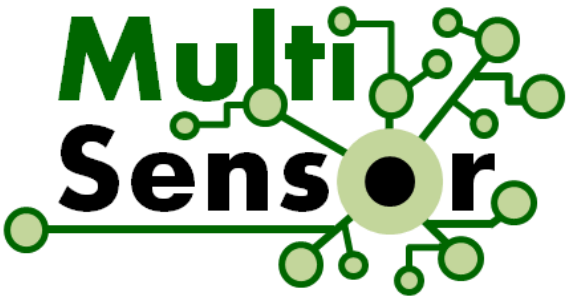

MULTISENSOR

Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation

FP7-610411

D9.7

Final exploitation report

Dissemination level:	Public
Contractual date of delivery:	Month 36, 31 October 2016
Actual date of delivery:	Month 37, 10 November 2016
Work package:	WP9 Dissemination and Exploitation
Tasks:	T9.4 Exploitation plans, T9.5 Business models
Type:	Report
Approval Status:	Final Draft
Version:	1.0
Number of pages:	37
Filename:	D9.7_FinallExploitationReport_2016-11-10_v2.0.pdf

Abstract

This document is the public version of the final exploitation deliverable of MULTISENSOR project. It aims at discussing the general exploitability of the project results by illustrating how MULTISENSOR technologies and tools can potentially contribute to the everyday tasks of journalists, media monitoring companies and business intelligence service partners targeting SMEs on the brink of internationalization. To this end, we describe general market conditions in the specific market segments and analyze existing tools already available that meet the specific needs. We also report on the exploitable assets developed in MULTISENSOR.

Co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	16-09-2016	Initial structure	Michael Jugov
0.2	29-09-2016	Contributions	ALL
0.3	06-10-2016	Integration of first round of contributions from partners	Michael Jugov
0.4	12-10-2016	Additional contributions	ALL
0.5	14-10-2016	Integrated version for internal review	Michael Jugov
0.6	21-10-2016	Internal reviews	Stefanos Vrochidis Reinhard Busch
0.7	26-10-2016	External review	Eric Karstens
0.8	28-10-2016	Final round of contributions from partners based on review comments	ALL
0.9	31-10-2016	Updated document	Stefanos Vrochidis
1.0	3-11-2016	Final integrated document	Michael Jugov
2.0	10-11-2016	Public version	Michael Jugov

Author list

Organisation	Name	Contact Information
pressrelations	Michael Jugov	michael.jugov@pressrelations.de
pressrelations	Mirja Eckhoff	mirja.eckhoff@pressrelations.de
pressrelations	Romina Gersuni	romina.gersuni@pressrelations.de
DW	Tilman Wagner	tilman.wagner@dw.de
DW	Nicolaus Heise	nicolaus.heise@dw.de
PIMEC	Teresa Forrellat	tforrellat@pimec.org
PIMEC	Marti Puigbo	mpuigbo@pimec.org
everis	Emmanuel Jamin	emmanuel.jean.jacques.jamin@everis.com
everis	Andriy Bilous	andriy.bilous@everis.com
EURECAT	Ioannis Arapakis	ioannis.arapakis@eurecat.org
Linguatec	Reinhard Busch	r.busch@linguatec.de
Ontotext	Boyan Simeonov	boyan.simeonov@ontotext.com
Ontotext	Vladimir Alexiev	vladimir.alexiev@ontotext.com
UPF	Leo Wanner	Leo.wanner@upf.edu
UPF	Gerard Casamayor	gerard.casamayor@upf.edu
CERTH	Stefanos Vrochidis	stefanos@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr
CERTH	Dimitris Liparas	dliparas@iti.gr
CERTH	Ilias Gialampoukidis	heliasgj@iti.gr

EXECUTIVE SUMMARY

In an evolving digital world, the need for relevant and analysed data for corporate decisions and management is all the more important. New companies are constantly emerging from a highly dynamic market environment, and this rapid development in novel media and business intelligence solutions drives new information needs. The role of Big Data technology is finding relevant information for customers based upon automatic or part-automatic data analysis. Discussing the potential for commercial exploitation of MULTISENSOR results takes into account the requirements and interdependencies of a unified platform designed to facilitate multidimensional content integration from heterogeneous sensors, and a thorough analysis of these requirements alongside business perspectives and market conditions.

The three MULTISENSOR user partners, Deutsche Welle, pressrelations and PIMEC have used their experience and expertise in order to define user requirements, and the three use cases, which reflect both the common challenge of having to deal with a large amount of heterogeneous data and information from different sources in different languages, as well as the industry-specific requirements.

When discussing exploitation strategies, the fundamental differences of the news industry, the media monitoring industry, and the business intelligence market need to be taken into account - which is why they are reflected in this document's structure. Our market analysis has found exploitable potential for commercial exploitation of MULTISENSOR in targeting the three relevant market segments.

Abbreviations and Acronyms

AMEC	International Association for Measurement and Evaluation of Communication
ARD	Association of public service broadcasters in Germany
ASR	Automatic Speech Recognition
BMCO	Broadcast Mobile Convergence
CAC	Cost of Acquiring a Customer
CEP	Content Extraction Pipeline
DAML	DARPA Agent Markup Language
DID	Digital Item Definition
DII	Digital Item Identification
DRM	Digital Rights Management
EBU	European Broadcast Union
EC	European Commission
ETSI	European Telecommunications Standards Institute
FIBEP	Fédération Internationale des Bureaux d'Extraits de Presse
ICEX	España Exportación e Inversiones
IEEE	Institute of Electrical and Electronics Engineers
IP	Integrated Project
IPR	Intellectual Property Rights
IPTC	International Press Telecommunications Council
IST	Information Society Technologies
JPEG	Joint Photographic Experts Group
KPI	Key Performance Indicator
LTV	Lifetime Value of a Customer
MAF	Multimedia Application Format
MMC	Media Monitoring Company
MPEG	Moving Picture Experts Group
MT	Machine Translation
NER	Names Entity Recognition
NITF	News Industry Text Format
NoE	Network of Excellence
OCR	Optical Character Recognition
OWL	Ontology Web Language
OWL-QL	Ontology Web Language Query Language
OWL-DL	Ontology Web Language Description Language
RDF	Resource Definition Framework
ROI	Return On Investment
RSS	Really Simple Syndication
SaaS	Software as a Service
SME	Small and Medium-Size Enterprises

STREP	Specific Targeted Research Projects
SWOT	Strengths Weaknesses Opportunities Threats
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
SWRL	Semantic Web Rule Language

Table of Contents

1 INTRODUCTION

The last three years, ever since the MULTISENSOR project started, have shown huge development in the area of digital media. In the context of this project, we developed and validated technologies with the aid of three main use cases (a) International mass media news, (b) media monitoring and (c) SME international investments.

While the initial results of our market studies that were documented in D9.4 have served as a starting point and continuous guideline during the project's lifespan, it is now time to take another in-depth look at what's available and what has changed. The question now is whether MULTISENSOR can still fit into this landscape, and if so, where it can exceed what is already out there and where it may fall short making further improvements necessary in order to exploit it further in the market.

When discussing exploitation strategies, the fundamental differences of the news industry, the media monitoring industry, and the business intelligence market need to be taken into account - which is why they are reflected in this document's structure.

Our market analysis has found both exploitable potential as well as significant drawbacks for commercial exploitation of MULTISENSOR in targeting the three relevant market segments. An exploitation inventory that outlines the project's foreground has also been prepared.

This report is structured as follows: in the second section we provide an updated overview market analysis for all the three targeted market segments. Next, we describe the exploitation inventory of the project in terms of modules created, licenses and IPR. Last, we conclude with a wrap up of lessons learnt from the current study.

2 MARKET SEGMENTS

2.1 Journalistic News Organizations

As already stated in Deliverable 9.4, the technological advancement in the digital media sectoralready has, and continues to have, a huge impact on the field of journalism. On the one hand, the number of sources for journalists is constantly increasing with social networks allowing more and more people to publish more and more information in the form of text, audio, images and video to the world. On the other hand, there is a constant movement of technology companies trying to tame these large numbers of information. With MULTISENSOR, we have tried to pick up some of the requirements journalists have in regards to dealing with this large number. When we started, we took a close look at the existing market, comparing available tools and services out there, to see where our project would fit in and where it could add additional value. In this deliverable, we retain an even more in depth look in the market to understand how MULTISENSOR can be positioned and what would be its unique selling point in the news industry.

We've done so in particular while keeping in mind the two main features identified in D9.4 as the major strengths of MULTISENSOR for the journalistic sector, especially in the areas of news and investigative journalism. These two features are the automatically generated summaries that help reduce the effort of pursuing a multitude of original sources manually, as well as the integrated machine translation of these summaries, allowing for a larger number of articles to be scanned by the users by also tapping into languages the users are not familiar with.

2.1.1 Market Analysis

The in-depth market analysis done in D9.4 focused on different aspects of the MULTISENSOR project. It quickly became clear that there was no tool that covered all envisioned features, which led us to separate the research into different sections to keep the analysis comparable. The categories covered *Social News Aggregators* (such as **virato**[1]), *Social Network Search and Analysis* (e.g. **topsy**[2]), different *online media monitoring services* (e.g. **Newsexplorer**[3]) as well as tools for *text analysis, extraction and comparison* (e.g. **Semantic Wire**[4]).

In order better to place the results of the MULTISENSOR projects in its current market environment, we took another look at the tools we had previously checked to see how much they had changed.

It turns out that many of those services haven't changed very much over the last 2 ½ years; some are not available any more, others have been bought by larger companies and been integrated into other services. But there are also a number of new offerings, with different features or simply better interfaces and services. Still there doesn't seem to be one tool covering the same range as MULTISENSOR does.

[1] http://www.virato.de/

[2] http://topsy.com/

[3] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

[4] http://www.semanticwire.com/

	automatic language detection and translation	multiple source integration	analysis				entity detection	categorization	enrichement	automatic summarisation
			sentiment	semantic	text structure	networks				
Social News Aggregator	NO	twitter, facebook, G+, Blogs	NO	partly	partly	YES	NO	YES	NO	NO
Network and Search analysis	NO	twitter only	partly	NO	NO	partly	NO	partly	NO	NO
Onlune MM Services	YES	social media & the web (to different extents)	NO	NO	NO	partly	partly	partly	mainly not	NO
Text analytics, extraction & comparison and web news filter	partly	mainly manual input necessary	mainly not	mainly	partly	mainly not	YES	YES	NO	NO
Social Networks	YES	NO	NO	partly	partly	partly	partly	partly	partly	NO
Multisensor (UC1)	WP2	WP2 & WP4	WP3	WP5	WP3	WP3	WP5	WP5	WP5	WP6

Figure 1: Comparing planned MULTISENSOR functionality with tools currently available

In the area of **Social News Aggregators** for example, *Virato* is still offering the same kind of service that it did almost three years ago, when we first looked at it for comparison. However, it has moved from being a free trends watch tool to a paid-for service. Looking at the feature list promoted on the website, it has stuck to its original focus of finding the most widely shared articles in social networks. Plus, it added some analytics features. But all in all, it has stayed the same. *Rivva*, as well, has practically stayed the same. It is still a freely available tool, open to use for everyone, listing the most popular articles from German media and highlighting the number of their shares in social networks.

In the area of **Social Network Search and Analysis,** *Topsy* has been shut down by its new owner, Apple, without any real replacement or alternative. This has taken one of the top Search & Analytics tools off the market. Apple has (so far) not tried to bring something similar into the market. It is speculated that the technology behind Topsy will be integrated into Siri or other Apple products.

Meanwhile nothing has changed for *Twitonomy*. It is still available as a search and analysis tool, solely for Twitter, allowing for in-depth analysis of what's happening around your twitter account. The main features remain unchanged; part of it is still free, for the rest there is a paywall.

Among **Online Media Monitoring Services,** *Mention* (formerly alert.io) is still around and has further improved its media monitoring experience, by improving their User Interface and probably some technological tweaks as well. But at the core, it is still the same service, offering some semi-automated search for mentions of certain keywords and pages, allowing for an overview of 7 languages (mainly the 'big' ones). Its most direct competitor, *Google Alerts*, is also still around, but hasn't evolved at all.

We've furthermore looked at competitors in this field like Brandwatch[5]. They also offer insights into media analytics, however, their focus is mainly on analyzing traffic across multiple networks and sources, finding influencers and tracking your own stories. What they lack is the possibility of finding relevant content for the users to use. The most convincing example we've found for those kinds of tasks is still the Newsexplorer[6] – a prototype from another EU research project. But this one as well hasn't changed.

There has been quite some change in the area of **text analysis, extraction and comparison.** *Churnalism*, the tool to detect similar text in different articles created by the Sunlight Foundation has been retired and is no longer available. *Opencalais*[7] has been developed further, now offering a better tagging system to extract entities, topic codes, events, relations and social tags. Also, *AlchemyAPI*[8] has extended its features, covering now also face detection, celebrity name-, age range- and gender recognition, as well as face position. *Semantic Wire*on the other hand has been discontinued and is no longer available.

A major change in this game are the social networks themselves – an aspect we had not looked at directly in the beginning of the project. Both Facebook and Twitter now offer built-in machine translation in their User Interfaces. Both networks, including some of their competitors/subnetworks (e.g. Instagram), also serve trending topics and in-depth search features to help users grasp the large amounts of data coming through these channels. Some of these efforts are fairly new, others have been around for a while, but are constantly improving user service, e.g. through artificial intelligence (AI) and machine learning (ML). But nevertheless, these services are still not perfect and only focused on each network specifically. There is no generic solution available, even though companies like Hootsuite are trying to become the one-for-all solution, building on top of the networks' APIs.

2.2 Media Monitoring Industry

2.2.1 Market Analysis

Our market analysis has been prepared with the goal of raising the consortium's awareness for general exploitability of the researched and implemented technologies, tools and services of MULTISENSOR specifically within the media monitoring industry. In our market analysis we sketch the exploitable foreground and overview the current and foreseen solutions available within the media monitoring industry. We analyse the range of offered services, prevalent business models and pricing, and by trying to identify potential demand and market niches for the project results we search for exploitable business potential taking into account opportunities and risks.

In servicing their customers, commercial media monitoring companies continuously monitor specific topics, brands or campaigns and perform deep media content analysis with the set of criteria always individually tailored to a customer's specifications. Competing on customers' parallel needs for speed, accuracy and insight, the general trend in the digital marketing industry, of which media monitoring is a part, has been on *Software-as-a-Service*

[5] https://www.brandwatch.com

[6] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html

[7] http://www.opencalais.com/

[8] http://www.alchemyapi.com/

(SaaS) platforms and the market is highly crowded with such SaaS solutions. The popularity of the SaaS business model stems in equal parts from the fact that the content these platforms usually are built around (online and social media) is inexpensive to harvest, as well as from the underlying SaaS business model by which the total Lifetime Value of a customer (LTV) usually greatly exceeds the Cost of Acquiring a Customer (CAC).

In contrast to fully automated SaaS solutions, media monitoring companies offer *managed services* performed through a hybrid production workflow, which combines data, technology and human data curation and analysis. Added value for the customer is created through a higher degree of accuracy, better customization of deep content analysis and highly customer-specific dissemination of both media monitoring as well as media analysis results. Generally speaking, SaaS solutions target customers who perform their media monitoring operations in house, while managed services target those who outsource such services.

We discuss potential exploitation of MULTISENSOR results from the perspective of integrating its results within a *hybrid media monitoring workflow*. We chose this view both because this is pressrelations' own market position, but also because the potential for exploitation as an implied SaaS tool is discussed elsewehere, namely within the journalistic as well as the SME internationalization use case.

In a hybrid media monitoring workflow, analysts perform data curation and extract relevant content from the selected media reports. This involves assessing topics, finding key messages, evaluating sentiment etc., as well as measuring and scoring media items according to customer-specific scoring criteria. Scoring and measurement are performed in order to calculate Key Performance Indicators (KPIs) designed to allow for a quantifiable overview of the Return On Investment (ROI) of specific actions, campaigns, media events or the marketing and public relations efforts as a whole. Data curation and data analysis are normally performed in a time-critical environment. This makes it essential that the analysis backend presents content to the analyst in an intelligent and structured way, using suitable visualisation techniques that allow the analyst to grasp the necessary information at little more than a glance in performing data analysis tasks quickly and efficiently.

To power hybrid media monitoring services, media content harvesting from the full 360° media spectrum (print, broadcast, online and social) must be performed. This requires specialized harvesting technologies for each media type, such as print digitization, OCR, speech-to-text, targeted crawling and integration of media content purchased from a variety of media data vendors. Through filtering and data curation, the media analysts clean the data of unwanted noise so as to deliver to customers only what fits. Additional data analysis is usually performed within a hybrid framework (computer-aided decision support plus human validation and curation). Finally, applications for media content delivery and dissemination to customers are purpose-built according to the specific requirements of customers normally looking at the consumption of media content primarily through media relations' glasses. They include features such as automatic generation and dissemination of media reviews (in various formats), alerts (in case of an unusual level of media activity or adverse events), and comprehensive dashboards showing aggregated analysis results.

2.3 Business Intelligence Service Industry

The MULTISENSOR prototype for SME internationalisation (Use Case 3) responds to the need of reducing the information burden that many companies encounter when they want to

start expanding their business to foreign countries. The platform's principal novelty is the Decision Support System that compares the countries and gives a prioritisation of the markets with the better conditions for international commercialisation. Here, we will firstly present a market analysis to situate the MULTISENSOR system in the market; secondly, we will determine the targeted market for exploitation purposes; thirdly, we will describe the commercialisation opportunities and risks and, last, we will present the exploitation path and strategy that we will follow.

2.3.1 Market Analysis

The initial country information gathering for internationalisation purposes starts by looking at their general conditions. Original sources such as the World Bank Data[9] and Eurostat[10] sites allow the user to navigate through the different countries and/or indicators in order to consult the needed data. Additionally, other websites extract indicators from a variety of sources to offer a complete picture of the countries. For instance, Trading Economics[11] allows you to view a complete set of indicators for each world country and it offers a premium membership to export the data and have a complete access to their database. The CIA World Factbook[12] also presents a country-by-country overview of the main characteristics. Similarly, Datos Macro[13] presents country indicators and it also delivers two-country comparisons so the user can easily visualise the differences between them. Euromonitor International[14] shows rather summarised country statistical information and offers paid market studies on specific industries and products of the different countries.

In this regard, Santander Trade[15] offers a complete site to discover world countries' information and focus on internationalisation. The portal presents indicators, but also some information on the political situation, consumer behaviour and international commerce. The site is directly linked with Banco Santander, and therefore it looks to create synergies between their clients. Furthermore, the portal offers paid information and studies on market trends or target markets. Indeed, being a Banco Santander client makes the portal much more attractive for the users as it opens more possibilities. However, for an outsider it offers a very complete overview of the countries' conditions for companies and consultants to assess.

Additionally, the Spanish public agency for exports and investment, ICEX, offers free country factsheets[16] that give an overview of the countries' economic situation as well as their main characteristics on international commerce. However, only a limited number of world countries are covered.

[9] http://data.worldbank.org/

[10] http://ec.europa.eu/eurostat

[11] http://www.tradingeconomics.com/

[12] https://www.cia.gov/library/publications/the-world-factbook/index.html

[13] http://www.datosmacro.com/paises/comparar

[14] http://www.euromonitor.com/

[15] https://en.portal.santandertrade.com/

[16] http://www.icex.es/icex/es/Navegacion-zona-contacto/libreria-icex/listado-libreria-icex/index.html?idTema=10707000&idColeccion=12060359

In all, from the platforms in the market that try to offer information on a large number of countries, Santander Trade is the most complete and the one focused on internationalisation. Nevertheless, not only it has a different approach compared to the MULTISENSOR system - as it falls into the bank strategy to generate a clients' network - but it also lacks the quick assessment module that MULTISENSOR can deliver. The project's Decision Support System incorporates indicators based on information mining from the web, and on a weighted sum of these indicators ranks the countries that have the better conditions for exporting a certain product. SME export managers with a vast experience assessing small companies to export configured this weighted sum. Such a tool is unique in the market and has good potential for commercialisation.

MULTISENSOR, in discussion with the Advisory Board, has defined three main profiles to target as potential clients: public administration and agencies, chambers of commerce and business organisations. These types of organisations offer internationalisation advice to a multiplicity of SMEs and can benefit from a tool that significantly increases the productivity and efficiency of the country selection process. SMEs themselves are not the main target of the MULTISENSOR system exploitation as they would need the tool only sporadically, if not once, to inform their decision on where to export; also, small companies are mainly focused on their internal product and dynamics and very often externalise the initial steps of guidance to internationalisation.

The MULTISENSOR decision advice falls in an intermediate step of the international strategy building. Indeed, the selection of the country or countries where to start exporting takes place at the beginning of the internationalisation process, in the context of the development of an International Strategic Plan. This Strategic Plan includes country prioritisation, and it is here where the Decision Support System can make a difference. By both offering the indicators in one same site and, more importantly, selecting the countries that present the better conditions for the selected product and the country of origin, the Strategic Plan provider can be much more productive and speed up the processes. In addition, MULTISENSOR can be useful not only for informing the decision of choosing the markets to target but also to discard countries.

More concretely, the targeted public administration and agencies are the departments or offices that support SMEs, as well as chambers of commerce. These administration bodies exist at a national, regional and local level all over Europe and the world. They offer assessment, advice and information to practically all levels an SME can need. As they need to deal with a large number of small and medium companies, a tool that simplifies a step in the process can make a difference in their day-to-day activity. Furthermore, the companies that more often demand this type of services from the administration tend to be smaller and have no experience on internationalisation. Thus, the Decision Support System can significantly help the public agencies to simplify and make their consultancy process more productive.

Specifically, regional and local agencies and chambers of commerce have a larger potential than national agencies, as they are usually closest to the SMEs needs. Also, they tend to have fewer resources and, thus, they can benefit much more from a tool that allows them to boost their productivity, make quick assessments and serve a larger number of clients.

Business organisations constitute another target group for our exploitation plan. Organisations that represent and give support to their associate companies also deliver

internationalisation consultancy. This assessment can be executed internally by the organisation but it can also be externalised to freelance internationalisation experts. Here, the improvement that MULTISENSOR Decision Support can offer has the potential to lower the costs of the business organisation and, at the same time, allow them to deal with a larger volume of services. Here, our strategy will highlight the economic advantages that Decision Support can provide by delivering a better service to the organisations' associates, by increasing the efficiency of SME assessment and, ultimately, by helping our client to grow its business.

3 EXPLOITATION INVENTORY

The following table lists the exploitation prospects for the modules developed in MULTISENSOR.

#	Outcome	Deliv.	Exploitation prospects
1	Name Entity extraction tool	D2.4	Exploitation by LT
2	Concept extraction module	D2.4	Freely available
3	Multimedia concept extraction framework	D2.4	Freely available
4	Machine translation module	D2.4	Exploitation by LT
5	Context analysis module	D3.4	Exploitation by Eurecat
6	Sentiment extraction module	D3.4	Exploitation by Eurecat
7	Social media mining module	D3.4	Freely available
8	Topic-based classification module	D4.3	Freely available
9	Semantic content integration framework	D4.4	Open source (OS)
10	Multimodal indexing and retrieval module	D4.3	Freely available
11	Semantic representation Infrastructure	D5.4	Open source (OS)
12	Hybrid reasoning module	D5.4	Exploitation by ONTO
13	Decision Support module	D5.4	Exploitation by ONTO
14	Summarisation module	D6.3	Open source (OS)
15	Final system	D7.7	Exploitation by the consortium

Table 1: Exploitation prospects for the modules developed in MULTISENSOR

The project outcomes have been updated according to D7.1 Roadmap deliverable as follows:

#	Module	Outcome
1	Content Extraction Module (WP2)	Named entities extraction component
2		Dependency parsing component
3		Concept extraction component
4		Automatic speech recognition component
5		Multimedia concept and event detection component
6		Machine translation component
7	User and context-centric content analysis module (WP3)	Context extraction and representation component
8		Polarity and sentiment extraction
9		Information propagation and social interaction analysis
10	Multidimensional content integration (WP4)	Multimodal indexing and retrieval component
11		Topic-based modelling component
12		Mapping discovery and validation component
13		Content alignment and integration component
14	Semantic reasoning and decision support (WP5)	Data infrastructure module
15		Semantic representation infrastructure management system module
16		Decision support system module
17	Content summarisation and delivery (WP6)	Extractive summarisation
18		Abstractive summarisation
19	System development and integration (WP7)	Data crawling module
20	Final System	MULTISENSOR

Table 2: Project Outcomes

In the following, we provide a detailed description and the exploitation perspective for each of these modules, which are considered as the main outcomes and the foreground of the project.

3.1 Content Extraction Module (WP2)

3.1.1 Named Entities Extraction Component

Module Description	The Named Entities Recognition Component identifies names in texts. Names are words, which identify objects, like 'Maastricht Treaty ', 'Berlin', 'Siemens'. Names belong to different types. The component can identify the following entity types: • persons • locations • organisations, divided into companies and institutions • amounts • dates It will be available for 5 languages (English, French, German, Spanish and Bulgarian).
Innovation Description	Unlike other NER components, which build on only shallow analysis techniques, the approach in MULTISENSOR is to choose a technology, which can be extended. This approach allows for deep analysis of texts, which is expected to result in higher precision of results and easier adaptability towards new domains.
IP rights	The NER component consists of three software components: Sentence splitting, tokenization, and NE recognition. The NE recognition uses three components: local parser, text analyser (for co reference determination etc.), and output generation. IP rights for the software components are with Linguatec. The lexica and gazetteers belong to the respective partners in the project who created them.
Foreseen license	Commercial licence
Alternative solution	• Stanford NER: Supports English, German, Spanish, Chinese. License only for non-commercial applications. • GATE Information Extraction Component ANNIE, with grammar JAPE. Grammar language supports only adjacent constituents, no longer-term dependencies. • OpeNER: Origin is an EU Research project, consisting of many components with unclear license conditions. Support status is also unclear. • Open Calais: a service offered by Thomson Reuters. Currently not able to support development of new domains or languages.
Adaptability and extensibility	Because of the modular NER architecture it can be adapted to other domains and languages with only limited effort. The software components are basically language-independent, so that only the resources need to be adapted, i.e. NER lexicon (annotated gazetteer) and grammar rules. During the project the component has run in the Linguatec cloud as this allows for easy access and improvement.

Maintenance after the duration of the project	During the project the NER component has run on Linguatec serversas this allows for easy access and improvement. After the end of the project the component is planned to be installed to CERTH's servers in order to maintain the platform and proceed according to the exploitation plan towards marketability. Linguatec plans to maintain a demo service for NER using existing infrastructure.

3.1.2 Dependency Parsing Component

Module Description	The statistical parsing module takes a natural language sentence and outputs either its surface- or deep-syntactic dependency structure (depending on the choice of the user). The module shows high quality performance for a number of languages – in particular English and Spanish. Its accuracy decisively depends on the size of the training corpus and the quality of its annotation.
Innovation Description	The deep parsing component of the module is a unique service to the market given the fact that no comparable parsers are available. We expect it to be of high interest to several downstream applications such as deep machine translation, information extraction, paraphrasing, etc.
IP rights	UPF is the owner of the deep parsing component. The IP right of the surface parsing component (which is used in a pipeline with the deep parser) belongs to a third party.
Foreseen license	GNU GPL v3[17]
Alternative solution	–
Adaptability and extensibility	The dependency-parsing module itself is language-independent. To apply it to other languages than those in MULTISENSOR, text corpora in the corresponding languages must be annotated with surface- and deep-syntactic structures. If a surface-treebank is already available, a semi-automatic mapping of the surface structures to deep-syntactic structures can be performed. The effort for the annotation may range thus from 3 PMs to 12 PMs.
Maintenance after the duration of the project	The research agenda of the UPF in the next years foresees further activities in the area of parsing. In the context of these activities, the parsing module will be maintained in UPF servers and further improved. Individual maintenance agreements will be set up with interested parties.

3.1.3 Concept Extraction Component

Module Description	The concept extraction module operates on plain text, drawing upon a number of standard semantic resources such as Babelnet, Babelfy, Framenet, etc. It outputs concepts of the analysed text and relations between them. In its basic variant, it consists of off-the-shelf components, including a surface dependency parser and the Semafor tool. In its advanced variant, it will consist of UPF's deep parser and an own Semafor-like component.
Innovation	The innovation of the advanced version of the module is in the combination of

[17] http://www.gnu.org/licenses/gpl.html

Description	deep parsing technologies with semantic processing. It can be considered a novel service for the ICT market.
IP rights	UPF is the owner of the IP rights of the advanced variant of the concept extraction module. The IPR rights of the components of the basic variant are with third parties.
Foreseen license	GNU GPL v3[18]
Alternative solution	
Adaptability and extensibility	In order to adapt the concept extraction module to new domains, sufficient language-specific training material is required. The annotation effort depends on the language and linguistic difference of the new domain with the MULTISENSOR domains. An effort of 3 to 6 PMs is realistic.
Maintenance after the duration of the project	The research agenda of the UPF in the next years foresees further activities in the area of parsing. In the context of these activities, the parsing module will be maintained and further improved. Individual maintenance agreements will be set up with interested parties.

3.1.4 Automatic Speech Recognition Component

Module Description	Automatic speech recognition (ASR) is employed within the MULTISENSOR project to provide a channel for analysis of spoken language in audio and video files. It transforms speech signals into a sequence of phonemes and words. The recognition quality depends on different factors such as speaker and channel variability, background noises, audio frequency spectrum, quality of microphone, or difficulty in differentiation between speech and non-speech events. The module will be available for 5 languages (English, French, German, Spanish and Bulgarian).
Innovation Description	The ASR technology used in MULTISENSOR is speaker-independent and uses an open-vocabulary approach (recognition of unknown words based on sub-word units). It employs a series of state-of-the-art techniques: continuous density HMMs for the acoustic modelling; MFCC or PLP feature extraction with support of LDA and VTLN; speaker adaptation by the means of CMLLR; time-synchronous left-to-right beam search strategy for the decoding. The advanced versions will be adapted by using in-domain data, as much as the project partners manage to collect. Another innovation is the integration of the results from the named entities recogniser. This is expected to result in better recognition of proper names in spoken language.
IP rights	IP rights for the software components and resources are with Linguatec. Some of the background components are with RWTH Aachen.
Foreseen license	Commercial licence
Alternative solution	A commercial ASR solution is available from Nuance Communications. As for Open Source alternatives, there are ASR engines such as SPHINX (from Carnegie Mellon University, USA) and JULIUS (from Nagoya Institute of

[18] http://www.gnu.org/licenses/gpl.html

	Technology, Japan). A more recent open source engine is KALDI. However, ASR engines still require access to audio and text data and the generation of domain specific language and acoustic models.
Adaptability and extensibility	Adaptation to other domains requires large in-domain data to train appropriate Language Models. Adaptation to other languages is substantially more complex as it involves extensive recordings of native speakers together with precise transcriptions.
Maintenance after the duration of the project	During the project the ASR component has run on Linguatec servers as this allows for easy access and improvement. Linguatec plans to maintain a demo service for ASR using existing infrastructure. After the end of the project the component is planned to be installed on CERTH premises for sustainability purposes.

3.1.5 Multimedia Concept and Event Detection Component

Module Description	**Functionality, Input/Output:** The multimedia concept and event detection component receives as input a multimedia file (i.e. image or video) and computes degrees of confidence for a predefined set of concepts and events. In order to achieve this, the component incorporates various procedures, such as video decoding (applicable for video files only), feature extraction and supervised classification. **Dependencies**: OpenCV and vlfeat libraries, ffmpeg, ffprobe, Python
Innovation Description	The main innovation of the component is the utilization of state-of-the-art techniques in all the above mentioned incorporated procedures (video decoding, feature extraction, classification) for automatically annotating a multimedia file based solely on visual content. This module could be exploited either as service or standalone and integrated in a media monitoring and multimedia management product.
IP rights	CERTH is the owner of the innovation
Foreseen license	Open Source (GNU General Public License, version 3 (GPL-3.0))
Alternative solution	N/A
Adaptability and extensibility	The component is language independent and given a reasonable amount of time (approximately 3 PMs/10 concepts), it could be trained and extended/adapted to other concepts and events. In order to extend it to additional concepts the module requires annotated data (i.e. images and videos that depict this concept/event and training of the predictive models).
Maintenance after the duration of the project	CERTH is already maintaining a framework (since 2004) for multimedia concept and -event detection, which is used for research purposes and for participation in evaluation benchmarks (e.g. TRECVID). Therefore, CERTH plans to continue supporting and maintaining this module after the end of the project for research purposes.

3.1.6 Machine Translation Component

Module Description	Automatic machine translation (MT) is employed within MULTISENSOR to provide the translation of the summarisation results in the end of the content analysis and summarisation chain and to enable full-text translation on-demand. In the first case, the translations are produced at the end of the analysis/summarisation process and are stored together with the summaries. In the second case, the translations produced by MT provide the input for the text analysis chain and follow the same analysis procedure as the input from original text sources in the required language. The languages covered by MT in the MULTISENSOR project will be English, French, German, Spanish and Bulgarian.
Innovation Description	In MULTISENSOR a Statistical Machine Translation (SMT) approach and the machine learning techniques associated to it are applied. A phrase-based translation model is used, i.e. instead of learning the translation word by word, larger word sequences (up to 7 words) are being taken into account. Thus larger contexts, different word orders in source and target, as well as distant dependencies are taken into account. Another innovation is the integration of the results from the named entities recognizer. This is expected to result in better translation of proper names.
IP rights	The MOSES translation decider is open source. The IP rights for the generated phrase tables and language models are with Linguatec.
Foreseen license	MOSES decoder: open source (LGPL license) Language Models, phrase tables and lexica: commercial license from Linguatec
Alternative solution	As an alternative to statistical MT there are several rule-based MT systems (e.g. Systran, Promt, Linguatec Personal Translator), all of which need a commercial license. With OpenLogos there is also an open source alternative for rule-based MT. However, it currently only supports English and German as source languages.
Adaptability and extensibility	Adaptation to other domains requires bilingual in-domain texts to train appropriate phrase tables. Adaptation to other languages requires in addition substantially larger bilingual training texts. By using English as a pivot language, however, it is possible to support 12 additional translation directions (e.g. French-German, Bulgarian-Spanish).
Maintenance after the duration of the project	During the project the MT component has run on Linguatec servers as this allows for easy access and improvement. Linguatec plans to maintain a demo service for MT using existing infrastructure.

3.2 User and Context-Centric Content Analysis (WP3)

3.2.1 Context Extraction and Representation Component

Module Description	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 its previous versions, the module extracts either from the text or the metadata

	the following information: • Author: an entity responsible for the creation of the item content • Source: an entity responsible for making the item available • Title: a name given to the media item • Keywords: a set of phrases describing the topic of the item • Genre: the style or type of the item • Category: a classification of the item according to its content • Date: a date associated with the creation or availability of the item • Location: a location indicating where the item was created • Language: the language of the content of item 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:** Readers perceive the vocabulary of rich articles as diverse and interesting; rich articles are not written in a plain and straightforward manner.
Innovation Description	The module can be used by a company to extract contextual features appropriate for describing an item such as a news article or a blog. The innovation of this component will lie in the potential use of analytics based on the contextual features.
IP rights	The owner of the innovation is Eurecat.
Foreseen license	The technology has been made available[19] "for non-commercial purposes" or "for research purposes" under the Apache Licence v2.0[20]. The consortium members have access to the source code and the final module and will be able to use the technology to fulfil the requirements of the project.
Alternative solution	N/A
Adaptability and extensibility	For the features extracted from the metadata, the module can be considered language-independent. For the features extracted from the text, we assume that the text would be in English (either originally or translated by the relevant component). However, the module can be extended to other languages given the availability of annotated training data.
Maintenance after the duration of the	Eurecat intends to look further into the technology, and its robustness and commercial attractiveness. Ongoing and future collaborations with European consortia, SMEs, and start-ups, provide an ideal environment for testing the

[19] https://gitlab.bigdata.eurecat.org/ioannis.arapakis/MULTISENSOR-user-and-context-centric-content-analysis

[20] http://www.apache.org/licenses/LICENSE-2.0

project	technology of this module in real-life scenarios. In addition, this will facilitate the maintenance and further advancement of the module beyond the duration of the project.

3.2.2 Polarity and Sentiment Extraction

Module Description	The "Polarity and Sentiment Extraction" module is designed to perform efficient and effective, in-domain, language-independent sentiment classification. More specifically, it involves two main subcomponents: a) Sentimentality (also known as subjectivity) detection - a text segment is classified as either subjective or objective b) Polarity detection - a text segment is classified as either having positive or negative sentiments
Innovation Description	The proposed module offers a tailor-built, domain-specific sentiment analysis solution, trained on specific domains of interest and value to the project, but also extensible to new ones. To this end, we employ a machine-learning approach using a wide range of shallow and syntactic NLP features, so that the classification of news content containing opinions is done with simple yet effective operations.
IP rights	The owner of the innovation is Eurecat.
Foreseen license	The technology has been made available[21] "for non-commercial purposes" or "for research purposes" under the Apache Licence v2.0[22]. The consortium members have access to the source code and the final module and will be able to use the technology to fulfil the requirements of the project.
Alternative solution	SentiStrength[23], SentiWordNet[24]
Adaptability and extensibility	Contrary to lexicon-based, domain-independent solutions, which are not straightforward in how they can be effectively extended to other domains, the proposed module is a tailor-built, ML-based solution that targets specific domains of interest. In principle, the Polarity and Sentiment Extraction module is language-independent and it is possible to extend to new domains, given the availability of annotated text corpora with sentiment features. However, certain techniques and features (e.g., syntactic features) may introduce language dependencies.
Maintenance after the duration of the project	Eurecat intends to exploit further the technology, its robustness and commercial attractiveness. Ongoing and future collaborations with European consortia, SMEs, and start-ups, provide an ideal environment for testing the technology of this module in real-life scenarios. In addition, this will facilitate the maintenance and further advancement of the module beyond the duration of the project.

[21] https://gitlab.bigdata.eurecat.org/ioannis.arapakis/MULTISENSOR-user-and-context-centric-content-analysis

[22] http://www.apache.org/licenses/LICENSE-2.0

[23] http://sentistrength.wlv.ac.uk

[24] http://sentiwordnet.isti.cnr.it

3.2.3 Social Media Mining Module

Module Description	The "Information propagation and social interaction analysis" module is part of the Social Media Analysis Pipeline (SMAP), which consists of a set of processes related to analysis of social network data. The SMAP pipeline performs social influence and interaction analysis on previously crawled Twitter data that includes 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, it outputs two ranked lists of users, one by decreasing order of PageRank and another one by decreasing order of influence in the Independent Cascade model. Moreover, the module computes a Consistency of Sphere of Influence (CSI) metric that quantifies the consistency of information propagation cascades in a social graph for a given user. The CSI metric measures the variability of the set of users influenced by the targeted user on different instances; thus, nodes with higher CSI are preferable: they are more reliable influencers. Finally, the module can determine the optimal set of users (of a social network) to target in order to maximize influence in a given context, such as targeted advertising. The module solves the influence maximisation problem based on spheres of influence and set cover.
Innovation Description	The module can be used to retrieve information publicly available on Twitter users. It provides multiple authority scores computed by leveraging different signals (number of followers, followees, and retweet counts). Similar scores are also available by social media analytics companies such as Klout; however: a) Scores on other platforms are available only for some and not all users b) Scores cannot be directly exported in other systems or knowledge bases c) The exact calculation algorithm of the score is not disclosed, while our computation is transparent.
IP rights	The owner of the innovation is Eurecat.
Foreseen license	The software supporting this module is available under the MIT free software licence[25].
Alternative solution	N/A
Adaptability and extensibility	The module is domain and language-independent and can be expanded in several ways like (1) integration of additional signals in the computation of authority score, and (2) refinement of the interest analysis. Moreover, the module may find applications in other contexts that we plan to explore, such as for example, in the viral marketing context. In this scenario, having the spheres of influence pre-computed and stored in an index, may provide a direct solution to several variants of influence maximization (for instance in the case where different segments of the market audience have different values for a

[25] https://opensource.org/licenses/MIT

	viral marketing campaign). Then when the next campaign is run, and the users have different values, we can again reuse the same spheres of influence. Other examples might include viral marketing campaigns under different types of constraints, such as, e.g., when different nodes have different costs to become a seed. Outside of viral marketing, we may consider the application of the module in contagion problems, or in the vaccination problem[26].
Maintenance after the duration of the project	The proper functionality of the module depends highly on the consistency and backward-compatibility of future versions of the Twitter's API, as well as the legitimacy of the API keys used. Eurecat intends to look further into the technology, and its robustness and commercial attractiveness. Based on these criteria it will be determined whether the technology of this module will be maintained and advanced beyond the duration of the project.

3.3 Multidimensional Content Integration (WP4)

3.3.1 Multimodal Indexing and Retrieval Component

Module Description	**Functionality, Input / output**: The multimodal indexing and retrieval module involves the development of a multimedia data representation framework that allows for the efficient storage and retrieval of socially connected multimedia objects. Specifically, it develops a representation model (called SIMMO) for holding several dimensions of the multimedia information. Moreover, it contains a fusion-based framework for retrieving efficiently the multimodal entities of the multimedia information. **Dependencies**: MongoDB, Morphia framework, Python
Innovation Description	The main innovation of the component is the development of the indexing and retrieval module based on SIMMO (Socially Interconnected MultiMedia-enriched Objects) model, which has the ability to fully capture all the content information of interconnected multimedia objects, while at the same time avoids the complexity of previously proposed models. This product will be used for media monitoring purposes.
IP rights	CERTH is the owner of the innovation
Foreseen license	Open Source (GNU General Public License, version 3 (GPL-3.0))
Alternative solution	N/A
Adaptability and extensibility	The component is language and domain independent. If the data representation requirements that stem from the characteristics of the SIMMO model are taken into account and met, then it is plausible to extend the component to include additional fields of information with relative ease (e.g. 1-2 PMs).
Maintenance after the duration of the project	CERTH plans to maintain a demo service for media monitoring using existing infrastructure. This media monitoring service will be based on this retrieval module and will run on a standalone server.

[26] Y. Zhang and B. A. Prakash. Dava: Distributing vaccines over networks under prior information. In SDM, 2014.

3.3.2 Topic-based Modelling Component

Module Description	**Functionality, Input/Output**: This component includes two basic functionalities: a) Category-based classification and b) Topic-event detection. The component receives as input multimodal features and provides as output a) the degree of confidence of a number of categories for a specific News Item (for category-based classification) b) a grouping for a list of News Items based on the existence or not of a number of topics / events (for topic-event detection). **Dependencies**: R (statistical programming language)
Innovation Description	The use of novel late classifier fusion approaches (for the category-based classification task) and the utilisation of multimodal clustering techniques (for the topic-event detection task). This module could be exploited either as service or standalone and integrated in a media monitoring and multimedia management in order to provide automatic category tagging (classification) and grouping of heterogeneous multimedia (clustering).
IP rights	CERTH is the owner of the innovation
Foreseen license	Open Source (GNU General Public License, version 3 (GPL-3.0))
Alternative solution	N/A
Adaptability and extensibility	The component, if trained accordingly, can receive as input features from any number of modalities. The extension/adaptation procedure of the component to other domains is considered relatively easy (About 2 PMs for any new domain set (regardless of the domain number)). It must also be taken into account that the component receives its input from the indexing structure of the multimodal indexing and retrieval component, therefore any extension/adaptation to other domains should also consider the indexing component. The approach is language agnostic and it assumes that each multimedia document is represented with vectors (e.g. bag of words).
Maintenance after the duration of the project	CERTH plans to maintain a demo service for media monitoring using existing infrastructure. This media monitoring service will integrate this topic-based modelling component and will be maintained after the end of the project for demo purposes. The service can run on a standalone server.

3.3.3 Mapping Discovery and Validation Component

Module Description	**Functionality, Input/Output**: The mapping discovery and validation component deals with discovering and registering in an automated way, valid mappings between the concepts and properties of two ontologies. The ontologies of WP5 are considered for mapping, since a manual discovery of the mappings between the ontologies is a tedious process, especially if the latter are big. The component uses string, lexical, structural and visual similarities combined in a late fusion approach in order to estimate the similarity between two concepts of the ontologies. The mappings are further semantically validated for consistency.
Innovation	The main innovation of this component lies in the development of a weighted

Description	late fusion approach for combining the different similarities for the metrics that are employed, as well as a novel algorithm for ontology alignment, based on visual similarity and Wordnet hierarchy. This application can be used as administrative support for ontology engineers and ICT companies that deal with solutions based on semantics.
IP rights	CERTH is the owner of the innovation
Foreseen license	Open Source (GNU General Public License, version 3 (GPL-3.0))
Alternative solution	N/A
Adaptability and extensibility	The component can be extended by integrating new similarity algorithms that take advantage of different metrics. For instance, in order to include a new existing matching algorithm a work of 2-3PMs would be required. This application is domain and language independent.
Maintenance after the duration of the project	CERTH plans to maintain the component as a standalone application and further improve it since this will be an important tool to be used in future projects relevant to ontology engineering and semantic representation. This component can run on a standalone server.

3.3.4 Content Alignment and Integration Component

Module Description	**Functionality, Input/Output**: The content alignment and integration component deals with the discovery of relations or inconsistencies between content items in the knowledge base. Identifying hidden relations in content enriches the knowledge base, which in turn enables enriched results. On the other hand, identifying inconsistencies between content allows identification of noisy entries in the knowledge base, e.g. false assertions, which should be further considered. Early versions of the module used querying strategies for content alignment, while more advanced methods that employ rule-based approaches have been developed. **Dependencies**: GraphDB or any semantic repository
Innovation Description	The main innovation of this component lies in the development of a content-oriented approach to identify hidden relations or inconsistencies between content items, as well as in the definition of measures for similar and contradictory content search (Framenet-based similarity, concept similarity and named entity comparison for similar content search, concept contradiction and sentiment contradiction for contradictory content search). The approach is generic so that it can be adapted to a variety of domains. It can be used as part of a media monitoring product.
IP rights	CERTH is the owner of the innovation
Foreseen license	Open Source (GNU General Public License, version 3 (GPL-3.0))
Alternative solution	N/A
Adaptability and extensibility	The component can be adapted to run on any semantic repository, while it can be tuned to specific domains of interest. In order to adapt it for a different domain an effort of 2-3PMs is expected. The component is language independent.

Maintenance after the duration of the project	CERTH plans to maintain the component after the end of the project and use it as part of an integrated media monitoring platform for demo purposes. This component can run on a standalone server.

3.4 Semantic Reasoning and Decision Support (WP5)

3.4.1 Data Infrastructure Module

Description	Ontotext provides GraphDB-Enterprise as a semantic data infrastructure layer for the purposes of MULTISENSOR. It is a high performance system implemented in Java, which support storing, querying and processing structured data formatted according to the RDF standards and is packaged as Storage and Inference Layer (SAIL) for the Sesame framework. It is based on the Ontotext's TRREE – native RDF rule-entailment engine. GraphDB is world leader among the structured data repositories in terms of volume of data and loading/inference speed. One of the main advantages is the in-memory reasoning implementation- the content of the repository is loaded and maintained in the main memory, which allow for efficient retrieval and query answering.
Innovation Description	Ontotext will implement and deliver four different types of reasoning, which currently are not supported by GraphDB. There is no product on the market that supports these inference techniques together. • Parallel inference • Hybrid reasoning • SPARQL-MM • GeoSPARQL
IP rights	Owner of these innovations is Ontotext.
Foreseen license	GraphDB is available under an RDBMS-like commercial license on a per-server-CPU basis.
Alternative solution	As alternative users can use Sesame, which is an open source framework for storing, querying and analysing RDF. It is implemented in Java by Aduna and is available under the GNU Lesser GPL license[27].
Adaptability and extensibility	GraphDB can be adapted to work on the cloud very easily. ONTO already has such kind of project named S4, where we offer Data-as-a-Service. GraphDB has the ability to work with many different languages, so all the innovations support these languages tool.
Maintenance after the duration of the project	These innovations are part of GraphDB, so they will be supported, maintained and improved after the end of the project by ONTO.

3.4.2 Semantic Representation Infrastructure Management System Module

Description	GraphDB Workbench is provided as Semantic representation infrastructure management system. It is our recommended web-based administration tool. The user interface is similar to the Sesame Workbench web app, but provide

[27] http://www.gnu.org/copyleft/lesser.html

	more functionalities, better user experience and clean design. Some of the additional features are: ● Query monitoring with the possibility to kill a long running query. ● Better SPARQL editor based on YASGUI[28] ● Connectors' administration presented only in the enterprise edition. GraphDB Workbench can be used as a SPARQL endpoint and as an administration tool for managing repositories, executing queries and updates. It contains user management module which allow to set different kind of restrictions over the repositories like read/write for different user groups. In addition to the SPARQL queries, you can perform Full Text Search over the data, but with arrangement that such index exists.
Innovation Description	The main innovation here is the new improved UI. This new UI will give us great new functionalities and management capabilities compared with Sesame Workbench.
IP rights	Owner of this innovation is Ontotext.
Foreseen license	GraphDB Workbench is part of the GraphDB distribution so is available under an RDBMS-like commercial license on a per-server-CPU basis.
Alternative solution	As alternative users can use Sesame Workbench, which provides all basic functionalities for managing, querying and updating the semantic repository. It is open source and is available under the GNU Lesser GPL.
Adaptability and extensibility	Currently the GraphDB Workbench is available only in English language. GraphDB Workbench can be easily extended depending on the specific use case.
Maintenance after the duration of the project	As an important part of the GraphDB distribution, ONTO will continue to improve and develop GraphDB Workbench after the end of the project.

3.4.3 Decision Support System Module

Description	ONTO is going to implement restful web services to expose the functionalities of the decision support system, which is built on the top of the semantic repository. This first version of the system will be very basic and will provide limited capabilities to support the third use case – Internationalization. In the process of development we identified the most important statistical indicators from The World Bank and Eurostat, which are well described in D3.2. Based on the data in this indicators and datasets like DBpedia and Geonames we can develop specific SPARQL query templates. We also added support for Google charts on the MULTISENSOR SPARQL endpoint[29], so the information of the queries can be very easily visualized. Another addition to the whole system but especially to the decision support is the implementation of GeoSPARQL standard which will help the users to work with geospatial objects. These template queries together with the restful web services; the GeoSPARQL support and the Google chars are in the core of our first version of the decision support system.

[28] http://laurensrietveld.nl/yasgui

[29] http://multisensor.ontotext.com/sparql

Innovation Description	• Develop SPARQL query templates which will help to get specific results depending on the decision support use cases • Develop new restful web service, to retrieve the results generated by the SPARQL templates. The queries are identified by ID • Develop new UI functionalities to support Google charts, so the results of the queries can be easily visualized, which will help in the process of decision making.
IP rights	Owner of these innovations is Ontotext.
Foreseen license	As part of the GraphDB Workbench it will be available under an RDBMS-like commercial license on a per-server-CPU basis.
Alternative solution	As an alternative, one can use Sesame Workbench or Yasgui[30], but neither of these alternatives has the needed functionalities and the capabilities ONTO is going provide.
Adaptability and extensibility	Currently GraphDB Workbench that takes the role of management application is available only in English language. GraphDB Workbench can be easily extended depending on the specific use case, but precise estimation can't be given because it highly depend by the specific use case.
Maintenance after the duration of the project	As an important part of the GraphDB distribution, ONTOwill continue maintaining, improving and developing GraphDB Workbench and its functionalities after the end of the project.

3.5 Content Summarisation and Delivery (WP6)

3.5.1 Extractive Summarisation

Module Description	The module operates on plain text or collection of texts, assessing the relevance of individual sentences to the summary accordance to a series of quantitative metrics. The most relevant sentences are selected and delivered to the user. The core of the module is the SUMMA summarization toolkit.
Innovation Description	The innovation of this module consists, first of all, in the metrics developed in the context of MULTISENSOR. From the perspective of the market, this module provides tuning facilities to an existing service.
IP rights	IP rights of the SUMMA summarization toolkit belong to a third party (Dr. Horacio Saggion)
Foreseen license	Proprietary license
Alternative solution	The abstractive summarisation component in section 3.5.2.
Adaptability and extensibility	To adapt the extractive summarization module to new domains, training material consisting of a sufficient number of sample summaries (along with the original texts) of high quality is needed. No estimation of the effort of the compilation of such summaries can be given, since it depends on existence of quality-annotated summaries for specific domains and languages. The effort for

[30] http://yasgui.org/

	retraining of the summarization module is minimal.
Maintenance after the duration of the project	The research agenda of the UPF in the next years foresees further activities in the area of concept extraction. In the context of these activities, the concept extraction module will be maintained and further improved by UPF. No targeted maintenance is foreseen. The service can run in a standalone server.

3.5.2 Abstractive Summarisation

Module Description	The module operates on the ontological representations of the content distilled from a given text and outputs a summary of this text in one of the languages of MULTISENSOR. It consists of three main components: (1) content selection, (2) discourse structuring, and (3) text generation. For sentence generation within the third component, a rule-based, statistical or hybrid generator can be chosen.
Innovation Description	The module possesses an innovative architecture (in that it is, de facto, a genuine content-to-text generator) and innovative realizations of the individual components. The module is thus a novel service for the ICT market.
IP rights	UPF possesses the IP rights of the abstractive summarization module.
Foreseen license	NU GPL v3
Alternative solution	N/A
Adaptability and extensibility	The content selection component of the module can be operated on any RDF/OWLIM content structures without the need of adaptation. The discourse structure module will require some adaptation to new domains; it is, however, language-independent. The adaptation of the rule-based text/sentence generator presupposes the compilation of language-specific grammatical and lexical resources at different levels of the linguistic description. The size (and thus coverage) of these resources depends on the nature and verbosity of the targeted summaries. The adaptation of the stochastic sentence generator requires the annotation of text corpora with linguistic structures at different levels of abstraction (surface-syntactic, deep-syntactic, and semantic) for each language that is to be covered by the summarizer (this was done for English, French, German, Spanish in the context of the project). For an acceptable performance of the stochastic sentence generator, training treebanks with between 3,500 sentences (with a very high quality of annotation) and 10,000 sentences are needed. The estimated cost is about 1PM per 1000 sentences of high quality annotation.
Maintenance after the duration of the project	The research agenda of UPF in the next years foresees further activities in the area of abstractive summarization. In the context of these activities, the abstractive summarization module will be maintained and further improved by UPF. Individual maintenance agreements will be set up with interested parties. The service can run in a standalone server.

3.6 System development and integration (WP7)

3.6.1 Crawlers and data channels infrastructure

Module Description	The Crawler component is the process responsible of collecting source material to be used by the MULTISENSOR platform. It runs regularly on a set schedule, going over a set of manually selected content sources, and sends the retrieved material through the analysis pipeline for extraction of intelligence. Sources include print and online news sites, social media, and audio and video sources (e.g. textual news, news with multimedia content and social media). Controlled by the cron job, raw data is crawled and is normalised into common JSON format and stored in the appropriate repository (ElasticSearch index). The crawler depends on two sub-components or "Collectors", the Media Collector and the Site Collector. In fact, the crawling process collects much more articles than are processed per day due to limitation of the CEP performance.
Innovation Description	The crawler infrastructure is an innovative product that may be used as a framework for implementing other "Collectors", providing a convenient system for scheduling the crawling process and converting the harvested content to a JSON output. The crawling tasks are performed using standard techniques and off-the-self tools. In that sense there is little room for innovation.
IP rights	everis, PR and EURECAT
Foreseen license	Open Source (BSD) for the Crawler infrastructure (everis) and Media Collector (BM) and proprietary rights for the Site Collector (PR).
Alternative solution	With **Nutch**[31] EURECAT has setup an open source crawler. Therefore, there are no proprietary rights involved. Other well-known open-source crawling applications are: **Scrappy**[32], and **Heritrix**[33]. Media currently delivered through PRs proprietary crawlers could be substituted through such an open source crawling solution.
Adaptability and extensibility	This component may be adapted to other "Collector" sub-components. If these respect the interface of the Crawler, integration should be perfectly feasible.
Maintenance after the duration of the project	Regular performance monitoring of EURECAT crawling operations account for 2 PD per month. Media currently delivered through PRs proprietary crawlers are based on PRs existing crawling infrastructure. This infrastructure, since simultaneously used in PRs on-going business operations will be maintained in any case for demonstration purposes. EURECAT crawling infrastructure is relevant to web information retrieval and it is therefore expected that the crawling module will be maintained to be used in future research projects.

[31] http://wiki.apache.org/nutch/FrontPage

[32] http://scrapy.org/

[33] https://webarchive.jira.com/wiki/display/Heritrix

3.6.2 Content Extraction Pipeline (CEP)

Module Description	The Content Extraction Pipeline (CEP) is the set of the software components that permits to process digital content through several services in order to extract the relevant information. All the extracted information is structured according to the ontological framework in RDF format and stored in the semantic repository. Before storing the SIMMO in the semantic repository, the CEP output is validated with the RDFUtils library. The pipeline can be used to process textual, multimedia and social content. For example, to process all the textual and social content, the pipeline is organised as a sequence of all the offline services developed in the project (language detection, machine translation, Named Entity recognition, concept extraction, dependency parsing, relation extraction, context extraction, category detection, sentiment analysis, entity linking, content alignment, etc.). To process the multimedia content, other services (concept and event detection and audio speech recognition) are automatically triggered to analyse the images and videos. For articles that are available in four different languages (French, Spanish, German and Bulgarian) they are automatically translated and processed in English. But the multilingual pipeline can use some services (to process them in their original language).
Innovation Description	The CEP testing tool is the web application that allows configuring and customizing the pipeline. With this tool, it is possible to test one service or more services and to process an article that was crawled by the MS platform or by loading custom text. The result is the extracted knowledge for each service that was selected. For this reason, the current version of the CEP can be considered as a technical framework that is able to produce Linked Open Data from unstructured content.
IP rights	everis and all the technical partners(CERTH, ONTO, eurecat, UPF, LT)
Foreseen license	Open Source (BSD) for the CEP (everis) and some fees to use the API to process data and structure the content.
Alternative solution	DataLift (French national project) is an alternative solution to structure the data as Linked Open Data but not so advanced NLP and knowledge extraction service were used in this project.
Adaptability and extensibility	The current version of the CEP is not easily adapted and extensible. The generic approach to extend the CEP integrating new services will require several improvements such as the possibility to select ontological data model that should be used to produce the output. The ontology used to structure the content should be compliant with the other services used in the pipeline. Also, the output of the new service should be automatically validated.
Maintenance after the duration of the project	The adaptation of the CEP could require an additional development that would imply an extra cost. For example, extra developments such as generic plug in mechanism to insert new services, customisation with parameters (service selection, language selection, export of the outputs, portfolio to upload multimedia content to be processed, etc.) would be needed. In terms of maintenance, it is not possible evaluate a reasonable estimation of time or cost before the adaptation of the CEP.

3.7 Final System (WP7)

Description	The final system includes the MULTISENSOR platform integrating the aforementioned components and modules. Specifically, the final system includes the 3 Use Cases Portals, integration services, business services, and associated repositories. The designed architecture follows the latest trends and developments in this area, by proposing a decoupled, layered, service-oriented system, and the use of lightweight RESTful API for accessing the different online services. The central infrastructure has also been designed to run in a cloud environment. The system makes use of JSON objects, RDF data, RESTful APIs and repositories such as Elastic Search, MongoDB and GraphDB has become an excellent opportunity to develop state-of-the-art solutions for integrating these elements.
Innovation Description	The final system includes 3 innovative products: a) portal for journalism, b) a portal for commercial media monitoring and c) a portal for SME internationalisation decision support. According to the market analysis performed in D8.2 and in chapter 2 it is clear that the 3 products have a very high innovation potential. With respect to a) and b) the MUTLISENSOR platform is an innovative product that allows for multilingual and multimedia search across several different sources, including both quick an in-depth analysis support through summarisation, sentiment analysis, named entities recognition, concept extraction, entities along with related content just as there doesn't seem to be anything as comprehensive in one single interface. With respect to c), MULTISENSOR platform comprises an innovative product that can provide, social dimension analytics like influential user or community detection for particulate sector or product, decision support in order to assess potential international investments by considering internationalisation indicators and by providing unified access to multilingual content.
IP rights	All partners
Foreseen license	The portals are open source (BSD), however each component is governed by each specific license as described above.
Alternative solution	As the system has been designed and developed originally following very specific requirements, no substitute products exist to our knowledge.
Adaptability and extensibility	The SOA approach that has been followed allows for further functional enhancements of the system, provided the interface standards defined in the project are respected. That is, new services and functionalities should adapt to the system, not the other way. With regard to adaptability to new domains or languages, the system depends on all the aforementioned modules for which specific details have been provided in the previous tables with respect to their extensibility and adaptability. The fact that the infrastructure is cloud-based also makes scalability possible without major challenges. On the other hand, it is also possible to implement the system in a local infrastructure. This approach will not limit the system's functionalities in any way, provided the infrastructure is properly dimensioned.

	However, the performance may be slower due to many factors like: network capacity, server potentiality, average content size especially applicable to multimedia, etc. During the project, a cloud infrastructure approach has been adopted. The main reason for this is that a flexible solution is needed to accommodate the system's evolving requirements. At the beginning of the project, it was rather difficult to foresee the required capacity, and therefore a flexible solution was essential. A cloud infrastructure can be scaled up and down easily and almost instantly. This has the benefit of aligning infrastructure and its associated costs with its usage. Moreover, everis does not have data centres that could fulfil the project's requirements. It is beyond the scope of everis business to own such infrastructures, and these are always subcontracted from third parties. All the above also applies to the commercialisation phase of MULTISENSOR. The user requirements and the load these may put into the system are unknown, and therefore a scalable solution is still required. A cloud infrastructure may be optimised, for example, to a situation where the system is put on "stand-by" i.e. neither crawling nor processing takes place. In this scenario, the system's capacity needs would be much lower than during the duration of the project, and cloud architecture will ensure that the infrastructure is not over-dimensioned. It is important to note that some of the services – e.g. Named Entities recognition, machine translation, GraphDB database API, abstractive summarization, the SMAP services – reside in the partner's premises and therefore outside the main platform. However, the maintenance of each service/module has been discussed above in detail and in the case of commercial products alternative open source solutions have been suggested.
Maintenance after the duration of the project	The maintenance costs of the cloud infrastructure are two-fold: • Direct infrastructure costs (Amazon EC2): These amount to approximately 2.250€ per month • Management costs (everis): These may be estimated as 3 PD per month After the end of the project, CERTH (as coordinator) plans to move the platform into their own servers in order to keep it running for demonstration and exploitation purposes. Since the implementation will move to a standalone server, a lower performance is foreseen due to network bandwidth capacity. In the case that all services will not be available, CERTH will integrate at least the most important services that will allow the platform to run (e.g. SIMMO, summarisation, content (NE, concepts, sentiment) extraction pipeline). The MULTISENSOR system builds upon 3 complementary (but also alternative) indexing mechanisms (elastic search repository, knowledge base, SIMMO-based representation), which means that platforms could be still functional even if we employ only one indexing structure.

4 CONCLUSION

In the last version of MULTISENSOR exploitation plan, reported in this deliverable, the trend and market analyses for all the three use cases namely the journalistic, the media monitoring and the SME Internationalization use case were described.

Figures on the digital media market were presented and analyzed per use case to provide an in-depth understanding of the market and MULTISENSOR's positioning in it. The potential competitors were listed to discuss the strengths and weaknesses of existing solutions per competing market.

To conclude, we iterate on the fact in an ever evolving digital world, the need for relevant and analysed data for corporate decisions and management is all the more important. Thereby, ICT based tools could hold the promise of providing the needed breakthroughs. MULTISENSOR outcomes obtained in the course of three years could set the standard for technology offerings for the digital media industry starting from SME Internationalization and Decision Support System. A major opportunity in the news sector for MULTISENSOR is to help journalists make quicker decisions on relevant articles and topics, covering more news items in less time. As per the media monitoring field, separate functional modules that have been developed in the context of the project are expected to transform day-to-day operations of data analysts upon being integrated in their media monitoring workflow.