P1 version	Version 1.0, fully functional
P2 version	Version 1.1, fully functional
P2 improvements	Extended to 18 additional languages for better discrimination purposes
FS version	Version 1.2, fully functional
FS improvements	Stable performance has been reached
Maximum amount of queries per second	7 queries per second
Supported languages	MULTSENSOR languages: English (en), French (fr), German (de), Spanish (es) and Bulgarian (bg). Plus 18 additional languages: Arabic (ar), Chinese (zh), Czech (cz), Danish (da), Dutch (nl), Finnish (fi), Greek (el), Hebrew (he), Italian (it), Japanese (ja), Korean (ko), Norwegian (no), Pashto (ps), Polish (pl), Portuguese (pt), Russian (ru), Swedish (sv), and Turkish (tr)
Known issues	N/A
Conclusion/deviation	N/A

Table 1: Integration description of the Language identifier service

3.3.3.2 Translation

The Translation service receives as input text in source language and translates it into specified target language. A more detailed description of the Machine Translation Module is provided in D2.3 “Advanced techniques for text analysis, machine translation and concept extraction”.

Service name	Machine Translation
SVN artefact	http://services.linguec.org/rest/lang/translate
Functional Description	Receives as input text in source language and translates it into specified target language.
Deployment status	Remotely deployed
P1 version	Version 0.8, fully functional
P2 version	Version 1.0, fully functional

P2 improvements	Improved translation quality resulting from: a) better homogenisation of training corpus b) reduction of the amounts of unknown words c) tuning of model parameters
FS version	Version 1.1, fully functional
FS improvements	<ul style="list-style-type: none"> Improved performance and stability of online services. Optimizing sentence segmentation according to new crawler results. Developed automatic monitoring and restarting of online services.
Maximum amount of queries per second	7 queries (sentences) per second
Supported languages	English to/from French; English to/from German; English to/from Spanish; English to/from Bulgarian.
Known issues	Continued testing
Conclusion/deviation	N/A

Table 2: Integration description of the Translation service

3.3.3.3 Named Entities recognition

Named Entities recognition is a service that recognises the names of persons, the locations, and the organisations & companies, as well as date, amount and measurements. A more detailed description of the Named Entity Recognition Module is provided in D2.3 “Advanced techniques for text analysis, machine translation and concept extraction”.

Service name	Named Entity Recognition
SVN artefact	http://services.linguattec.org/rest/ner/recognize
Functional description	Identifies names (named entities), i.e.: persons, locations, organisations & companies, date, amounts, measurements
Deployment status	Remotely deployed
P1 version	Version 0.8, fully functional
P2 version	Version 0.9, fully functional, except Bulgarian language
P2 improvements	<ul style="list-style-type: none"> Adaptation of lexicons towards the three use case scenarios. Extensions of grammars to a more context-sensitive approach. Output given in the CEP workbook format.
FS version	Version 1.0, fully functional
FS improvements	<ul style="list-style-type: none"> Improved performance and stability of online services. Optimised sentence segmentation according to new crawler results.

Maximum amount of queries per second	7 queries per second
Supported languages	English, German, Spanish, Bulgarian
Known issues	N/A
Conclusion/deviation	N/A

Table 3: Integration description of the Name Entities Recognition service

3.3.3.4 Entity Linking service

Babelfy is used to obtain disambiguated references to BabelNet senses, covering both NEs and concepts. References to concepts in BabelNet are annotated as generic to differentiate them from the UC-specific concept annotations produced by the concept extraction service.

Service name	Entity linking service
SVN artefact	ms-svc-el
Functional description	Annotates texts with mentions of NEs and concepts in BabelNet
Deployment status	Deployed
P1 version	-
P2 version	-
P2 improvements	-
FS version	Version 1.0, fully functional
FS improvements	-
Maximum amount of queries per second	Unknown
Supported languages	Bulgarian, English, French, German and Spanish
Known issues	N/A
Conclusion/deviation	Babelfy is a third party service free for research purposes, but that might change in the future.

Table 4: Integration description of the Entity Linking service

3.3.3.5 Concept extraction

Concept extraction annotates mentions to concepts relevant to each UC. These concepts are often referred to by terminological expressions, which are used with a special meaning within the domain of the UC. Concept annotations are marked as specific to differentiate them with annotations produced by the EL service

Service name	Concept extraction service
SVN artefact	ms-svc-extr
Functional description	Annotates texts with mentions of UC-specific concepts
Deployment status	Deployed
P1 version	Version 0.2, baseline version

P2 version	Version 0.3, baseline version
P2 improvements	Term detection has been implemented using terminology-lists obtained by applying the TermRaider GATE plugin. NIF RDF annotations validated by Ontotext.
FS version	Version 1.0, fully functional version
FS improvements	Concept extraction has been rewritten from scratch and is based on a Solr index for each UC. Counts in the indexes are updated after every document is processed by the service.
Maximum amount of queries per second	The performance of the service is fast.
Supported languages	English, French, German and Spanish.
Known issues	N/A
Conclusion/deviation	Bulgarian is not supported.

Table 5: Integration description of the Concept extraction service

3.3.3.6 Dependency parsing

The dependency parsing service annotates texts with the syntactic structure of their sentences. Two different structures are annotated, (i) a surface syntactic structure indicating language-specific grammatical relations between all words in a sentence, and a (ii) deep syntactic structure with language-independent predicate-argument relations between content words. Both structures follow the dependency syntax formalism.

Service name	Dependency parsing service
SVN artefact	ms-svc-dep
Functional description	Annotates texts with syntactic parses of their sentences
Deployment status	Deployed
P1 version	Version 1.0, fully functional (for English language)
P2 version	Version 2.0, fully functional (for English and Spanish languages)
P2 improvements	Added support for Spanish language. NIF RDF annotations validated by Ontotext.
FS version	Version 3.0, fully functional (for English, French, German and Spanish)
FS improvements	Added support for French and German.
Maximum amount of queries per second	Although the service is pretty fast, no more than a few sentences can be processed per second.
Supported languages	English, French, German, Spanish.
Known issues	N/A
Conclusion/deviation	Bulgarian is not supported.

Table 6: Integration description of the Dependency Parsing service

3.3.3.7 Relation extraction

The relation extraction finds and annotates n-ary relations between entities, concepts and other relations in the text. The service associates linguistic predicates with semantic types from FrameNet a lexical repository of relational meanings (i.e. FrameNet)

Service name	Relation extraction service
SVN artefact	ms-svc-rel
Functional description	Annotates texts with n-ary semantic relations.
Deployment status	Deployed
P1 version	Version 0.2, baseline version
P2 version	Version 0.5, partly functional version
P2 improvements	Replaced the third-party library, Semafor, with a much faster deterministic tool developed by UPF. NIF RDF annotations validated by Ontotext.
FS version	Version 1.0, fully functional version
FS improvements	Added support for French, German and Spanish. Removed coreference resolution as it wasn't performing adequately and its output wasn't used by other MULTISENSOR services.
Maximum amount of queries per second	The performance of the service is fast.
Supported languages	English, French, German and Spanish.
Known issues	The lack of disambiguation affects the quality of the output.
Conclusion/deviation	No disambiguation against FrameNet has been implemented. No support for Bulgarian.

Table 7: Integration description of the Relation extraction service

3.3.3.8 Polarity and sentiment extraction

The sentiment analysis service detects sentiment in English text. More specifically, it improves existing sentiment analysis algorithms for the social web, as well as models a robust opinion mining system, by applying a linguistic analysis that is applicable to large datasets and considers the interdependencies observed between expressions and opinion holders in different sentences. The implemented service is using a machine-learned, domain-specific classifier that has been trained using syntactical features, such as standard BoW (e.g., uni-grams, bi-grams, tri-grams) and shallow (Shallow Kincaid, Coleman-Liau) features, extracted from an annotated, in-domain news corpus. The sentiment is extracted from both the body- and sentence-level of the news articles. This machine-based classifier provides for a piece of text two sentiment scores, a positive score in the range from 1.0 to 5.0 and a negative score in the range from -1.0 to -5.0. Based on those, it then derives the following sentiment features:

- Negative polarity value: negative sentiment expressed in the news article.
- Positive polarity value: positive sentiment expressed in the news article.

- Sentimentality¹: $|score_{pos}| + |score_{neg}| - 2 = score_{sent}$
- Polarity: $score_{pos} + score_{neg} = score_{pol}$
- Minimum polarity score (Pol_{min})
- Maximum polarity score (Pol_{max})

Improvements were made to the module to receive an object in the RDF format and return the updated object with the sentiment information in RDF format.

Service name	Sentiment Analysis
SVN artefact	ms-svc-sa
Functional description	This service aims to provide an analysis of the sentiment that is expressed in a news article. The extracted sentiment is given by SentiStregth sentiment lexicon.
Deployment status	Deployed
P1 version	Version 0.2, baseline version
P2 version	Version 1.0, fully functional extracting the sentiment features for a given news article
P2 improvements	Developments for additional features, such as negative and positive polarity. Validation of the results at sentence and full news article level. The received and returned object is RDF (before it was implemented for a JSON object).
FS version	Version 1.1, fully functional extracting the advanced sentiment features for a given news article.
FS improvements	Current version does not rely on a dictionary-based sentiment extraction approach; instead, it uses a machine-learned classifier with superior performance. A validation of the results at sentence and full news article level has been performed. The received and returned object is RDF.
Maximum amount of queries per second	Unknown
Supported languages	English (additional languages can be supported given the availability of machine translation)
Known issues	Requires NIF wrapper class to access the RDF part of the article object (full text of the news article or sentences)
Conclusion/deviation	The performance of the sentiment extraction classifier has met the highest expectations (D1.2) and has improved the classification accuracy by more than 5% (an approximate 48%

¹ For a better understanding of the sentiment values, the sentiment score is given in a range from 0.0 to 4.0.

	has been reported in D3.4) over the baseline.
--	---

Table 8: Integration description of the Sentiment analysis service

3.3.3.9 Extractive summary and query-based extractive summarisation

Extractive summarisation refers to the generation of summaries from texts by selecting sentences from the original summaries and composing a summary from these (unchanged) sentences. The language of the summary is the same as that of the original documents. The service has been extended to incorporate metrics based on semantic features such as NE, concept and relation annotations.

Service name	Extractive summary
SVN artefact	ms-svc-summ
Functional description	Generates extractive summaries from either a single document or a whole collection.
Deployment status	Deployed
P1 version	Version 0.2, baseline version
P2 version	Version 1.0, fully functional version
P2 improvements	Current version incorporates first methods resulting from research in T6.4. Relevance metrics for sentences are now based on semantic features in addition to keywords.
FS version	Version 2.0, fully functional version
FS improvements	Incorporates all previous methods with optimised performance
Maximum amount of queries per second	The performance of the service is good, meaning that up to a summary can be processed per second.
Supported languages	English
Known issues	No support for multilingual summarisation due to lack of corpora.
Conclusion/deviation	N/A

Table 9: Integration description of the Extractive summary service

3.3.3.10 Classification

This activity deals with the classification of News Items retrieved from the CNR into categories by means of a supervised learning technique called Random Forests (RF). The final fully functional version of the category classification service makes use of two sets of textual features, namely N-grams and word2vec, which are extracted from the text body of the News Items.

Service name	Category classification
SVN artefact	wp4/ms-svc-categoryClassification
Functional description	The category classification service receives as input the text body of a News Item that is retrieved from the CNR, extracts two sets of textual features (N-grams and word2vec) and utilises two RF

	classification models in order to provide as output the category, to which the News Item belongs. One RF model has been trained for each set of features. Next, the predicted probabilities from each model on the test set are aggregated, so as to calculate the final predictions. These probabilities are not equally weighted, as weights are individually calculated for each class based on a late fusion strategy that relies on the operational capabilities of RF, namely the Out-Of-Bag (OOB) error weighting scheme.
Deployment status	Remotely deployed
P1 version	Version 0.2, baseline version
P2 version	Version 0.8, baseline version
FS version	Version 1.0, fully functional version
FS improvements	An additional textual-based modality (word2vec), apart from the textual modality of the P2 version (N-grams), is utilised for the prediction of the category of a News Item.
Supported languages	English
Known issues	R (statistical computing software) needs to be installed in the computer where the service runs, along with the packages randomForest, tm and stringr.
Conclusion/deviation	<p>The planned visual modality has been replaced by an additional textual-based one, due to the fact that a) one main finding from the conducted experiments is that the textual modality is more reliable and suitable for the topic-based classification task than the visual one and b) it is not guaranteed that all News Items contain one or more images in order to extract visual features from them.</p> <p>The classification service has been integrated into the interface of UC2.</p>

Table 10: Integration description of the Category classification service

3.3.3.11 Context extraction

The context extraction service requires as input the textual content and the metadata that is stored in the html source of the media item. Similar to the previous version of the context extraction module (version 1.2), the module extracts either from the text or the metadata the following information: *author* (or creator of the content item); a set of *keywords* characterising the content item; the *genre* of the item (if found in the metadata), and other features like *date*, *location*, and *source*. In addition, the context extraction service offers valuable insights with respect to what constitutes an engaging, good quality news article by identifying benchmarks for characterising editorial-based news article quality. More specifically, it identifies the following proxies that can be learned and predicted in an automatic and scalable manner:

Fluency: Fluent articles are built from sentence to sentence, forming a coherent body of information; consecutive sentences are meaningfully connected; similarly, paragraphs are written in a logical sequence.

Formality: Formal articles are written by following certain writing guidelines; they are more likely to contain formal words and obey punctuation/grammar rules.

Richness: The vocabulary of rich articles is perceived as diverse and interesting by the readers; rich articles are not written in a plain and straightforward manner.

Service name	Context extraction
SVN artefact	ms-svc-context
Functional description	Given a media item, the context extraction service extracts or collects from the output of other services contextual features, such as <i>author, source, title, keywords, genre, category, date, location, literary style, language</i> , as well as editorial quality features of news articles (e.g., formality, richness)
Deployment status	Deployed
P1 version	Version 1.0, fully functional extracting a subset of the contextual features
P2 version	Version 1.2, fully functional extracting a subset of the contextual features
P2 improvements	The output of this service has been validated and improvements were made to the module. At the moment, the module receives and returns a RDF object
FS version	Version 1.3, fully functional extracting a set of basic and advanced contextual features
FS improvements	The advanced context extraction service builds on top of the previous module and additionally extracts a set of editorial quality indicators of news articles: formality, fluency, and richness. For the modelling process we learned a Generalized Linear Model using a diverse set of novel features, such as bag-of-words, shallow lexical, and syntactic (cohesion, word vectors, generative features). Our GLM regression model improves the performance (in terms of the RMSE metric) by at least 40%, compared to two baselines (model trained on shallow features and mean baseline).
Maximum amount of queries per second	Unknown
Supported languages	English (additional languages can be supported given the availability of machine translation)
Known issues	Requires NIF wrapper class to access the RDF part of the article object

Conclusion/deviation	N/A
-----------------------------	-----

Table 11: Integration description of the Context extraction service

3.3.3.12 **Audio extraction and ASR**

This service extracts audio from video files and recognises text from audio files. The resulting text files can then be processed by the textual dimension of the CEP. More detailed description of ASR Module is described in the D2.3 Advanced techniques for text analysis, machine translation and concept extraction.

Service name	Speech recognition
SVN artefact	http://voicepro.linguatec.org/rest/documents/2075/
Functional description	Extracts audio from video files and recognises test from audio files.
Deployment status	Remotely deployed
P1 version	Version 0.8, fully functional
P2 version	Version 0.9, fully functional
P2 improvements	Extension of language coverage to German
FS version	Version 1.0, fully functional
FS improvements	Service client version was improved and integrated in the CEP module to automatically process multimedia items from crawled articles. Visualisation of the results was already available.
Maximum amount of queries per second	ASR works as an asynchronous process. 4 simultaneous uploads / 100 words per minute
Supported languages	English, German
Known issues	Continued testing
Conclusion/deviation	N/A

Table 12: Integration description of the Speech recognition service

3.3.3.13 **Concept and Event detection**

This activity involves the detection of a set of predefined concepts/events in multimedia files (including videos and images), by considering visual features. To this end, various procedures are involved, such as video decoding (applicable for video files only), feature extraction and supervised classification. The video decoding procedure is responsible for extracting a predefined number of frames from a video file. The feature extraction step refers to the extraction of descriptors that describe visually images by capturing either global or local information out of the images. Finally, the classification step refers to the development of models used for classifying images or video frames to the set of predefined concepts/events.

Service name	Concept and Event detection
SVN artefact	wp2/ms-svc-conceptEventDetection
Functional description	The service receives as input a multimedia file (i.e. image or video) and computes degrees of confidence for a predefined set of concepts/events. In case the file is a video, the video decoding step extracts specific frames from the file. In the feature extraction step, deep convolutional neural networks (DCNNs) are utilised for the extraction of visual features.. Finally, all the concept/event detection models have been trained by means of the Support Vector Machines (SVM) classification algorithm. The models provide confidence scores indicating the belief of each model that the corresponding concept/event appears in the image or video file.
Deployment status	Remotely deployed
P1 version	Version 1.0, fully functional version
P2 version	Version 1.1, fully functional version
P2 improvements	Concept detection models for approximately 30 new concepts have been developed and integrated into the service. Regarding the feature extraction procedure and for a specific set of concepts, a procedure called hierarchical saliency detection (aims at finding and isolating the most important information of the image) was applied to the training image datasets, in order to pre-process them before the extraction of the features and the training of the concept detection models. Finally, in the classification step, an improved late fusion strategy (compared to the corresponding strategy used in the P1 version) was utilised for fusing the prediction results of the concept detection models.
FS version	Version 1.2, fully functional version
FS improvements	The FS version of the service makes use of deep convolutional neural networks (DCNNs), one of the most successful and widely used forms of deep networks, in order to learn features directly from the raw key frame pixels. This results in the extraction of more sophisticated visual representations, compared to the previous versions of the service.
Maximum amount of queries per second	Unknown
Supported languages	N/A

Known issues	Python needs to be installed in the computer where the service runs, along with several packages, such as numpy ² and sklearn ³ .
Conclusion/deviation	The concept and event detection service has been integrated into the interface of UC1.

Table 13: Integration description of the Concept and Event detection service

3.3.3.14 Indexing (Simmo Mongo Storing)

In the Indexing service, a multimedia data representation framework that allows for the efficient storage and retrieval of SIMMO objects is developed. The service stores the News Items of CNR into MongoDB and allows the user to send a simple or more complicated query.

The current status of the service is different compared to the previous version. This is due to the adjustments that were realised in the SIMMO model, in order to be able to hold efficiently all the required information (i.e. named entities, concepts, sentiment etc.) and thus perform more complicated questions to the Mongo database.

First of all, additional fields were added regarding the concepts, the named entities and connections between the different content types of a SIMMO. The goal of this change was to simplify the retrieval from the services using MongoDB. Moreover, in this version, the SIMMO model supports storing of visual concepts and visual annotations for the multimedia. This information was significant in order to add another option for the similarity search service as in the current version similar articles are not only retrieved considering the textual content similarity but also the visual content similarity. Finally, temporal data from the videos are saved in the SIMMO representation (if an article contains any of them).

Service name	Indexing (Simmo Mongo Storing)
SVN artefact	wp4/ms-svc-simmoMongoStoring
Functional description	The service receives as input the contents of an article, including the data produced from the pipeline services, which are used for storing it as a SIMMO representation in MongoDB.
Deployment status	Deployed
P1 version	Version 1.0, fully functional version
P2 version	Version 1.0, fully functional version
P2 improvements	Performance was significantly improved
FS version	Version 1.1, fully functional version
FS improvements	Several modifications were made in the SIMMO model in order to be able to hold efficiently all the required information (i.e. named entities, concepts, sentiment, visual concepts etc.).
Maximum amount of	Normally, the number of queries that can be handled by

² Numpy: <http://www.numpy.org/>

³ Sklearn: <http://scikit-learn.org/stable/>

queries per second	MongoDB per second is in the order of 10000+.
Supported languages	N/A
Known issues	None
Conclusion/deviation	The topic-event detection and similarity search services make use of the data stored by the indexing service.

Table 14: Integration description of the Indexing service

3.3.3.15 RDF Validation

To improve the RDF data quality Ontotext integrated validation tool named RDFUnit. It is an open source test driven data-debugging framework that can run automatically or manually generated tests against SPARQL endpoint or file in one of the recommended by W3C data formats. In MULTISENSOR is used adapted version for NIF format. This helps all the partners to align their RDF output with the standard. There are two ways to use the RDF validator:

- By command line
- By user interface

History of 500 validated files in two different formats – turtle and HTML is kept on the server. This provides the users with easy way to check the results of the validated files.

Service name	Storing RDF
SVN artefact	http://multisensor.ontotext.com/cep
Functional description	Validate the RDF content produced by CEP and each service separately. Produce results in two different formats – turtle and HTML and store these results on the server. Provide the users with easy access to these results.
Deployment status	Remotely deployed
P1 version	Implemented in for P2
P2 version	Version 1.0, fully functional
FS version	Version 1.1, fully functional
FS improvements	Extended version to validate the final version of the data models.
Maximum amount of queries per second	20 files per minute
Supported languages	The validation service validate the schema of the data, so it does not depend on particular language. It supports all recommended RDF formats by W3C.
Known issues	None
Conclusion/deviation	N/A

Table 15: Integration description of the Storing RDF service

3.3.3.16 Storing RDF

The RDF Storing service is designed to handle input in the form of SIMMO JSON objects, parse it and store it in the knowledge base. Each SIMMO is stored in different context. The supported request methods are PUT and POST. For the First Prototype, this service was designed to extract particular fields from the SIMMO, extract also the json-Id part of the object and store them both in the repository. For the Second Prototype, the RDF validation service has been integrated, so each SIMMO object first is validated and if the validation is successful, the data is stored in GraphDB. This service is fully functional and implemented.

The final version of the storing service has modified to insert timestamp, when the document is processed and stored. There is also a new mapping which handle the new quality field of the documents. New functionality is added to handle SIMMOs which need to be processed by the Bulgarian Extraction Pipeline. In this case the storing service serve as a bridge. It get the object, unpack the simmo, send data to the pipeline, then get the response, encode the result as JSONLD, pack the SIMMO and send it back to MS services.

Service name	Storing RDF
SVN artefact	http://multisensor.ontotext.com/cep
Functional description	RDF storing service is designed to handle SIMMO JSON objects, parse, validate and store then in the semantic repository.
Deployment status	Remotely deployed
P1 version	Version 1.0, fully functional
P2 version	Version 1.1, fully functional
P3 version	Version 1.2, fully functional
P2 improvements	For the Second Prototype the storing service is integrated to work with the RDF validation service. This produce better data quality.
FS version	Version 2.0, fully functional
FS improvements	For the final version of the system, the storing service is improved to add timestamps, handle properly the SIMMO quality fields and work as an intermediate level between the MS service and the Bulgarian NLP Pipeline.
Maximum amount of queries per second	10 SIMMOs per minute
Supported languages	The RDF storing service does not depend on any particular language but on the data format.
Known issues	N/A
Conclusion/deviation	N/A

Table 16: Integration description of the Storing RDF service

3.3.4 Content Alignment Pipeline (CAP)

The Content Alignment Pipeline (CAP) is a different processing flow from the CEP. While the CEP is running online for every article that is retrieved from the various sources that are

crawled by MULTISENSOR, the CAP is running offline at fixed intervals and performs across the Knowledge Base (KB), finding relations between articles. Relations are considered either similarities or contradictions:

- **Similarity:** A number of articles are referring to the same subject, the same persons or locations or organisations, similar concepts or similar events are described that take place.
- **Contradiction/opposing views:** Articles that talk about the same subject use different expressions that result in different sentimentality or polarity.

The workflow of the CAP is displayed in Figure 9.

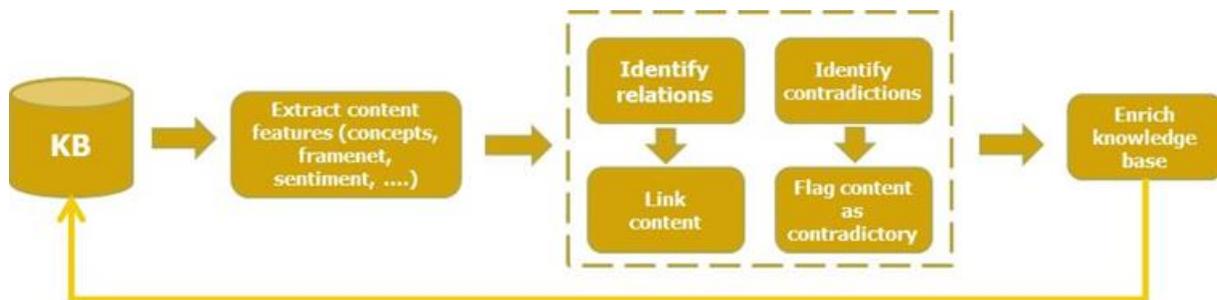


Figure 9: Content Alignment Pipeline

Service name	Content Alignment
SVN artefact	wp4/ms-svc-contentAlignment
Functional description	The CAP receives as input data retrieved from the KB. It can be considered as a meta-process as it operates on information that has been extracted through processing of original data (retrieved articles). Relation measures (referring to either similarities or contradictions) are combined in order to assess the relation score between articles.
Deployment status	Deployed
P1 version	Version 0.2, baseline version
P2 version	Version 0.5, extended baseline version
P2 improvements	The P2 version works on the actual KB data. An improved relevance score calculation method has been implemented. The service is developed as a REST service and produces JSON output.
FS version	Version 1.0, fully functional version
FS improvements	A number of novel measures for assessing the relation of items based on RDF content have been defined. The development and usage of these measures was based on ontology alignment approaches, where instead of assessing the similarity between ontological entities, the similarity between content items is calculated. The rationale is that for each information source

	different comparison methods are applied for determining relation scores. Finally, all relation scores are combined in order to end up with a single relation score for each pair of articles.
Maximum amount of queries per second	Depends on GraphDB capabilities. The average time for a query response is 918.1ms.
Supported languages	N/A
Known issues	None
Conclusion/deviation	N/A

Table 17: Integration description of the Content Alignment Pipeline

3.3.5 Social Media Analysis Pipeline (SMAP)

The Social Media Analysis pipeline (SMAP) is a set of processes related to analysis of social network data stored into the MULTISENSOR repositories. It is executed periodically in the background by the Supervisor. The SMAP pipeline performs social influence and interaction analysis on previously crawled Twitter data. The data is collected using the Twitter collector (see Section 3.3.1.2). The collector makes use of Twitter's streaming API, in order to produce JSON-encoded data containing the set of posts relevant to a given set of hashtags, together with information about the profiles of the posters and the associations among them.

Given this data, the Graph Extraction service builds a topic-dependent network of contributors based on the mentions in the set of monitored tweets. It also computes retweet probabilities between users in this network, and finally the Influential User Detection service outputs a ranked list of users by decreasing order of influence. The Community Detection service, on the other hand, makes use of the Twitter posts collected by the Twitter collector, in order to detect online dynamic communities by means of an appropriate community detection algorithm, which is applied to each graph snapshot defined by the user network of mentions. The flowchart of SMAP is depicted in Figure 10.

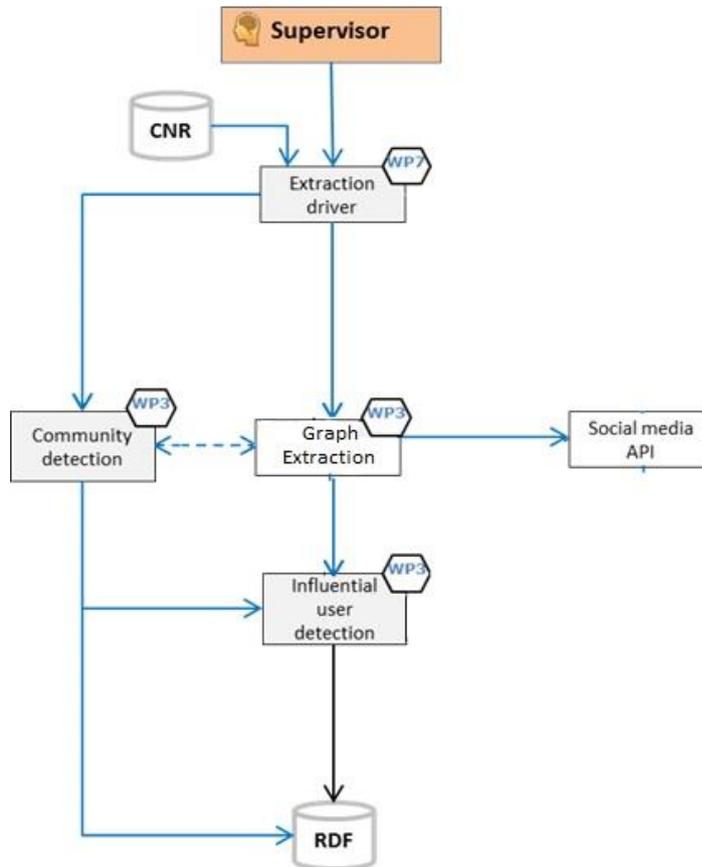


Figure 10: SMAP execution

Service name	Social Media Analysis Pipeline (SMAP)
SVN artefact	Influential User detection: wp3/ms-svc-socialMediaAnalysis Community detection: wp3/ms-svc-communityDetection
Functional description	<p>Influential User detection: The Influential User detection service receives as input Twitter posts retrieved by the Twitter collector for a given set of hashtags concerning each Use Case separately and a list containing the 10 most influential users in descending order for each Use Case is provided as output.</p> <p>Community detection: The community detection service receives as input tweets retrieved by the Twitter collector for a given set of hashtags concerning each Use Case separately, utilises the Infomap method, in order to process the tweets posted in a given date that is added as an input parameter and provides as output user communities based on the associations that arise when a user mentions another user.</p>
Deployment status	Remotely deployed
P1 version	NA
P2 version	Version 1.0, fully functional version
P2 improvements	N/A

FS version	Version 1.1, fully functional version
FS improvements	<p>Influential User detection: Implemented the measure Consistency of Sphere of Influence (CSI), a metric that quantifies the consistency of information propagation cascades in a social graph for a given user (measures the variability of the set of users influenced by the targeted user on different instances). Additionally, the Global Influence Index (GIN) was improved by incorporating CSI.</p> <p>Community detection: Introduced some improvements with respect to the visualisation of the service’s output (keep the network structure as a list of links among Twitter IDs and list the Twitter IDs within each community).</p>
Maximum amount of queries per second	Unknown
Supported languages	Multilingual
Known issues	<p>Influential User detection: Python needs to be installed in the computer where the service runs.</p> <p>Community detection: R (statistical computing software) needs to be installed in the computer where the service runs, along with the packages igraph and rjson.</p> <p>Additionally, a MongoDB database, along with the Twitter collector, need to be running on the computer where both services run.</p>
Conclusion/deviation	The influential user detection and community detection services have been integrated into the interfaces of UC2 and UC3.

Table 18: Integration description of the Social Media Analysis Pipeline (SMAP)

3.3.6 Platform Security

The whole platform, as well as individual modules, run under data security frameworks. Endpoints are protected both with HTTPS bidirectional encryption secure communication protocol (protection against DDoS & man-in-the-middle attacks) and Nginx.

Nginx is an HTTP and reverse proxy server, a mail proxy server, and a generic TCP/UDP proxy server. In addition, its modular event-driven architecture can provide a more predictable performance under high loads.

With respect to the CERTH services, security measures are applied to the machine hosting them in order to be safe from probable attacks. The port of the apache tomcat server that was set up in this machine to deploy the services is blocked in a way that it can listen for requests only by the IP of the EVERIS endpoint hosting the MULTISENSOR platform. Also, the collections existing in the mongoDB of the same machine are secured by using an authentication mechanism that blocks access to unauthorised users. The required credentials to connect to the mongoDB collections are used only by the CERTH services,

therefore they do not exist in any other machine apart from the one hosting tomcat server and mongoDB.

The online services of NER, MT, ASR and Language Identifier are implemented in REST architectural style and could be secured by using session-based authentication, either by establishing a session token via a POST or by using an API key as a POST body argument or as a cookie. Also by configuring the additional security headers on the web server layer like Access-Control-Allow-Origin (ACAO) and Access-Control-Allow-Methods (ACAH) which a part of the Cross-Origin Resource Sharing (CORS) W3C Recommendation the unauthorised and undesired calls from foreign web sites.

The Server 2 Server communication between MULTISENSOR main platform and Online/Offline services is secured by using the IP verification of unicast reverse path interface on the front end router or firewall (Unicast RPF). The effect of Unicast RPF is that it stops SMURF, DDOS and other attacks that depend on source IP address spoofing.

Ontotext provides an access to the knowledge base in two different ways.

- Open RESTful API
- GraphDB Workbench (SPARQL)

GraphDB Workbench has its own security level. All interactions, queries, updates, inserts go through it. There are different levels of access based on user accounts. The administrator has full access to the whole system. It can manage other users, accounts and access levels. The open API is exposed through GraphDB Workbench and its security is provided by the security level.

3.3.7 Platform monitoring

The server is monitored with tools MRTG4 and Nagios5. These tools allow the control of many aspects, like performance and reliability of the infrastructure. Additionally, a set of monitoring logs are kept alive and periodically managed and reviewed to ensure the correct performance and accessibility of the whole platform.

The Named Entity Recognition, Machine Translation and ASR Servers are monitored and restarted by special modules, developed by Linguattec. In comparison to the classical monitoring systems like MRTG, Anturis, Datadog, Nagios etc., the Linguattec monitoring modules are adopted to each service separately in accounting to the specific critical characteristics, like CPU and/or memory usage and the response times. The self- and cross-server monitoring allows the automatic re-initialisation and restarting of a given service and sends the emergency message to the correspondent application administrator.

All services provided by Ontotext are deployed on its own servers. The system monitoring is provided by different tools - Nagios2 and custom linux scripts. Nagios is a system for notification while the scripts are sending regular pings to check the knowledge base health. If there are delays in the responses the system will be restarted.

Services developed by CERTH are stored and deployed in the CERTH server. System performance is monitored using the Windows Performance Monitor. This tool is useful to

⁴ <http://oss.oetiker.ch/mrtg/doc/mrtg.en.html>

⁵ <https://www.nagios.org/>

analyze how programs affect CERTH server’s performance both in real time and by saving log information. Windows Performance Monitor uses performance counters, event trace data, and configuration information, which can be combined into Data Collector Sets.

3.3.8 Platform testing services

The solution for testing CEP services was developed and configured to run on development environment. This is online-based testing tool assessable via:

- <http://grinder1.multisensorproject.eu/cepTesting>

With the help of the CEP testing tool the partners can:

- Test services individually and independently of the other CEP services;
- Execute entire CEP pipeline;
- Obtain output of the services on the same page or download output stored in JSON file, and
- Access/Download logs.

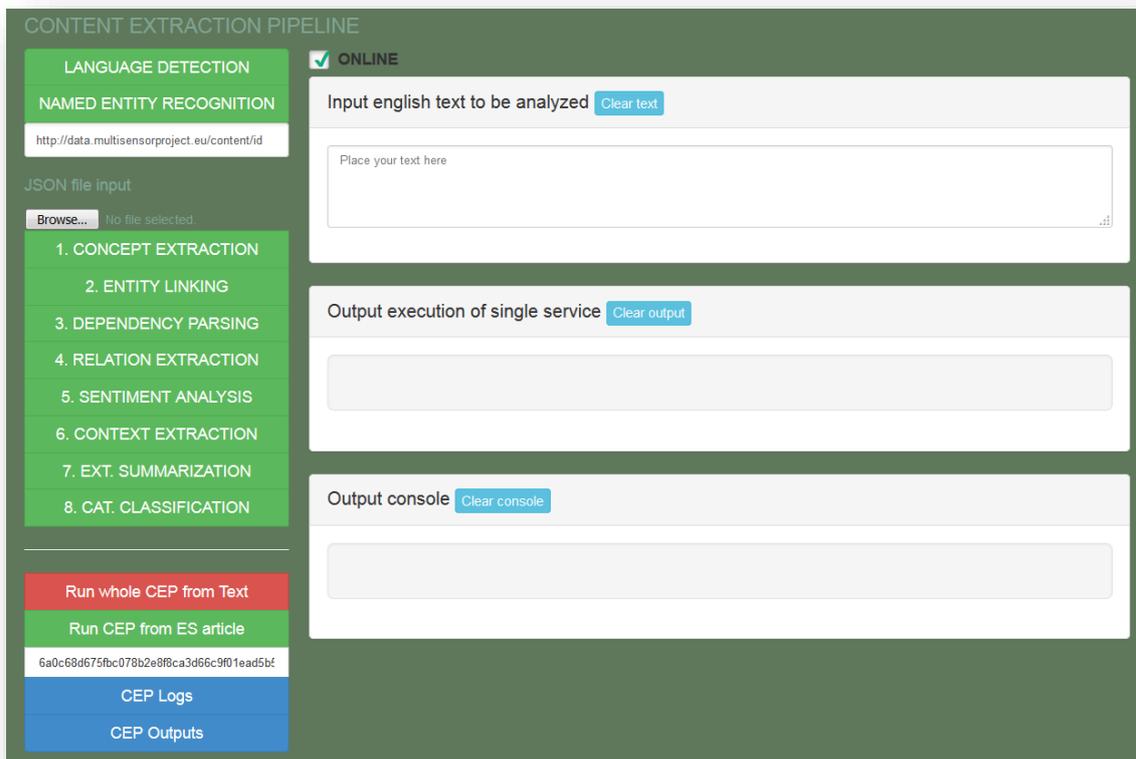


Figure 11: Final System CEP testing tool

In Figure 11, the initial view of the tool is displayed. On left side menu, services buttons are marked with green colour. By clicking on each button the respective service will be triggered, and the output of the service will be displayed on the right part of the CEP testing tool (see Figure 11). Output represents a series of containers RDF content produced by the service.

3.4 Online modality

3.4.1 Business Shared Services

3.4.1.1 Content delivery

The Content delivery service provides access to multiple information stored in the repositories of MULTISENSOR (mainly GraphDB and CNR). In other words, it grants access to enriched available data which is accessed through an internal REST API layer.

For the FP extra methods were defined along with improvements on existing endpoints in order to obtain RDF information for the whole MS platform (including all Use Cases). Specifically, Content delivery supplies results for:

On the Table below there is a representation of the complete list of the functions offered:

Base URL: <https://grinder1.multisensorproject.eu/onlineapi>

Description	Endpoint	Status
Retrieves the original SIMMO article, including title, language, etc	/news/ID/raw	Integrated
Retrieves the multimedia concepts detected per simmo (ID).	/news/ID/mediaConcepts	New integration
Retrieves Specific Concepts detected for a set of Simmos	/concepts	New integration
Returns Specific Concepts in keyword cloud format to be represented in the semantic view	/conceptsCloud	New integration
Retrieves Generic Concepts detected for a particular Simmo	/genericConcepts	New integration
Retrieves intel available on GraphDB: Sentiment, Polarity, Contextual Features (title in the original language, Summary and Entities recognised).	/news/ID/all	Improved
Retrieves the English processed intel from GraphDB per item. Similar to the /news/id/all method above.	/news/ID/all/multilingual	New integration

Retrieves total number of Simmos inserted in graphDB <i>ms-final-repository</i>	/graphTotal	New integration
Retrieves total number of Simmos in every language inserted in graphDB <i>ms-final-repository</i>	/graphLanguages	New integration
Retrieves total number of Simmos in every UC inserted in graphDB <i>ms-final-repository</i>	/graphUC	New integration
Retrieves the distance between 2 countries (ISO) used as arguments	/graphDistance/ISO1/ISO2	New integration
Retrieves numbers of articles stored in CNR	/CNR	New integration
Retrieves the textual transcript from ASR module stored in CNR	/asr/ID	New integration

Table 19: List of the online services

Service name	Content delivery
Functional description	The Online Service provides information and an API layer for all UC
Deployment status	Deployed
P1 version	Version 0.8
P2 version	Version 0.9
FS version	Version 1.0, fully functional
FS improvements	<ul style="list-style-type: none"> ▪ Multilingual RDF processed information ▪ Multilingual original raw article’s information ▪ Indicators with the population numbers represented in the Welcome view of MS ▪ Specific and generic concepts detected per one or more articles processed ▪ Distance values for two selected countries (used in UC3 – Assessment view) ▪ Multimedia concepts detected per image/video within a processed article ▪ Audio Speech Recognition transcript per article

Maximum amount of queries per second	Up to 20
Known issues	N/A
Conclusion/deviation	By using this architectural approach, the server’s sensitive information is protected

Table 20: Integration description of the Content delivery service

3.4.1.2 Semantic search

Due to the specific requirements of each use case in terms of RDF information, for the FP it has been defined two different approaches as far as the semantic retrieval is concerned:

- Ontotext’s RESTful API
- Semantic Search Online API (node js online layer)

The first strategy (Ontotext’ RESTful API) has been developed and integrated during the last stage of the project to harmonise the information retrieval for UC1 and UC3. Usage and support was already included in the base url endpoint provided:

<http://multisensor.ontotext.com/searchapi/apidocs/#!/search-controller/search>

A complete range of querying criteria is available to convert the user’s selection into querying parameters. Including but not limited to:

- Keywords
- Concepts (understood as entities)
- Concept type
- Offsets
- Total count
- Language
- Quality
- Use Case
- Date (from, to)

The corresponding results returned include not only the documents (processed articles with its contextual fields) ordered by relevance but also the most prominent entities encountered in the KB.

On the other hand, the Semantic Search Online API developed within the business layer for the SP has been improved and upgraded to meet UC2’s requirements.

The following changes were made in comparison to the Second Prototype:

- Add quality parameter and filter.
- Fixed total count values returned.
- Possibility to obtain the original title and article’s body from processed articles facetNames.
- Fixed minor issues when searching for different fields at the same time.
- Mandatory use of Content-type header which was not clear in 1st version.

- Search operator (AND & OR) can be set when searching for multiple words (e.g., querying for energy AND agency will provide different output than energy OR agency).

The endpoint for the service is the following:

<http://grinder1.multisensorproject.eu/onlineapi/search/rdf>

The method for querying against it is always POST, so the different fields will be specified on the request's body. The fields that are currently available for the search are the following one:

- subject
- title
- source
- language
- country
- description
- entity
- body
- quality

The Request parameters are listed below. In this list, the mandatory parameters are in bold.

- **queryFields**: fields to be used;
- **queryWords**: words to search on the different queryFields;
- offset: from where to start querying;
- limit: amount of articles to retrieve at the same time;
- use_case: filter by use_case field from PR API;
- pr_feed: filter by pr_feed field from PR API.

Important: the fields “*queryFields*” and “*queryWords*” must have the same length, since they are mapped together. If you use more than one field, specify the different fields using a CSV format (queryFields=title,country & queryWords=energy,en).

The possible values for the faceted search are:

- **facets**: set to **true**, and
- **facetFields**: fields to be used as facets [example: facetFields=title,subject,country]

Service name	Semantic search
Functional description	Service provides API access for semantic search
Deployment status	Deployed
P1 version	N/A
P2 version	Version 0.8, partly functional
FS version	Version 1.0, fully functional
FS improvements	Separation of searching modules per UC. Definition of new methods and parameters throughout the business online layer

	to render RDF information in the MS platform
Maximum amount of queries per second	Up to 100
Supported languages	English, Spanish, German, French, Bulgarian
Known issues	N/A
Conclusion/deviation	N/A

Table 21: Integration description of the Semantic search service

3.4.1.3 Topic-Event detection

Topic-event detection is tackled as a clustering problem on the space of concepts and named entities. The goal of this activity is to provide a grouping for a list of News Items without a priori knowledge of the number of topics. The current version of the topic-event detection service makes use of the named entities and concepts, extracted offline as described in Sections 3.3.3.3 (Named Entities recognition) and 3.3.3.5 (Concept extraction). Each detected topic is presented as a list of article IDs, ordered from the most topic-relevant to the less topic-relevant. The irrelevant news items, if any, are presented as a topic, namely “noise”, after the last detected topic.

Service name	Topic detection
SVN artefact	wp4/ms-svc-topicDetection
Functional description	The service receives as input the concepts and named entities of a list of News Items retrieved from the CNR. The topic detection service estimates the number of topics using one realisation of the DBSCAN-Martingale (i.e. a density-based clustering algorithm), extracts irrelevant News Items as noise and re-orders the News Items in each topic-event, using Latent Dirichlet Allocation.
Deployment status	Remotely deployed
P1 version	Version 0.2, baseline version
P2 version	Version 1.0, fully functional version
P2 improvements	Concepts and Named Entities are employed, so the monolingual previous baseline version became multilingual.
FS version	Version 1.1, fully functional version
FS improvements	Introduced some improvements with respect to the labelling of the topics and a faster implementation using kd-trees for nearest neighbor search in density-based clustering.
Query list size	Minimum: 100 News Items (at least 2 topics have to be detected) Maximum: 2000 News Items (for real-time results)
Supported languages	Multilingual

Known issues	R (statistical computing software) needs to be installed in the computer where the service runs, along with the packages tm, RWeka, fpc, topicmodels and rjson.
Conclusion/deviation	The topic detection service has been integrated into the interface of UC1.

Table 22: Integration description of the Topic detection service

3.4.1.4 Similarity search

The Similarity search service involves the retrieval of similar articles/documents given a query. Depending on the query, which can be an image, video or an article that may include images or videos, similarity based on a single or multiple modalities is realised. The service involves the creation and update of indexing structures for every modality (i.e. visual features, visual concepts, textual concepts, and named entities). These structures use different monomedia similarity measures. It should be noted that the indexing structures are updated regularly, that is every time a new item is stored into MongoDB holding the SIMMO objects. Finally, the similarities between the query item and the indexed objects are calculated and the top *k* results are returned.

Service name	Similarity search
SVN artefact	wp4/ms-svc-similaritySearch
Functional description	The service receives as input the ID (Elastic Search ID) of an article and outputs a list of similar article IDs, by constructing the similarity matrices of the multimedia retrieval framework (developed for the purposes of MULTISENSOR) and fusing them for the computation of one relevance score vector, which is uniform for all considered modalities.
Deployment status	Remotely deployed
P1 version	Version 0.1, dummy version
P2 version	Version 0.1, dummy version
FS version	Version 1.0, fully functional version
FS improvements	The fully functional version of the service has been implemented, by integrating the fusion-based multimedia retrieval framework developed for the purposes of MULTISENSOR.
Maximum amount of queries per second	Normally, the number of queries that can be handled by MongoDB per second is in the order of 10000+.
Supported languages	Multilingual
Known issues	Python needs to be installed in the computer where the service runs.
Conclusion/deviation	The similarity search service has been integrated into the interface of UC1.

Table 23: Integration description of the Similarity search service

3.4.1.5 Machine Translation

The Translation service receives as input text in source language and translates it into specified target language. More detailed description of Machine Translation Module is described in the D2.3 Advanced techniques for text analysis, machine translation and concept extraction in the Task T.7.

Service name	Machine Translation
SVN artefact	http://services.linguec.org/rest/lang/translate
Functional Description	Receives as input text in source language and translates it into specified target language.
Deployment status	Remotely deployed
P1 version	Version 0.8, fully functional
P2 version	Version 1.0, fully functional
P2 improvements	Improved translation quality resulting from: a) better homogenisation of training corpus b) reduction of the amounts of unknown words c) tuning of model parameters
FS version	Version 1.1, fully functional
FS improvements	Improved performance and stability of online services Optimizing sentence segmentation according to new crawler results Developed automatic monitoring and restarting of online services
Maximum amount of queries per second	7 queries (sentences) per second
Supported languages	English to/from French; English to/from German; English to/from Spanish; English to/from Bulgarian.
Known issues	N/A
Conclusion/deviation	N/A

Table 24: Integration description of the Machine translation service

3.4.1.6 Abstractive summary

Abstractive summarisation refers to the generation of multilingual summaries from contents in the RDF KB using natural language generation methods. The abstractive summarisation service can generate text-based summaries about one entity in the KB. The summarisation process consists of a text planning stage, where the most relevant contents about the entity are selected and sorted in the order in which they will appear in the text, and a multilingual generation stage where the contents are rendered in natural language.

Service name	Abstractive Summarisation service
SVN artefact	none

Functional description	Produces an abstractive summary of a document tailored to one of the entities annotated in the document.
Deployment status	Deployed
P1 version	Version 0.1, dummy
P2 version	Version 0.5, partly functional
P2 improvements	Template-based text planning and rule-based surface generation modules have been deployed for the production of summaries in English.
FS version	Version 1.0, fully functional
FS improvements	Template-based text planner has been replaced with a statistical module. Multilingual generation module now supports French, German and Spanish.
Maximum amount of queries per second	The module can produce a short summary in no less than a few seconds. Processing time depends on the length of the summary.
Supported languages	English, French, German, Spanish.
Known issues	The summariser runs on a separate server maintained by UPF due to technical requirements, which could not be met by the main EVERIS server.
Conclusion/deviation	N/A

Table 25: Integration description of the Abstractive summarisation service

3.4.1.7 Hybrid Search

Similarly to the Semantic Search, the FP comprises the integration and display of the Hybrid Search functionality, which can be understood as a list of suggested entities and concepts returned for certain characters used as input.

Within the scope of Ontotext’s Semantic RESTful API (already explained above), the *suggest* method has been successfully integrated into UC1 platform:

<http://multisensor.ontotext.com/searchapi/apidocs/#!/search-controller/suggest>

Service name	Hybrid search
Functional description	The service suggests entities and concepts to be used in the main semantic search.
Deployment status	Deployed
P1 version	N/A
P2 version	N/A
FS version	Version 1.0, fully functional
FS improvements	Complete development and integration on different views of Use Case 1 : Journalism.

Maximum amount of queries per second	Up to 10
Supported languages	English (including brands, personalities and location names)
Known issues	N/A
Conclusion/deviation	N/A

Table 26: Integration description of the Hybrid Search

3.4.1.8 Contributor analysis

The contributor analysis module receives as input a twitter handle (e.g., "@barackobama"), and then queries the Twitter API for information about the user and his immediate connections, including measures of the user's authority. The authority scores are based on three criteria: reach (number of followers and size of the ego network), relevance to a given set of keywords and retweet influence score (average fraction of followers that retweet a random post by the user).

Specifically, the service allows a user with a legitimate Twitter application and user authentication keys to crawl the profile of particular users and compute basic statistics on network and retweeting influence. Instead of giving as input a specific twitter handle, the service can work alternatively given a specific search key as input, e.g., "Barack Obama". Given this search key, the service retrieves the top 10 relevant Twitter accounts with this string and proceeds as before with each of them. The service has been fully developed and deployed. Since this an online service operating given a specific user input (i.e., a specific Twitter handle or a search key for retrieving relevant twitter handles), the service is integrated but with specific limitations. Depending of the Twitter API, the users of this service are allowed to check 3 accounts per hour.

The contributor analysis module was fully functional in the First Prototype. Hence, no further changes were required. It should be noted that the functionality of the contributor analysis module in the FS has been replaced by that of the SMAP influential user detection and community detection services.

Service name	Contributor analysis
SVN artefact	ms-svc-contributorAnalysis
Functional description	Retrieve information about a Twitter user and compute local authority scores
Deployment status	Deployed
P1 version	N/A
P2 version	Version 1.0, fully functional
FS version	Version 1.0, fully functional
FS improvements	The service has been fully developed and deployed; thus, no further improvements were planned.
Maximum amount of queries per second	N/A

Supported languages	This service is language-independent.
Known issues	Limitations of the Twitter API (3 accounts per minutes)
Conclusion/deviation	N/A

Table 27: Integration description of the Contributor analysis service

3.4.2 Other Online Services

3.4.2.1 User profile

The User Profile service was designed in order to control user accounts over the different UCx. The service is supported by all use cases and allows registration of the new record, authorise users and profile edition.

Profile information data is stored in the OPS repository.

Profile API available and it has following functionalities:

PATH	METH	DESCRIPTION	PARAMETERS	RESPONSE
/register	- GET - POST	Renders registration view with a form (once integrated in the portals) to post credentials (username, email and password) to be saved in DB (MongoDB). GET for retrieving the register page and POST for registering a new user into the system. Creates a new user in OPS DB (Mongo).	- email - username - password	Response redirects to login page.
/login	- GET - POST	Insert user's details on database. MongoDB's methods to save and insert new users and their fields. For UC2 decide redirect view (home for example). GET for retrieving the login page and POST for logging into the corresponding portal.	- email - password	JSON Object (user) + Redirect to profile page on UC portals
/home	- GET	Renders home page. For the UC portals login and registration links will be displayed.	---	Redirect to the login page could be the response or simply show the login form in the same view.
/profile	-GET -POST -PUT -DELETE	Renders users dashboard. Allows to save and manage multiple search profiles associated to that user. GET for retrieving the profile info. POST for saving search profile data. PUT for updating the user's information. DELETE for removing specific content.	-user_language -email -profile_name - keywords -country -language -media_source - relevant -irrelevant	JSON Object
/profile/profile_name_to_be_changed	-PUT	Possibility to modify a particular profile_name stored for a set of filters. PUT for updating that profile's name. All the other filters stored remain unchanged.	-profile_name	JSON Object

/logout	-GET	Sign out path to end the session and redirect to initial home page. Logs out the existing user from the system.	---	Redirect to home page.
---------	------	---	-----	------------------------

Table 28: Methods of the User Profile service

The Profile service model is implemented with the npm module a part of node.js.

Service name	User Profile
SVN artefact	ms-svc-profile
Functional description	Service provides API for profiling services
Deployment status	Deployed
P1 version	Version 1.0, fully functional
P2 version	Version 1.1, fully functional
P2 improvements	Improvements to store the relevant articles in the user folder
FS version	Version 1.2, fully functional
FS improvements	Minor code improvements and refactoring
Maximum amount of queries per second	Up to 20
Supported languages	N/A
Known issues	N/A
Conclusion/deviation	N/A

Table 29: Integration description of the User Profile service

3.4.2.2 Reference Data

The Reference Data is a service that permits to collect many indicators about the countries. Those indicators were selected by PIMEC to help the SMEs to understand what are the relevant internationalisation factors and conditions. Those indicators are organised by categories and the following Table depicts all the indicators:

Category	Sub-category	Indicators
Economic indicators	GDP	* GDP growth
		Real GDP growth rate – volume (tec00115)
		GDP per capita in PPS (tec00114)
		GDP per capita – quarterly Data (namq_aux_gph)
		Exports of goods and services in % of GDP (tet00003)
		Imports of goods and services in % of GDP (tet00004)
		Export to import ratio (tet00011)
		Inward FDI stocks in % of GDP (tec00105)
	Importation / exportation	Customs and tariffs
		Structure of taxes by economic function (gov_a_tax_str)
		Export and Import
		Current account – quarterly data (ei_bpca_q)
		Harmonised indices – monthly data (ei_cphi_m)
Political	---	Foreign Direct Investment
		Government type
		Political instability index

indicators		Corruption perception index
		General government deficit (-) and surplus (+) – quarterly data (ei_nagd_q_r2)
Social indicators	Population	Life table (demo_mlifetable)
		Human Development Index
		Population with tertiary education attainment by sex and age (edat_lfse_07)
	Work	Unemployment rate
		Harmonised unemployment rates (%) – monthly data (ei_lmhr_m)
	Health	Life expectancy
		Life expectancy by age and sex (demo_mlexpec)
Population distribution		
Cultural indicators	Urbanisation	Distribution of population by degree of 51 urbanisation, dwelling type and income group (source: SILC) (ilc_lvho01)
	Consumption habits	Economic sentiment indicator (teibs010)
		Households having access to the internet at home (isoc_pibi_hiac)
		Easiness of doing business

Table 30: List of the indicators

Most of the indicators are provided by EuroStat⁶ and WorldBank⁷. These organisations publish the data as Linked Open Data in RDF format. Therefore, the EuroStat dataset⁸ is stored in the knowledge base (GraphDB) and the indicators values are collecting by SPARQL queries.

For each indicator, SPARQL queries have been created to collect the values. To be completely adaptive to the user browsing, those queries are formalised as template to take into account specific parameters (country, time frame, etc.). In the next Table, an example of a SPARQL query is presented:

<p>SPARQL Query Template to get the Economic indicator called Inward FDI stocks in % of GDP</p> <pre># Inward FDI stocks in % of GDP (tec00105) PREFIX qb: <http://purl.org/linked-data/cube#> PREFIX eudata: <http://eurostat.linked-statistics.org/data/> PREFIX prop: <http://eurostat.linked-statistics.org/property#> PREFIX eugeo: <http://eurostat.linked-statistics.org/dic/geo#> PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> PREFIX sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> SELECT ?datePretty ?value { ?s qb:dataSet eudata:tec00105; prop:geo ?country; sdmx-dimension:timePeriod ?date; sdmx-measure:obsValue ?value; FILTER(?country = eugeo:ES) BIND(((substr(str(?date),1,4+1+2))) as ?datePretty) # Monthly returns YYYY-MM</pre>
--

Table 31: SPARQL Query Template to get the Inward FDI stocks in % of GDP

⁶ <http://ec.europa.eu/eurostat/web/main/home>

⁷ <http://databank.worldbank.org/data/home.aspx>

⁸ <http://datahub.io/es/dataset/eurostat-rdf>

For the Final System, two new indicators were added - UN COMTRADE and Google distances. The data about these indicators is available and can be downloaded through open APIs. There are some limitations on the number of requests you can send per day. According to these limitations, the data was downloaded in a two weeks period.

Service name	Reference data
SVN artefact	ms-svc-refdata
Functional description	This service permits to collect indicators on Linked Open Data datasets through SPARQL queries.
Deployment status	Deployed
P1 version	Version 0.5, baseline version
P2 version	Version 1.0, fully functional version
P2 improvements	More indicators have been identified and integrated in UC3.
FS version	Version 2.0, fully functional version
FS improvements	New indicators were added and the application was modified to handle the UN COMTRADE and Google distance APIs
Maximum amount of queries per second	Unknown
Supported languages	N/A
Known issues	The data is downloaded in csv format and after that has to be converted to RDF
Conclusion/deviation	N/A

Table 32: Integration description of the Reference data service

3.4.2.3 Decision support

The decision support system is part of Task 5.4. In order to handle this task, we decided to use two different approaches. The first one is based on loading and using statistical indicators retrieved from World Bank and Eurostat. These indicators are from different areas like – social, economic, political, sector and products. They will be used to compare all this parameters between the different countries. Based on that information, users will be able to make better decision and build their companies strategy.

The DSS system was improved to generate decisions based on specific user input. The input is based on country of origin, destination country and the product type. Based on that information, the system gets all available data from the statistical indicators, performs calculations and generates a list of the most appropriate countries for export. For more information, please see D5.4.

New indicators were added - UN COMTRADE, Google distance.

Service name	Decision support
SVN artefact	ms-svc-decsupport
Functional description	The semantic recommender will provide the users with highly

	relevant recommended articles. This will help journalists and press clipping agents to do their research, drill down on specific topic and find relevant content.
Deployment status	Deployed
P1 version	---
P2 version	Version 0.5, baseline version
FS version	Version 1.0, fully functional version
FS improvements	New decision generation process
Maximum amount of queries per second	24
Supported languages	For the first version of the recommender, we will support only English language.
Known issues	The provided data in the form of RDF dumps from Eurostat and The World Bank is old (2011 - 2014)
Conclusion/deviation	N/A

Table 33: Integration description of the Decision support service

4 PROTOTYPE APPLICATIONS

4.1 UC1: Journalism Use Case

The UC1 application is an application that should support media professionals (e.g. journalist, media expert) to find relevant information in different formats, coming from different sources, and according the social activities that were produced around.

As implemented in the Second Prototype, the access to the application is managed by a user profile service, which controls the user account (credential and preferences). The user has to login with his/her credentials to the UC1 application. When the user is logged in the system, he can access his folder that contains his/her favourite documents.

The main improvements of the UC1 are the integration of final version services (extractive summarisation, hybrid search and semantic search along with a refurbished UI design to merge all the RDF information (entities, concepts, similar articles, sentiment, categories, etc) generated on the offline modality. UX design was improved according to the user partner's suggestions.

Search section: With a simple selection of keywords and filtering criteria, the user can make a textual search by querying the updated Semantic Search online service explained in Sections 3.4.1.2 and 3.4.1.7 of this document. The available search methods are:

- **Main Semantic Search:** This is the basic possibility to search for relevant content using some keywords.
- **Hybrid Search:** When the user starts writing the query, some entities and concepts are suggested in real time with auto-completion mechanism. Even if some entities and concepts are selected, some keywords can be added as well to complement the query.
- **Multimedia Search** consists in the retrieval of textual articles which possess at least one multimedia element (image, audio or video) which is also analysed

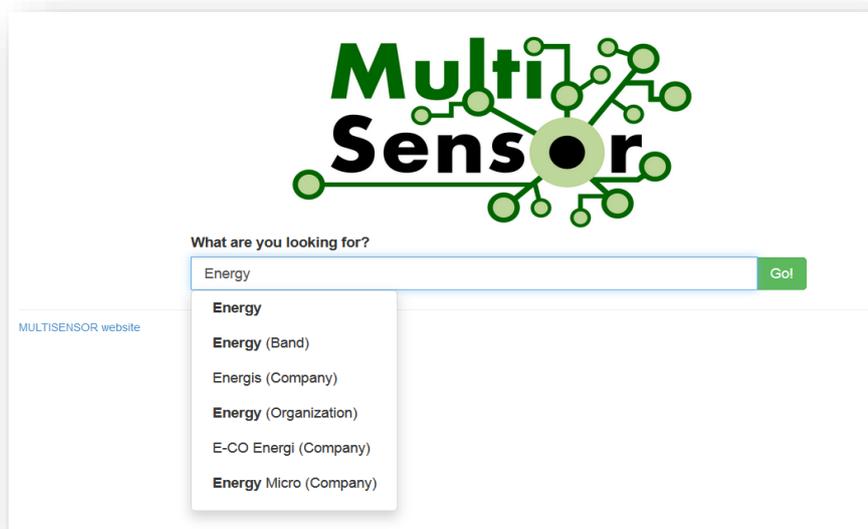


Figure 12: Main page Hybrid Search

Results section: The result page has two parts; it also has dynamic header and static footer. The header of the page displays user related information and depends on status of the user. Non authorised user will have access to authorisation dialog, while the authorised will be able to access his “Portfolio”.

On the left side, advanced search features are available (a search field that supports hybrid functionality). Search can be done on full text, entities or both. Bellow, filters like searching language, date, as well as multimedia filter, can be seen (see Figure 13).

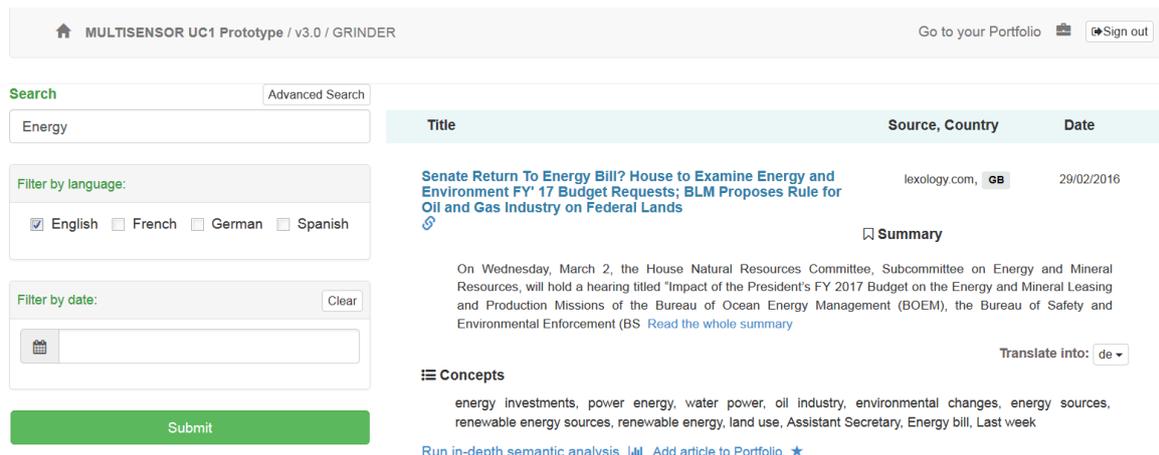


Figure 13: Result page (Header, Advanced search and result listing)

Further bottom on the left side, search related entities are displayed. By clicking on an entity it will be added to the search query. Then these entities can be used to extend the search query. Also, each entity has a link to DBpedia to obtain more information about the entity: An example is shown in Figure 14.

The right side displays results information:

- Context - Contextual features per article (title, source, etc), this information is provided by the CEP Context Extraction service;
- Summarisation – Display of the output of the Summarisation CEP service;
- Translation - The Online Machine Translation service operates on this functionality in order to translate summary to one of the five available languages (English, French, Spanish, German and Bulgarian);
- Run in-depth semantic analysis – Displays Semantic page view;
- Add article to Portfolio - The link to add article to the portfolio, for further analysis.

The screenshot shows a user interface with two main sections. On the left, there are two lists: 'Top Persons' and 'Top Organizations'. 'Top Persons' includes David Cameron, Barack Obama, Donald Trump, and Angela Merkel, each with a small profile picture and a link to 'Explore DBpedia information'. 'Top Organizations' includes the European Union and Twitter, also with links to 'Explore DBpedia information'. On the right, there are two news article cards. The top card is titled 'Energy policy U-turns 'may cost households £120 a year'' from Guardian, GB, dated 03/03/2016. It includes a 'Summary' section, a 'Translate into: de' dropdown, and a 'Concepts' section listing terms like 'energy efficiency programme' and 'autumn statement'. Below it is another article titled 'From airport expansion to energy policy, our politicians are shirking the big decisions - and it's partly our fault' from City A.M., GB, dated 02/03/2016. It also has a 'Summary', 'Translate into: de' dropdown, and 'Concepts' section listing terms like 'energy policy' and 'national interest'.

Figure 14: Result page (Entities and result listing)

Semantic analytics section: The result listing on the right side displays processed information like context (source, language, time of publishing), title, summary, translation, specific concepts and also a link to more deep analytics page that is called the “semantic analytics” (see Figure 15). On this page, more information extracted from the text is displayed (the list of named entities, the sentiment polarity, a cloud of specific concepts and the related articles are listed at the bottom of the page). In addition, there is a link to add an article to the portfolio, for further analysis.

The screenshot shows a detailed view of an article. At the top, there's a breadcrumb 'MULTISENSOR UC1 Prototype / v3.0 / GRINDER' and a 'Go to your Portfolio' link. Below that is a 'Back' button and the article source '2016-02-29 - lexology.com' with an 'Add to Portfolio' button. The article title is 'Senate Return To Energy Bill? House to Examine Energy and Environment FY' 17 Budget Requests; BLM Proposes Rule for Oil and Gas Industry on Federal Lands'. It includes a 'Summary' section and a 'Translate into: en' dropdown. The main body of the article contains text about legislative action and Senate hearings. On the right side, there are two sections: 'Entities mentioned' listing Bureau, Doe, Doi, Good, Joseph, Loan, Mike Lee, Neil, Power, and Water; and 'Extracted Text Concepts' showing a word cloud with terms like 'oil industry', 'water power', 'energy investments', 'Last week', 'renewable energy', 'land use', 'energy sources', 'environmental changes', 'Energy bill', 'Assistant Secretary', and 'power energy'.

Figure 15: Semantic analytics page (textual dimension)

2016-07-21 - [Guardian](#) [Add to Portfolio](#)

Sections of Great Barrier Reef suffering from 'complete ecosystem collapse'

Summary

"Complete ecosystem collapse" is being seen on parts of the Great Barrier Reef, as fish numbers tumble and surviving corals continue to bleach into winter, according to a scientist returning from one of the worst-hit areas. "The lack of fish was the most shocking thing," said Justin Marshall, of the University of Queensland and the chief investigator of citizen science program Coral Watch.

Complete article

Sections of Great Barrier Reef suffering from 'complete ecosystem collapse' A scuba diver inspects a section of coral for bleaching in a section of the Great Barrier Reef of Queensland. Photograph: Justin Marshall/University of Queensland

"Complete ecosystem collapse" is being seen on parts of the Great Barrier Reef, as fish numbers tumble and surviving corals continue to bleach into winter, according to a scientist returning from one of the worst-hit areas. The lack of fish was the most shocking thing," said Justin Marshall, of the University of Queensland and the chief investigator of citizen science program Coral Watch.

"In broad terms, I was seeing a lot less than 50% of what was there [before the bleaching] some species I wasn't seeing at all." Marshall spent a week this month conducting surveys on the reefs around Lizard Island. The Great Barrier Reef: a catastrophe laid bare Marshall said many of the fish species that were commonly seen around branching coral had completely disappeared from the area, including the black-and-white striped humbug damselfish.

He said in his time there he saw only one school of green chromis, which were previously seen all over the area. Marshall said the lack of fish was an indication that there was "complete ecosystem collapse" without enough surviving corals, the fish didn't have the shelter and food sources they needed and had died or moved elsewhere.

Without many of those fish, Marshall said the coral would face a harder time recovering, since the entire ecosystem had been degraded. Marshall said he was also surprised to see that some of the surviving corals continued to bleach, despite the southern hemisphere winter bringing cooler waters to the Great Barrier Reef. "There are still corals bleaching," Marshall said.

"Especially noticeable on Lizard Island were the soft corals. Some of them have remained bleached and some of the hard corals are still white." Related: Coral graveyard: the aftermath of bleaching on the Great Barrier Reef – in pictures He said many of them were probably not bleaching for the first time now but rather have remained bleached since it began.

"They're just holding on by the fingernails," he said. Marshall said he also saw some corals that had recovered, as well as some anemones that had bleached but not died. Overall, Marshall estimate that more than 90% of the branching corals had died around Lizard Island.

Videos

[Click to see Speech Recognition results](#)

Entities mentioned

- Barrier
- Coral
- Torres

Extracted Text Concepts

climate change Great Barrier Reef
scientist water
energy species world
lot anemones
food sources citizen science

Abstractive Summary Entities

Retrieving entities for abstractive summary ...

Extracted Multimedia Concepts

Legs Overlaid_Text
Animation_Cartoon Text
Eukaryotic_Organism
Animation_Cartoon

Figure 16: Semantic analytics page (multimedia dimension)

For the multimedia content, the semantic analytics page contains the video player of the images portfolio and the multimedia concepts detected are displayed as a key cloud (under the specific concept one). The ASR transcript is displayed as subtitles to the video.

Portfolio analysis section:

During the searching process, any article can be added to “Portfolio” for further analysis. The “Portfolio” can be accessed by clicking on the “Go to portfolio” link on the top left corner of the header.

Figure 17 shows the portfolio page.

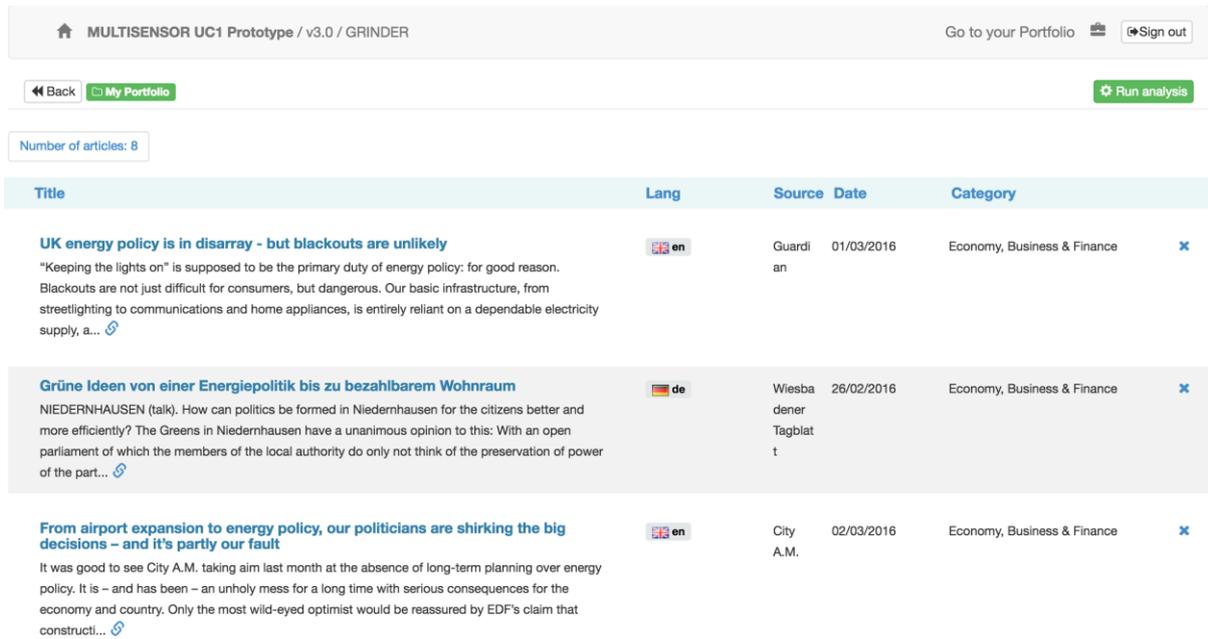


Figure 17: Portfolio home page with selected articles

By clicking the “Run analysis” button, the aggregated analytical view of the portfolio content can be generated (as shown in Figure 18), where we can see the entities, the most frequent words, the extracted topic and similar articles that the topic from this analysis contains.

- **The tag cloud:** The system analyses the term frequency over all the texts of the selected articles in the folder. It displays a graphical summary of the folder content.
- **Entity aggregates:** The analysis services retrieve all the entities that are present in the folder’s documents.
- **Topic and event detection:** Extracted topics from the portfolio articles.

Firstly, the named entities of all the documents present in the portfolio are aggregated. The same approach is followed for the specific concepts. Secondly, the results of the topic and event detection service are aggregated and displayed. For this, the labels of the clusters are shown in the keyword cloud and the list of related articles is presented on the right.

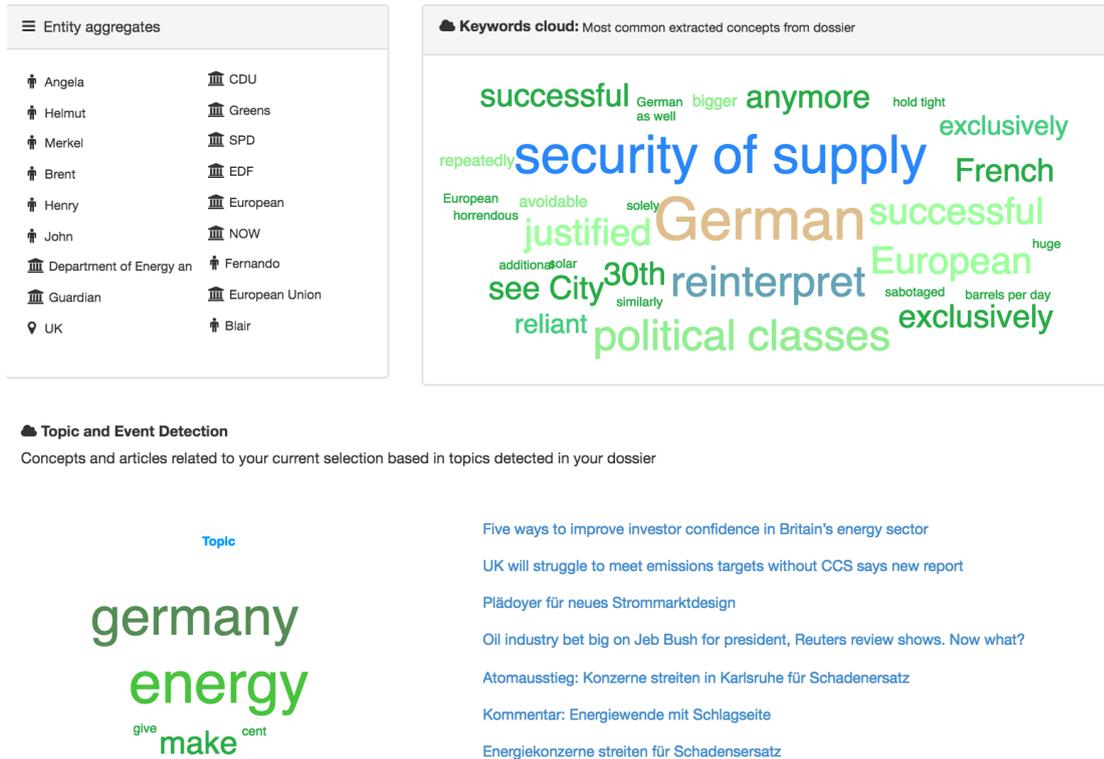


Figure 18: Portfolio aggregated portfolio view

Below is the summary of the integration of the Online Services within the UC1 application and how the extracted knowledge is presented to the end user:

- **Profile service:** This service manages the profiling functionality and the storage of the relevant articles in the user's folder;
- **Semantic search view:** The textual, multimedia, semantic and hybrid search functionalities are integrated into the KB. It displays all extracted knowledge in an intuitive manner;
- **Content delivery service:** Serves for easy access to the whole content of the SIMMO or only some specific fields from different applications;
- **Similarity search service:** This service provides a list of related articles in the semantic view and the analysis view of multi-documents;
- **Translations service:** This service translates the summary of an article in 5 different languages;
- **Summarisation service:** Works in Offline and Online mode, able to produce customisable summaries with different compression ratios and with possibility to adjust to specific keywords;
- **Abstractive summarisation service:** Works as Online service, able to produce abstractive summaries;

4.2 UC2: Media Monitoring Use Case

As previously described in D7.2, D7.4 and D7.6, the UC2 application (Media Monitoring) will replicate the workflow of a media monitoring professional to execute an analysis for a client.

This includes checking articles for relevance by various indicators and saving the relevant articles for a client’s profile. The relevant articles will then be analysed, so that conclusions can be drawn from this analysis.

Search section:

After logging in into the Prototype, the user is presented with a view to search for keywords and filter languages and countries. Alternatively, he can select a profile from the upper dropdown menu. Profiles have search settings stored for recurring searches in order to quickly populate the search mask.

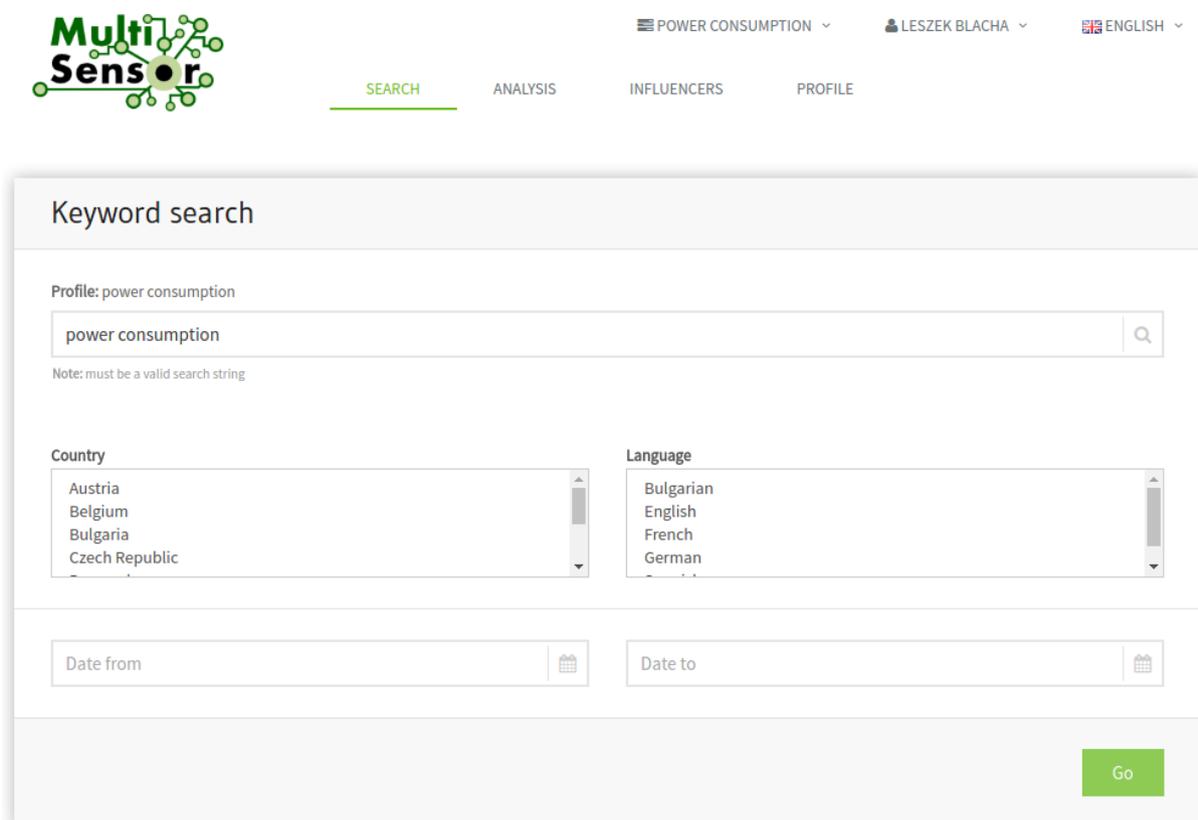


Figure 19: Search section

The search has been expanded to include semantic features, so that not only articles that contain the search term are returned, but also articles that are semantically relevant.

The results of a search query are displayed in a single article view. In the Second Prototype, the results were displayed grouped by topics first. This option is still available, but users have regarded it subordinate during evaluation.

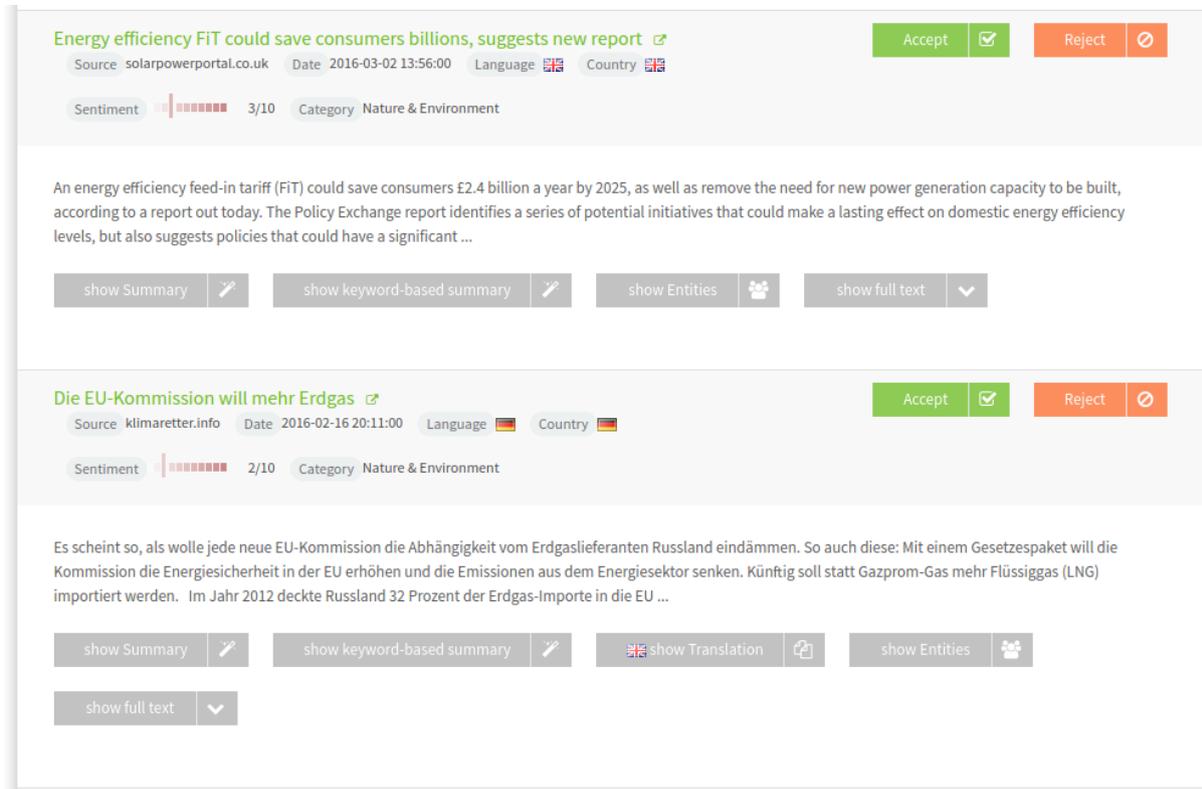


Figure 20: Results page

In order to evaluate whether an article is relevant for the client, the user can use additional functionalities like calling the summarisation and/or translation service. In addition, he can take a look at the entities extracted from the text and read the article's full text.

The new feature “keyword-based summarisation” has been implemented to create not only extractive summaries, but also summaries that are tailored to the clients' needs. When creating a keyword-based summary, detected entities and search terms can be selected. These terms are then regarded more significant and sentences including them are ranked higher when creating the summary.

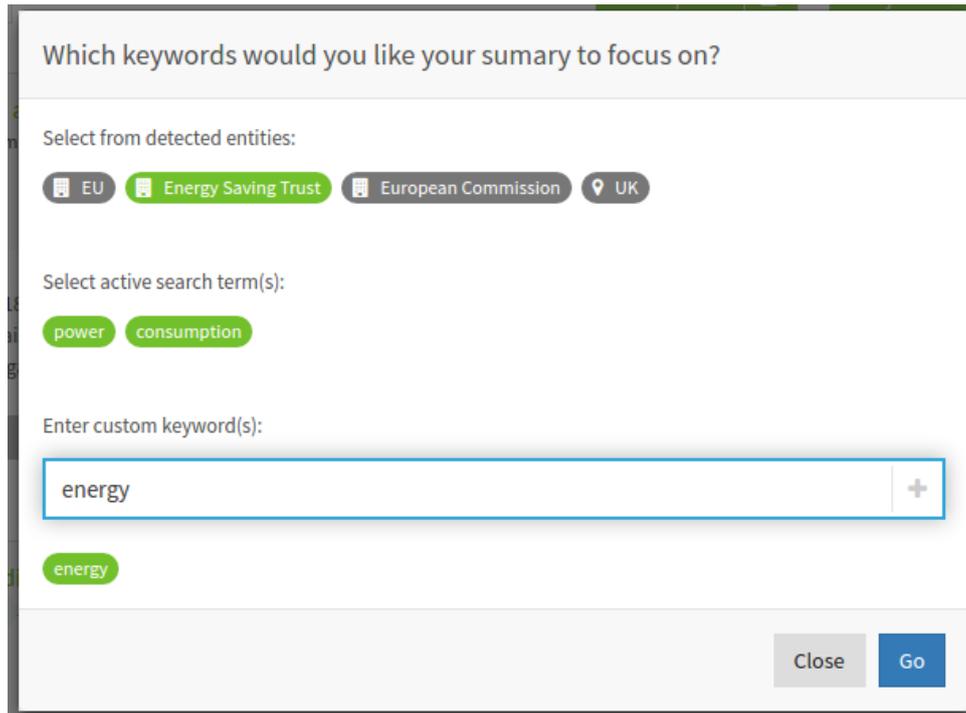


Figure 21: Keyword summary dialog

In addition to the entities and search terms, the user can also add individual terms (e.g. “energy”).

In the single article view, information for sentiment and category is now displayed. The scale for sentiment has been enlarged from a scale of 1-3 to a scale of 1-10. Sentiment is also displayed in a graphical form instead of a numeric value.



Figure 22: Single article view

As mentioned above, grouping articles by category is still available for users. It is an easy and convenient way to mark groups of articles as relevant or irrelevant.

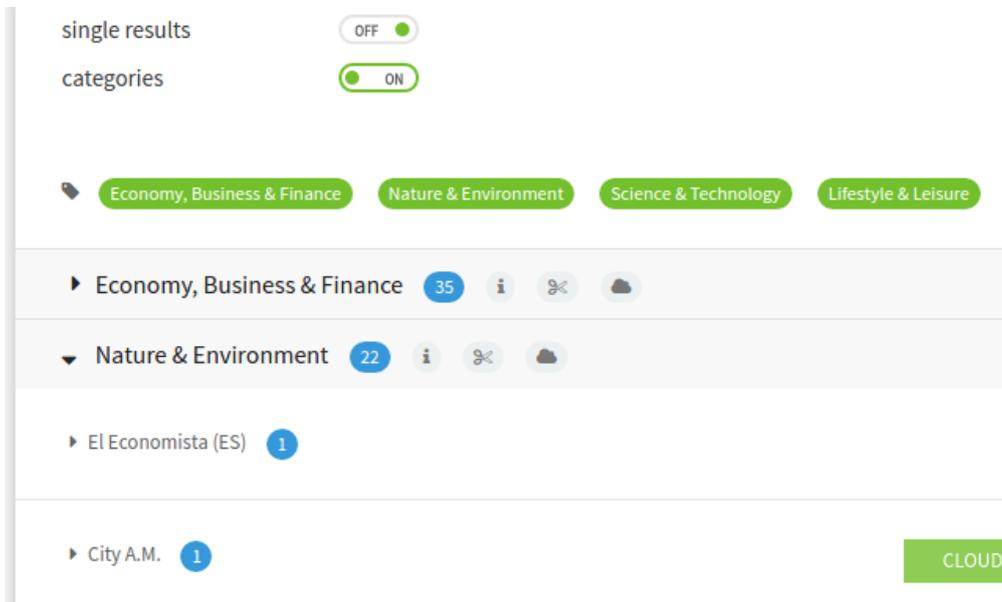


Figure 23: Article grouping

For each category, a short explanation will be displayed by clicking on the information icon.

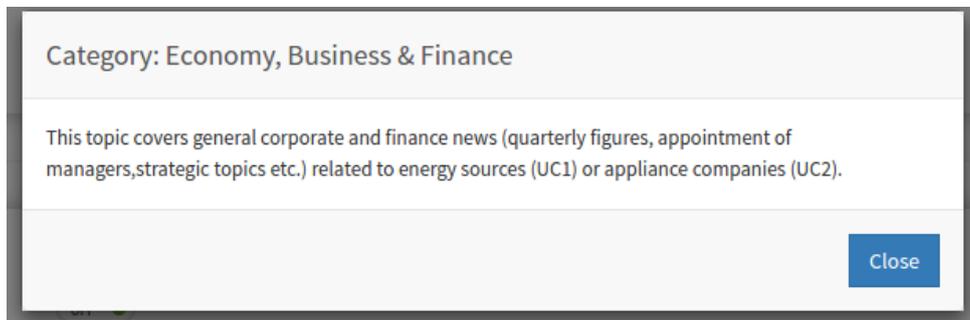


Figure 24: Help dialog

Clicking on the scissor icon will create a multi-document extractive summary.

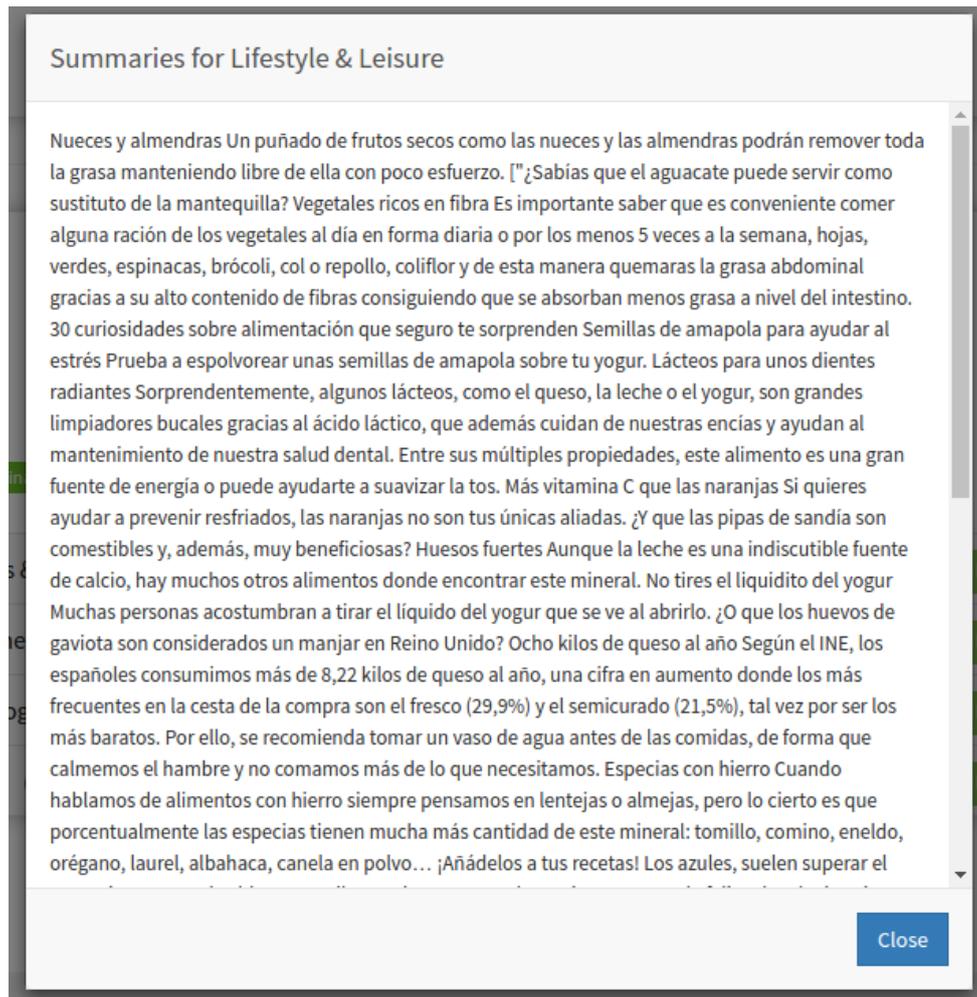


Figure 25: Multi-document extractive summary

If a user clicks on the cloud icon, the user will be presented with the main keywords from all texts within that set of articles. It is visualised in the form of a tag cloud.

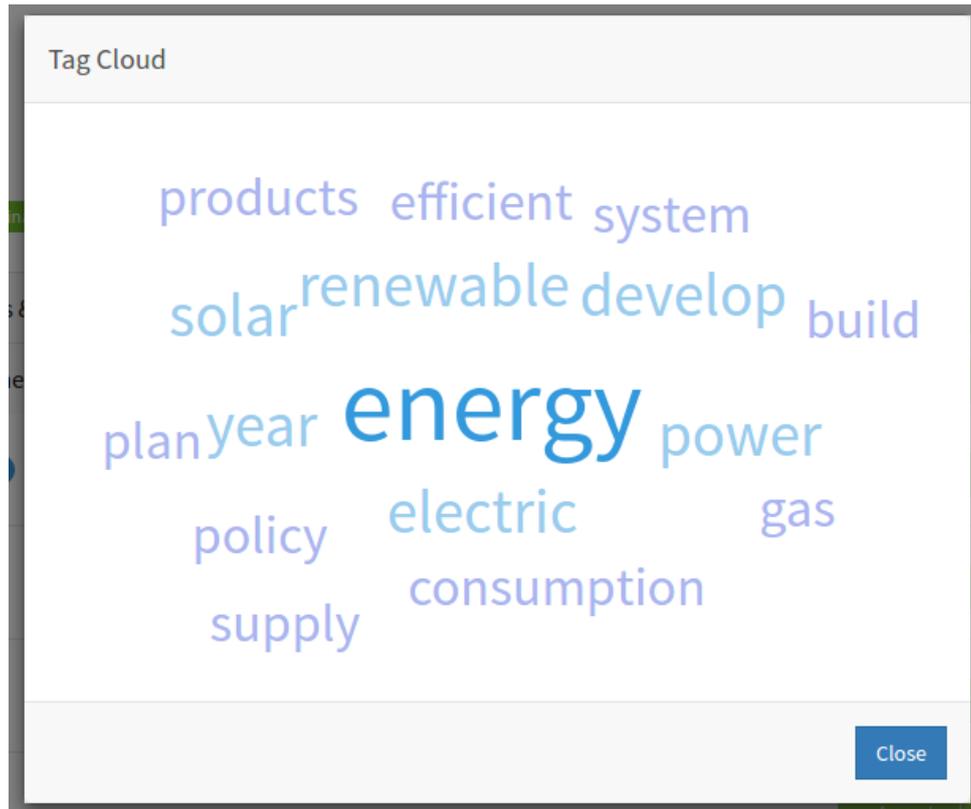


Figure 26: Aggregated most frequent words cloud

Analysis section:

In the analysis section, visual results are shown for all articles that have been marked as relevant in the search section. Rejected articles are not taken into account.

The previous tree map chart for entities provided by the semantic search service has been discarded and replaced by separate charts for locations, persons and organisations, so that the entities are not mixed, as it has been the case in the Second Prototype. In addition, the chart type has been changed from a tree map to bar charts, as recommended by users during evaluation.

The already existing charts for countries and categories mentioned in D7.6 are kept as is.

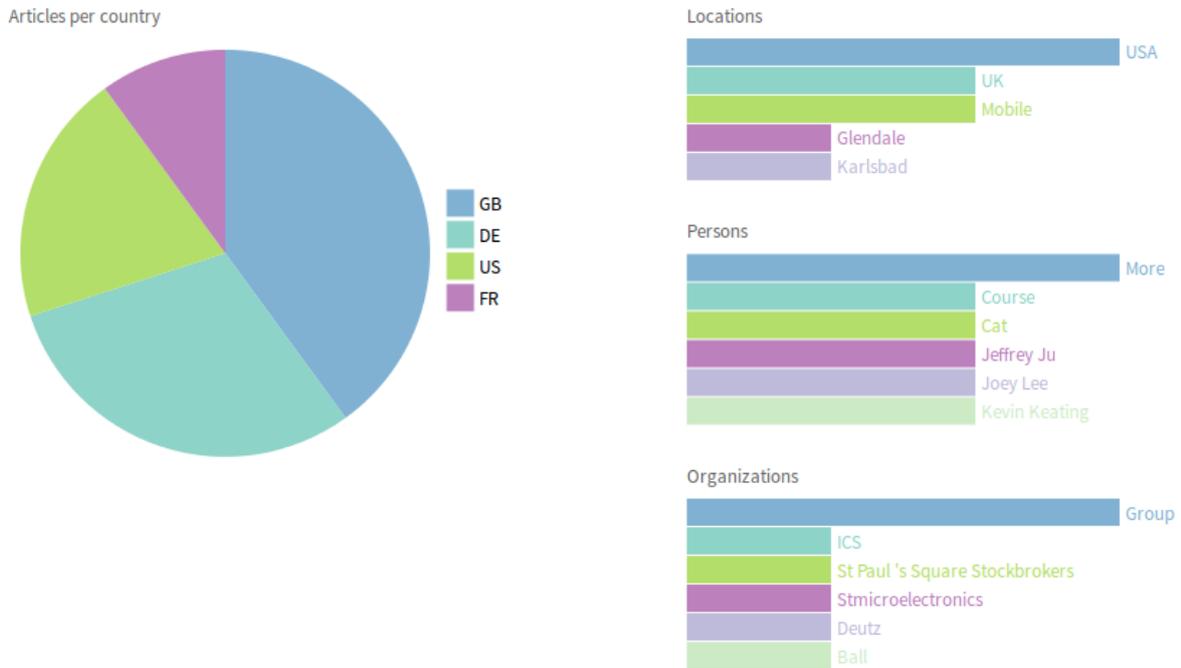


Figure 27: Analysis section

A new chart has been introduced to display the scatter of sentiment over time in order to see how the sentiment of the set of articles has developed.

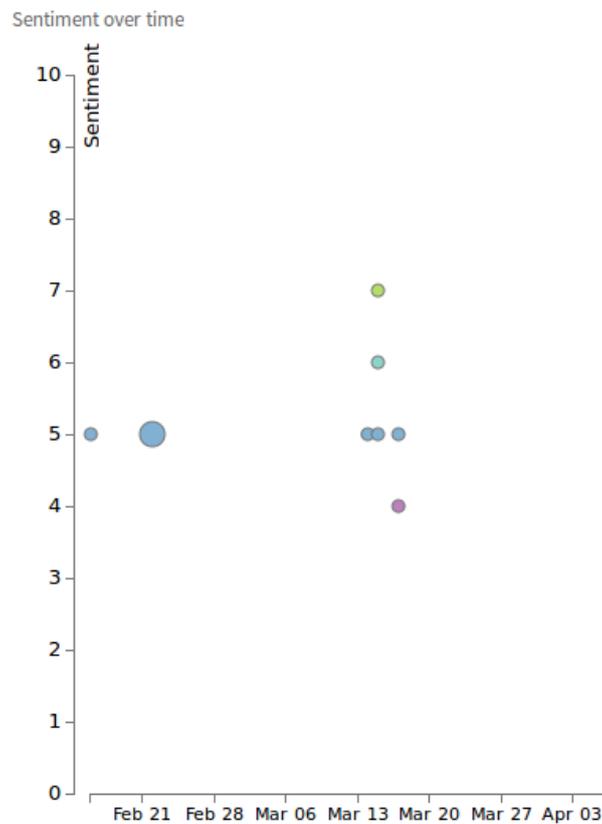


Figure 28: Scatter of the sentiment chart

By clicking on any chart section, the user is provided with the articles that fulfil the criteria, e.g. you can click on the section “GB” in the countries pie chart to see the articles from Great Britain. These are shown below the charts. Additionally a multi-document summary is provided for these articles.

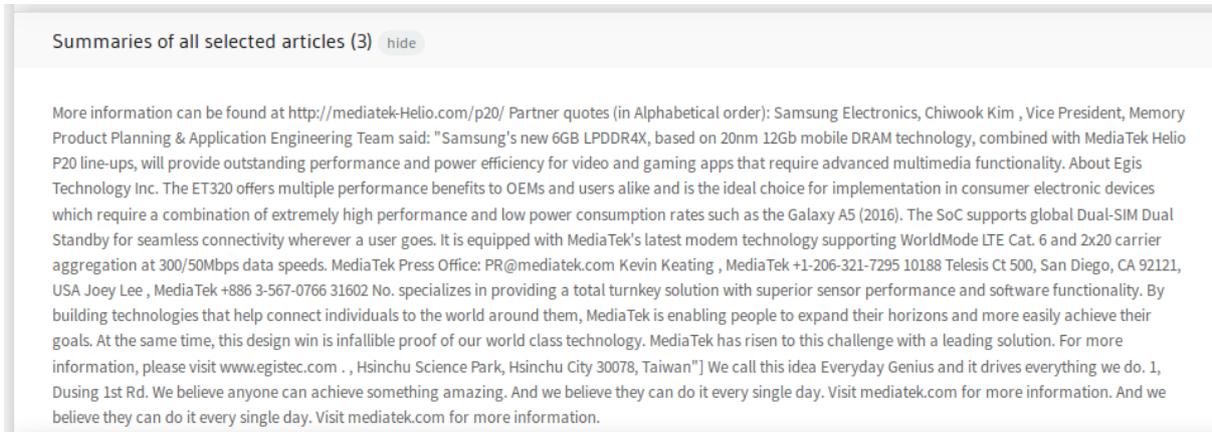


Figure 29: Multi-document summary

Influencer section:

The influencer section has been improved to add value for the user. Not only the relevant hashtags for the users and the top 10 Twitter users are displayed in a tree map, but also a new grid (Influencer meta-data) with additional data has been included. Here, not only the influence score is shown, but also the number of tweets, number of persons following and followers. By clicking on the avatar/image of the Twitter user, the Twitter page will be opened to see directly the latest activities of the particular Twitter user.

Image	Username	Name	Influence	Tweets	Following	Followers
	DiegoCusano_	Diego Cusano	0.099	1857	1041	3669
	gabrielesalari	Gabriele Salari	0.079	10874	1135	2120
	CNET	CNET	0.074	119670	295	1256593
	HWarlow	helen warlow	0.057	43413	13137	14786
	PaperGeekFr	PaperGeek	0.044	847	29	666

Figure 30: Twitter Most influential user

Finally, a new chart for community detection has been implemented. Twitter users are linked through mentioning. The more a user is mentioned by other users, the more edges target the specific user.

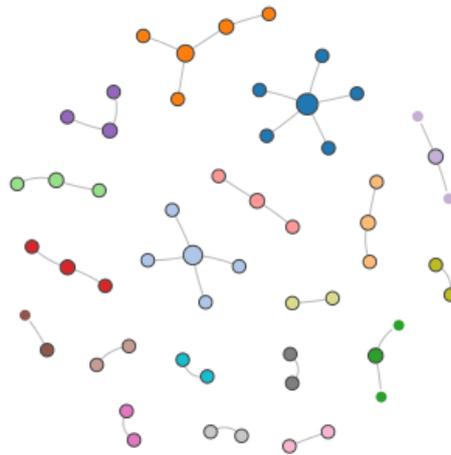


Figure 31: Detected communities chart

List of the integrated services:

- Profile service: Fully implemented
- Semantic search service: Provides sentiment, categories and entities for each article. These facets are also available for the analysis section. Fully implemented.
- Translation service: Fully implemented.
- Extractive summarisation service: Fully implemented, as well as keyword-based and multi-document summarisation.
- Sentiment analysis service: Overall sentiment per article is implemented.
- Filtering service: Implemented for relevant/irrelevant articles. Also filtering of countries and languages had already been implemented in the previous prototype.
- Influential user detection service: For pre-defined hashtags, the most influential Twitter users are retrieved. Also displaying additional data for each Twitter user is fully implemented.
- Community detection service: Fully implemented.

4.3 UC3: SME internationalisation Use Case

The UC3 application is an application that should support SMEs in order to start a process of internationalisation with any kind of products. Relevant information related to the countries, the economic situation of the market, the legal information, and the exportation/importation conditions should be retrieved easily to support decision making.

Since the Second Prototype, the scenario of the SME internationalisation has the same sectors and products. In addition, the number of indicators have been increased. This should help the user take a better decision on which country it could be interesting to export his/her products. There have been some changes to the global design in order to reflect the new content structure.

Based on the NACE taxonomy, 3 sectors have been selected and for each sector, the list of products has been modified in this Prototype. The list of the sectors and products is structured in the following Table:

Sector category	Sectors	Products
C - Manufacturing	C10 - Manufacturing of Beverages	Tea Coffee Beer Soft drinks Juice
	C11 – Manufacturing of Food products	Dairy products Cheese Meat Ice cream Olive oil Bakery Vegetables Sugar Chocolate
	C13 - Manufacturing of Textiles	Animal Plant Mineral Synthetic

Table 34: List of the sectors and the corresponding products

When a user selects a specific sector, articles about that sector are shown. After the selection of a product, the search will contain specific information about it. The whole procedure is displayed in the following images:

Country Portugal

Politics

Economy

Society

Culture

Sector Food products

Sector information

Product Choose...

Product information

Social Media

Internationalization Support

Assessment

Title	Country	Source	Date
These healthy products are her in Food	DE	Abendzeitung	24/03/2016
<p>Summary</p> <p>A way to prepare a fresh meal is to tie my children and I also can try so to eat healthily " boil ", the actress declares in the interview with "thenewpotato.com". As a real in Food brave yoghurts, maple water, are for Spelling - "I am addicted to it" - and avocados - I "am" addicted also to it.</p> <p style="text-align: right;">Translate into: de</p> <p>Specific Concepts: Hills, MAC, beauty, products, avocados, Food, chickens, much time, juice, snack, herbs, way</p>			
Probi Acquires License for New Probiotic Strains – Strengthening Its Offer in Functional Food	GB	businesswire.co.uk	11/03/2016
<p>Summary</p> <p>Probi has signed an agreement with the Swedish biotech company, Probac, to acquire a license for two probiotic strains for use in food applications, primarily dairy products. "We regard this acquisition of new probiotic strains as a great opportunity to grow our Functional Food business area by adding the dairy segment to our offer," says Peter Nahlstedt, CEO of Probi. The strains Read the whole summary</p> <p style="text-align: right;">Translate into: de</p> <p>Specific Concepts: yoghurt, Functional Food, Verum, immune system, great opportunity, product development, dairy products, sales, customers, product portfolio, Financial Instruments, market segments</p>			
Food: Minister wants to abolish Mindesthaltbarkeits date, this makes meaning?	DE	Hamburger Morgenpost Online	27/03/2016
<p>Summary</p>			

Figure 32: Selection of a specific sector

Country Portugal

Politics

Economy

Society

Culture

Sector Food products

Sector information

Product Cheese

Product information

Social Media

Internationalization Support

Assessment

Title	Country	Source	Date
Cheese: When may one the bark con-chimneys?	DE	Berlin.de	17/03/2016
<p>Summary</p> <p>This one often still contributes Bavarian dairy farming the bark according to information from the regional association in Munich of cheese with a surface from noble mould or grease to the special flavour of the cheese type. Sliced cheese or sliced cheese often has more semi-solidly an artificial bark of liquid paraffin, wax or synthetic material. Read the whole summary</p> <p style="text-align: right;">Translate into: de</p> <p>Specific Concepts: information, Consumers, substance, account, identification, dairy, salt, storage, hard cheese, type, grease, surface</p>			
Building a 'cathedral of cheese'	IE	Irish Independent	09/03/2016
<p>Summary</p> <p>Earlier this year the Haslams carried off another prestigious prize at the Randwick cheese festival in Britain from what Ralph likes to describe as the family's "cathedral of cheese", on the 240ac family farm outside Birr. "It was hard at times but I have no regrets," he said last week as he collected yet another award for his Mossfield cheese brand at the Irish Food Writers Guild Read the whole summary</p> <p style="text-align: right;">Translate into: de</p> <p>Specific Concepts: family farm, home market, milk bottling, domestic demand, Mossfield, bank loans, milk quotas, organic milk, dairy farmers, yoghurts, agricultural products, mass market</p>			
Despite lactose intolerance cheese may be feasted on perhaps	DE	CityNEWS	26/03/2016
<p>Summary</p>			

Figure 33: Selection of a specific product

The web application keeps the same structure as in the previous Prototype. However, a new tag or field has been added in the "Product" part. The reason for this change is the fact that new services have been added. The services added to the Social Media view are two (both from the online modality). They are listed in the following Table:

Service	Description
Community Detection	It retrieves the different communities from Twitter related to a selected product.
Influential	It retrieves the influential users from

User	Twitter.
------	----------

Table 35: Services in Social Media view

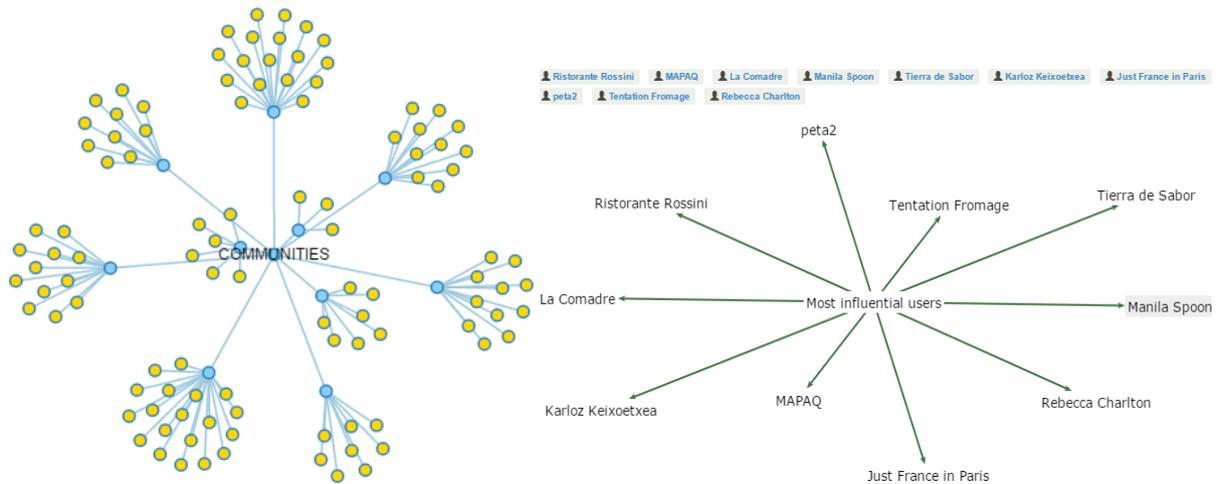


Figure 34: Social Media Services: Communities & Influential Users

The indicators are the same as in the Second Prototype. What is more, some new indicators have been added to Assessment – Table of Indicators. They are selected and organised by categories to depict the relevant information related to the target country: Politics, Economy, Society and Culture. All the categories and the corresponding indicators are presented in the following Table:

Category	Sub-category	Indicators	Graphical representation
Economic indicators	GDP	GDP growth	Line chart
		Real GDP growth rate – volume (tec00115)	Line chart
		GDP per capita in PPS (tec00114)	Line chart
		GDP per capita – quarterly Data (namq_aux_gph)	Line chart
		Exports of goods and services in % of GDP (tet00003)	Line chart
		Imports of goods and services in % of GDP (tet00004)	Line chart
		Export to import ratio (tet00011)	Line chart
		Inward FDI stocks in % of GDP (tec00105)	Line chart
	Importation / exportation	Customs and tariffs	Multidimensional lines chart
		Structure of taxes by economic function (gov_a_tax_str)	Multidimensional lines chart
		Export and Import	Multidimensional lines chart
		Current account – quarterly data (ei_bpca_q)	Line chart
		Harmonised indices – monthly data (ei_cphi_m)	Line chart
Foreign Direct Investment	Line chart		

Political indicators	---	Government type	Bar chart
		Political instability index	Bar chart
		Corruption perception index	Bar chart
		General government deficit (-) and surplus (+) – quarterly data (ei_nagd_q_r2)	Bar chart
Social indicators	Population	Life table (demo_mlifetable)	Bar chart vertical
		Human Development Index	Line chart
		Population with tertiary education attainment by sex and age (edat_lfse_07)	Bar chart with age groups
	Work	Unemployment rate	Line chart
		Harmonised unemployment rates (%) – monthly data (ei_lmhr_m)	Line chart
	Health	Life expectancy	Bar chart with age groups
		Life expectancy by age and sex (demo_mlexpec)	Bar chart with age groups
Population distribution		Line chart	
Cultural indicators	Urbanisation	Distribution of population by degree of urbanisation, dwelling type and income group (source: SILC) (ilc_lvho01)	Bar chart
	Consumption habits	Economic sentiment indicator (teibs010)	Line chart
		Households having access to the internet at home (isoc_pibi_hiac)	Histogram
		Easiness of doing business	Bar chart

Table 36: List of the indicators displayed per category

For all information about every category and sub-category, please see deliverable D7.6.

In the FS, there have been some User Interface changes in the “Assessment” view in order to be more user-friendly. Now, there is a step-by-step guide in order to select the appropriate information. This can be seen in the following Figure:

Decision Support

Select your product and country of origin. Then choose two countries to compare and obtain the Decision Support results.

Step 1: Select product you want to export. Cereals ▾

Step 2: Select your country of origin. Choose a country... ▾

Step 3: Select the two countries you are considering export to.
Portugal ▾
Spain ▾
Submit

Figure 35: Step by step user guide in Decision Support, Assessment view

In the Comparison Table, some of the indicators mentioned previously appear. In addition, this list of indicators has been modified. Some indicators have different names and others have been added. This can be seen in the following Figures:

Table of indicators

The selection of the correct country for the international investment depends on a number of indicators. These indicators are very important in order to make a first analysis of the different options that can be presented in a global market.

#	Indicator	Portugal	Spain
1	Average days to import goods ⓘ - Economy (days)	15 (2011)	10 (2011)
2	Average days to export goods ⓘ - Economy (days)	16 (2011)	9 (2011)
3	GDP Growth ⓘ - Economy (%)	-1.4 (2013)	-1.2 (2013)
4	GDP-PPS ⓘ - Economy (Index)	75 (2013)	95 (2013)
5	GDP-Capita ⓘ - Economy (€)	3,800 (2014)	5,800 (2014)
6	GDP-Exports of goods and services ⓘ - Economy (%)	40.7 (2013)	34.1 (2013)
7	Balance of trade ⓘ - Economy (ratio)	1.03 (2013)	1.08 (2013)
8	Unemployment ⓘ - Society (%)	15.3 (2014)	26 (2014)
9	Total Population ⓘ - Society (people)	10,427,301 (2014)	46,507,760 (2014)
10	Economic Sentiment ⓘ - Society (Index)	102.3 (2014)	104.2 (2014)
11	Internet households ⓘ - Culture (%)	38 (2013)	48 (2013)
12	Ease of doing business ⓘ - Culture (World ranking)	30 (2012)	44 (2012)
13	Stability Index ⓘ - Politics (Percentile)	73 (2014)	58 (2014)
14	Government Effectiveness ⓘ - Politics (Percentile)	79 (2014)	84 (2014)
15	Government Deficit ⓘ - Politics (Percentile)	-2,389.39 (2014)	-4,954 (2014)
16	Distance - From country of origin (Km) ⓘ	2,046.836	1,580.489
17	Import - from origin country to selected country (US \$) ⓘ	744,152 (2014)	57,057,105 (2014)
18	Export - from origin country to selected country (US\$) ⓘ	555,938 (2014)	2,048,017 (2014)

Figures 36-38: List of indicators – Assessment – Compare Table

In this view, an extra indicator and a new service have also been added, both provided by Ontotext:

- **Distance indicator:** It gets the distance between every selected country in the Table and the country of origin. These values are represented in the comparison Table as another indicator, “Distance – From country of origin”.
- **Countries suggestion service:** It returns a suggestion of the 3 best countries to introduce the selected product from the country of origin. The result of this service is displayed in the “Final results” section. This can be seen in the following Figure:

The display of the final result in the comparison section has changed in the FS.

Final results			
Final conclusion:	Winner of the comparison:  Spain		
Suggested countries:	Country 1: Netherlands	Country 2: Germany	Country 3: Hungary

Figure 39: Final results example

At the end of the Assessment view, there is a graphic summary of the indicators about the selected and two best suggested countries. Here, main indicators representing Economy, Culture, Politics, Import and Export are displayed in a spider chart, as it can be seen in the following Figure:

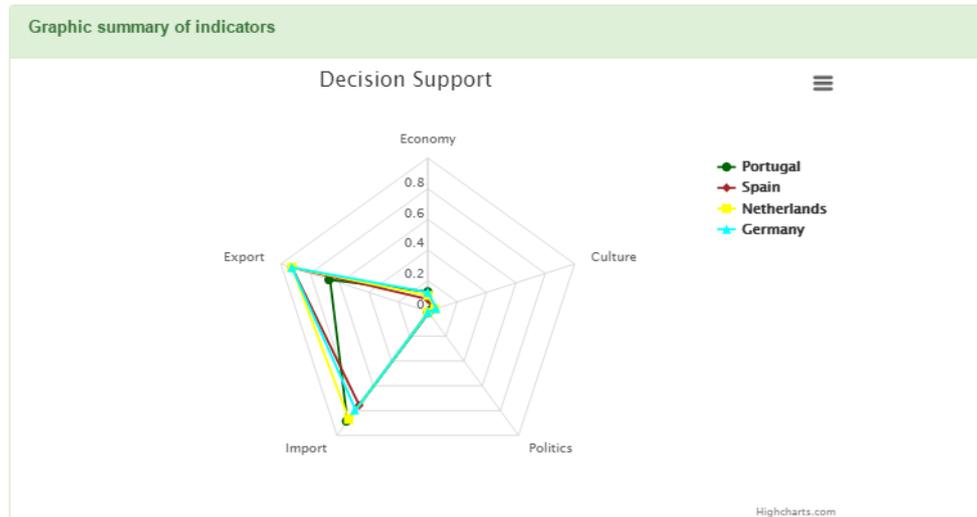


Figure 40: Example of Graphic summary of indicators

The last functionality of the UC3 application is the information search about sectors and products, and its representation. The MULTISENSOR search engine can retrieve specific information about UC3 topics. Then, the user is able to search for any keywords to retrieve the list of relevant articles.

The display of the information has changed in comparison with the previous Prototype. The new display format of articles allows seeing the summary with the “Read the whole summary/Read less” function that hides or shows part of it. After the summary, there is a drop list that allows choosing a language (FR, EN, DE, ES and BG) to translate this text and the hide and show button. At the bottom of every article, the Specific Concepts (if any) are displayed.

Country Portugal ▾

Politics

Economy

Society

Culture

Sector Food products ▾

Sector information

Product Cheese ▾

Product information

Social Media

Internationalization Support

Assessment

Title	Country	Source	Date
Cheese: When may one the bark con-chimneys?	DE	Berlin.de	17/03/2016
<p>Summary</p> <p>This one often still contributes Bavarian dairy farming the bark according to information from the regional association in Munich of cheese with a surface from noble mould or grease to the special flavour of the cheese type. Sliced cheese or sliced cheese often has more semi-solidly an artificial bark of liquid paraffin, wax or synthetic material. Read less</p> <p style="text-align: right;">Translate into: de ▾</p> <p>Specific Concepts: information, Consumers, substance, account, identification, dairy, salt, storage, hard cheese, type, grease, surface</p>			
Building a 'cathedral of cheese'	IE	Irish Independent	09/03/2016
<p>Summary</p> <p>Earlier this year the Haslams carried off another prestigious prize at the Randwick cheese festival in Britain from what Ralph likes to describe as the family's "cathedral of cheese", on the 240ac family farm outside Birr. "It was hard at times but I have no regrets," he said last week as he collected yet another award for his Mossfield cheese brand at the Irish Food Writers Guild Read the whole summary</p> <p style="text-align: right;">Translate into: de ▾</p> <p>Specific Concepts: family farm, home market, milk bottling, domestic demand, Mossfield, bank loans, milk quotas, organic milk, dairy farmers, yoghurts, agricultural products, mass market</p>			
Despite lactose intolerance cheese may be feasted on perhaps	DE	CityNEWS	26/03/2016
<p>Summary</p> <p>Who has to fight with differently strong digestive trouble after the pleasure of dairy products probably suffers a stomach ache, feeling of fullness, flatulences, diarrhea, blockages and nausea, from a lactose incompatibility. The pleasure of dairy products nevertheless does not have to be renounced; many putatively obvious products like cheese, buttermilk, yoghurt and quark can b Read the whole summary</p> <p style="text-align: right;">Translate into: de ▾</p> <p>Specific Concepts: water content, Soft cheese, dairy product, hard cheese, free products, sour cream, City centre, first notes, small intestine, natural process, world population, Cream cheese</p>			

Figure 41: Example of articles display

There are three ways to get to the search view:

- By selecting a sector on the left column: this search will show information about the sector.
- By selecting a product on the left column: this search will show information about the product.
- By entering a key word in the search box at the top: this search will show information about the entered word.

Next, a summary of the integration of the Online Services with the UC3 application is provided:

- Profile service: Not required for UC3.
- Semantic search service: The search functionality is provided by ElasticSearch (CNR).
- Similarity search service: Not required for UC3.
- Translations service: Integrated to translate the summaries in the search view.
- Summaries service: Fully integrated.
- Content delivery service: Fully integrated.
- Clustering + Filtering service: Not required for UC3.
- Reference data service: Integrated as a data wrapper that collects the data from the Linked Open Data datasets uploaded in GraphDB. Results are obtained by querying the SPARQL endpoint with query templates. The list of indicators has been extended.
- Decision support service: Integrated as a SPARQL query to compare specific indicators between two countries. The list of indicators has been extended.
- Community detection service: This service has been modified. By selecting a specific product, it retrieves a list of the detected communities.

- Influential user detection service: Improved functionality. Now, it can retrieve the most influential Twitter users by selecting a product.

5 CODE ORGANISATION

5.1 Source tree layout (D7.4 updates)

All the MULTISENSOR code and related artefacts are kept in a Subversion⁹ repository in EVERIS premises and organised on a per-Work Package basis. The root of the source tree is located at <https://quark.everis.com/svn/MULTISENSOR/trunk>.

A breakdown of the repository's layout is as follows:

- **wp1:** WP1 artefacts
 - **NA**
- **wp2:** WP2 artefacts
 - **ms-svc-dep:** Dependency Parsing Service, Maven package (see Section 3.3.3.6)
 - **ms-svc-extr:** Concept Extraction Service (see Section 3.3.3.5)
 - **ms-svc-rel:** Relation Extraction Service (see Section 3.3.3.7)
 - **nifutils:** Library to manage NIF formats, Maven package (complementary tool for the Concept extraction, Dependency parsing and the Relation extraction).
 - **ms-svc-conceptEventDetection:** : Concept and Event Detection Service, Maven package (see Section 3.3.3.13)
 - **ms-svc-ner:** NER (see Section 3.3.3.3)
 - **ms-svc-el:** Entity Linking (see Section 3.3.3.4)
 - **ms-svc-mt:** Machine translation (see Section 3.4.1.5)
- **wp3:** WP3 artefacts
 - **ms-svc-context:** Context Extraction Service (see Section 3.3.3.11)
 - **ms-svc-contributorAnalysis:** Social Graph Service (see D7.2, pp. 42)
 - **ms-svc-sa:** Sentiment Analysis service (see Section 3.3.3.8)
 - **ms-svc-communityDetection:** Community Detection module a part of SMAP service (see Section 3.3.5)
 - **ms-svc-socialMediaAnalysis:** Influential User Detection (Social Media Analysis service, see Section 3.3.5)
- **wp4:** WP4 artefacts
 - **ms-svc-categoryClassification:** Category Classification Service (see Section 3.3.3.10)
 - **ms-svc-contentAlignment:** code related to Content Alignment pipeline (see Section 3.3.4)
 - **ms-svc-simmoMongoStoring:** Indexing Service (see Section 3.3.3.14)
 - **ms-svc-topicDetection:** Topic detection service (see Section 3.4.1.3)
 - **ms-svc-entityAlignment:** Entity Alignment
 - **ms-svc-similaritySearch:** Similarity Search (see Section 3.4.1.4)

⁹ <https://subversion.apache.org/>

- **wp5:** WP5 artefacts
 - **ms-svc-decsupport:** UC3 Decision Support Service (see Section 3.4.2.3)
- **wp6:** WP6 artefacts
 - **ms-vc-summ:** Summarisation service (see Section 3.3.3.9)
 - **ms-svc-abs:** Abstractive Summary (see Section 3.4.1.6)
- **wp7:** WP7 artefacts
 - **crawler:** Crawler engine (see D7.2, Section 4.2.2.2)
 - **ms-common:** Shared Java library and services for services
 - **ms-crawler-socialmedia:** Yahoo! Crawler, Maven package (see D7.2, pp. 32)
 - **ms-js-common:** Shared Node.js modules and utilities
 - **ms-parent:** Parent Maven package for all MULTISENSOR packages
 - **ms-svc-cdelivery:** Content Delivery service (see Section 3.4.1.1)
 - **ms-svc-refdata:** UC3 Reference Data service (see Section 3.4.2.2)
 - **supervisor:** Supervisor Node.js (see D7.2, Section 4.2.2.1)
 - **uc:** Use Case portals
 - **landing:** UCx live repository population information
 - **uc1:** UC1 Node.js application and related artefacts
 - **uc2:** UC2 Angular JS application and related artefacts
 - **uc3:** UC3 Node.js application and related artefacts
 - **uclib:** Shared Node.js modules and libraries for UC applications
- **wp8:** WP8 artefacts
 - N/A
- **wp9:** WP9 artefacts
 - N/A
-

5.2 Continuous integration environment

During the development of the FS, the Jenkins open source continuous integration server was used. It allows continuous integration by pulling newly committed code from SVN. Builds can be triggered either on a schedule or by hitting a URL.

5.3 Packaging

5.3.1 Java modules

All Java modules are packaged as Maven¹⁰ artefacts for automated build, test and deployment capabilities.

In order to keep dependency management in check and ensure consistent use of package and library versions, all packages in the MULTISENSOR platform use a parent package,

¹⁰ <http://maven.apache.org>

wp7/ms-parent. This package provides versions for common dependencies and specifies shared build properties etc.

Additionally, a transversal module, **wp7/ms-common**, provides shared features for all services. This includes constants, common classes and interfaces, access to shared resources, wrappers to access common services, and more. All services must depend on this package.

Most notably, the **ms-common** package contains a Bootstrap class, which calls the supervisor to bootstrap into the platform, retrieving the shared configuration for coordination with the rest of the services.

5.3.2 Node.js modules

The Node.js modules are built as self-contained applications. They all have a package.json file, which describes their dependencies and allows using npm¹¹ to download and install them. A special module, named **ms-js-common**, contains shared modules across the rest of Node.js applications.

¹¹ <https://www.npmjs.org/>

6 INFRASTRUCTURE

The Final System is running in Amazon EC2 cloud infrastructure provisioned by EVERIS. Rationale and plans for scaling and provisioning are discussed in D7.2, Section 5.

6.1 Current farm (D7.6 updates)

All servers run Ubuntu Linux 14.04.1 LTS (“Trusty”) on x64 architecture. Ubuntu is hugely popular and as such, Personal Package Archives (PPAs) and vendor repositories are readily available providing very recent versions of core packages of MULTISENSOR (mongodb, elasticsearch, nodejs, maven, nginx).

The main server, called **msgrinder1**, is hosting the Content Extraction Pipeline services, the repositories and the three UC applications.

1. The server grinder1 has the following specifications:

- 16x x64 core (52 ECUs).
- 122 GB RAM.
- 300 GB local SSD storage (xfs).
- 100 GB EBS SSD storage (ext4).

The servers provided by Ontotext have the following specifications:

2. MULTISENSOR-prod:

- 24 cores, 2xIntel(R) Xeon(R) CPU X5680 @ 3.33GHz
- 70G RAM
- 630G SSD (zfs)
- 500 GB HDD (zfs)

3. MULTISENSOR:

- 24 cores, 2xIntel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz
- 40G RAM
- 300G HDD (zfs)

The servers provided by LT have the following specifications:

4. LT server1 (Named entities recognition, Language detection, Machine translation, ASR):

- CPU: 2 x Intel Core 2 (2.66 Ghz, 128K cache)
- RAM: 4GB
- HDD space: 30GB

5. LT Database server1:

- CPU: 2 x Intel Core 2 (2.66 Ghz, 128K cache)
- RAM: 8GB
- HDD space: 2000GB

The servers provided by CERTH have the following specifications:

6. CERTH server:

- CPU: 4 cores, Intel i7-4790K CPU @ 4.00GHz
- RAM: 32GB RAM
- 500GB SSD

- 2TB HDD

7. Twitter collector server:

- CPU: 2 x Intel Xeon E5-2620v3 6-Core (2.40GHz 15MB)
- Memory: 128GB (8 x 16GB) PC4-17000P-R 2133MHz RDIMM
- 2 x Samsung Pro 1TB

7 DEMONSTRATOR URLS AND INFORMATION

The following URLs can be used to access the different parts of the MULTISENSOR Final System:

Population information: <http://grinder1.multisensorproject.eu/>

UC1 Application: <http://grinder1.multisensorproject.eu/uc1/>

UC2 Application: <http://grinder1.multisensorproject.eu/uc2/>

UC3 Application: <http://grinder1.multisensorproject.eu/uc3/>

SVN repository: <https://quark.everis.com/svn/MULTISENSOR/trunk/>

CEP testing tool: <http://grinder1.multisensorproject.eu/cepTesting/>

Summary of the available data in the repositories:

<http://grinder1.multisensorproject.eu/>

No credentials are required to access the environments.

8 SUMMARY AND CONCLUSIONS

In D7.7, the status of the Final System is presented. It explains the system scope, repositories, services, processes and workflows. In addition, it includes some updates regarding the architecture and the integration of all the Offline and Online services. This can be summarised as follows:

- The services have been provided in their final versions.
- In addition, some new services have been developed, deployed and integrated.
- The development infrastructure has been optimised.
- The RDF repository has been populated with textual, multimedia, multilingual and social information.
- The UX for the three Use Cases has been improved based on the user partners suggestions after the evaluation of the second prototype.