Scalable Understanding of Multilingual MediA (SUMMA)

http://www.summa-project.eu

H2020 Research and Innovation Action
Number: 688139
D8.3 – Dissemination and Exploitation Plan and Initial Report

Nature	Report	**Work Package**	WP8
Due Date	31/07/2017	**Submission Date**	31/07/2017
Main authors	Dominic Tinley (BBC)		
Co-authors	Peggy van der Kreeft (DW)		
Reviewers	Steve Renals (UEDIN)		
Keywords	dissemination, exploitation, communication		
Version Control			
v0.1	**Status**	Draft	26/07/2017
v0.2	**Status**	Reviewed	28/07/2017
v1.0	**Status**	Final	31/07/2017

Contents

List of Figures

Abstract

This deliverable D8.3 relates to Work Package 8 “Dissemination and Exploitation” and provides an overview of the plans in terms of dissemination and exploitation for SUMMA and reports on the achievements for the reporting period of M1-M18.

This report is divided into two main parts: section 2 (Dissemination) and section 3 (Exploitation). These relate to Tasks 8.1 and 8.2 respectively.

1 Introduction

This deliverable D8.3 relates to Work Package 8 ”Dissemination and Exploitation” and provides an overview of the plans in terms of dissemination and exploitation for SUMMA and reports on the achievements for the reporting period of M1-M18.

It is a public document and, whilst the initial target audience of this deliverable is largely internal to the project, since it is a planning and results document, the real target audience of the dissemination strategy and its individual actions are the wider scientific, industrial and general audience who can be perceived as stakeholders and are approached and encouraged to take an interest in the project. All SUMMA public deliverables are made available publicly through the SUMMA website.

This report is divided into two main parts: section 2 (Dissemination) and section 3 (Exploitation). These relate to Tasks 8.1 and 8.2 respectively.

Section 2 (Dissemination) shows plans and efforts to inform and inspire other researchers and potential users of the SUMMA platform about the project's intentions and results. SUMMA wants to establish feedback loops and engage potential users, early adopters and technology providers right from the beginning of the project. The project expects to contribute to and inspire other projects with the goal of building networks and showing that such a media monitoring platform is easy to integrate, usable and efficient.

Section 3 (Exploitation) outlines the activities aimed at successful exploitation of SUMMA. Two foundation tasks are essential for supporting exploitation activities: convening an Exploitation Committee and taking charge of IPR Management. Building on this foundation there are three main ways in which SUMMA can be exploited: Knowledge (for new applications), Component-Based Exploitation (for other applications) and Multi-Strand Exploitation (SUMMA as a whole). This document outlines progress in each of these areas.

2 Dissemination Plan and Initial Report

This section of the deliverable relates to Task 8.1 (Dissemination), focusing on providing visibility of the project results in the scientific community, the broader community of users and stakeholders and other related research and innovation projects.

2.1 Overview

This section describes the different dissemination channels, identifies planned dissemination tools, provides a survey of dissemination events that promote related fields of research and suggests journals, press, and mass media channels reaching the targeted audiences.

The purpose of this section of the document is to provide a project dissemination and communication strategy by highlighting targeted groups and communities, define internal dissemination/communication guidelines and procedures, outline the foreseen channels, and report on the efforts in the first year and a half. Dissemination of the project is a collaborative effort of all project partners and this document describes available tools and partners' responsibilities.

This part will:

- Document actions related to expected dissemination and communication including their priorities, responsibilities and outcomes
- Guide the project's awareness and engagement programme
- Be a living document, with two reporting iterations on dissemination efforts
- Ensure a plan for post-project dissemination and communication

Section 2 has three main parts:

- Description of the planned dissemination and communication strategy
- Overview of the planned tools, materials and activities
- Reporting of initial dissemination results

Thus, after outlining the dissemination plan in terms of strategy and the planned tools, materials and activities, a third chapter provides lists and tables depicting the status of the actual dissemination results to date. These lists and tables will be updated in the next iteration of the deliverable (D8.4) to report subsequent dissemination and communication activities and results.

2.2 Dissemination Plan: Strategy

SUMMA's dissemination and communication strategy is briefly outlined here. It provides the guidelines and procedures for communicating internally and externally and disseminating the results of the project.

2.2.1 Strategy and Tactics

The project's communication strategy lays down how we communicate internally, with peer researchers and other stakeholders, and with the public at large - and how this communication can be implemented efficiently. It is important to ensure the entire consortium is kept informed of user requirements and priorities, of the development status and details, the challenges and efforts to overcome these, at different levels (technological components, integrated protoype, UX), and user testing and feedback. Equally important is communication towards other stakeholders, the research community, related project groups, the media production and monitoring world, to open up our efforts and achievements to professional communities outside of the consortium and have them profit from the results and provide their feedback. Finally, in order to ensure continued use of the platform, potential end users of the integrated SUMMA solution or any of the technological components (and possibly customisation) need to be made aware of the advantages - and restrictions - of the outcome. The strategy describes what channels and methods will be used for this purpose.

The dissemination strategy determines what kind of events will be covered and organised, what materials will be created to describe and show project results, what publications are targeted.

For simplicity's sake, this report uses the term "dissemination" to include the concept of communication.

The consortium aims to disseminate the project goals, research, results and experiences to industrial communities (SME and Industry), to academic and research institutions, as well as to the generic audience interested in the project.

Project results will be promoted and disseminated during the entire project, as an appropriate prerequisite for a successful exploitation, and at the end of the project in order to engage its stakeholders. The dissemination is both a collective activity managed by the entire consortium and an individual set of actions handled by each single partner on a local level. All partners are aware that a broad dissemination of results carries a great importance and it is committed to allow access to the results achieved in the project to various kinds of audiences and users: information and research results will be considered to be available to the public unless the consortium decides otherwise.

2.2.2 Target Groups

The project has established a central digital communications and promotional committee, which is responsible for the definition of the target audiences for communications, and the creation, management and execution of a rolling communications strategy. The consortium plans diffusion of information on the project to its diverse target groups encompassing:

- Peer research groups in academia in the EU
- Peer research groups in academic globally
- Peer research groups in other H2020 projects
- Academic periodicals (note this is different to, but allied with, the publication of academic papers)
- Specialist press in research and technology sectors

- Specialist correspondents for science and technology in general press
- International broadcast technology users
- Semantic analysis technology users
- Key stakeholder groups for all project partners
- Diverse industry sectors: broadcasting, media monitoring, translation, business intelligence

- Policy makers and interest groups
- Users of the SUMMA platform, including journalists, editors, monitors, analysts
- Business intelligence consumers
- Educational outreach audiences (the project seeks to encourage young people into scientific careers by showcasing the interesting challenges tackling by SUMMA)

2.3 Dissemination Plan: Tools, Materials, Activities

A module-based "dissemination kit" has been developed, adapted to changing requirements of events, target groups and communication channels. This kit contains a set of key visuals of the SUMMA project, including a project website, providing information about the project mission and objectives, descriptions of the application scenarios and use cases, and covering news and results, to support the consortium's presence at events, to promote the project's own events, and to engage in a dialogue with user group members. Academic and industrial activities will engage stakeholder communities in the project outcome.

2.3.1 Project Identity

The project identity is expressed by consistent use of the project logo in all its communication. It creates a direct visual recognition of the SUMMA project. Therefore, the logo is simple, transparent and easy to recognise. The logo obviously appears on the project website, as well as all other external communication, including deliverables, flyers and posters.

In addition to the logo, flyers, posters and website have a common style and appearance, to increase project visibility, recognition, and familiarisation.

2.3.1.1 Contact Info The standard contact information to be used on dissemination material for the SUMMA project is:

- SUMMA Project Website: summa-project.eu
- SUMMA Project Coordinator
  - Steve Renals, Professor of Speech Technology, School of Informatics, University of Edinburgh, UK
  - s.renals@ed.ac.uk

- http://homepages.inf.ed.ac.uk/srenals/

The above contact info will be used on the project website, flyers, generic poster, etc.

Of course, partners publishing research on or demonstrating their components or communicating project results within their network can provide their own contact information, using their names or business cards. All key team members appear on the project website, with their name, short description and picture. Obviously, the profile of all partner organisations, with their logo and link to their company website, also feature there.

2.3.2 Project Fact Sheet

A project fact sheet was produced at project start, will be updated throughout the project to take account of project progress and achievements.

2.3.3 Website and Blog

A project website, summa-project.eu was created and made available from the early stages of the project, providing information about the project mission and objectives, descriptions of the application scenarios and use cases, and covering news and results, to support the consortium's presence at events, to promote the project's own events, and to engage in a dialogue with user group members.

The website was made such that it is both secure and easy, fast and efficient to add content. It is updated regularly and contains a blog section with short news items or announcements as well as longer articles. The project website is the primary channel of communication towards the outside world.

2.3.4 SUMMA Wiki

A wiki was set up by the coordinator as a major communication channel within the consortium. It is used as a repository for guidelines and procedures, contains contact information, mailing lists, links to other repositories used in SUMMA, describes status and progress in the different work packages, and has the agenda and minutes of all meetings. It has a JIRA section which serves as a feedback reporting tool for user evaluation. In short, it is the primary internal reporting and communication channel for the consortium.

2.3.5 Slide Pack

A generic project presentation was produced at the early stages, and updated throughout the project to take account of project progress and achievements. This presentation can be used for short-term requests for presentations and as a model to create customised presentations. It will also be added to the project website.

This will be supplemented by demonstration presentations which will be compiled in the course of the project.

2.3.6 Social Networks

The project will also communicate via social networks; the current plan is to focus on LinkedIn (via a SUMMA group) and Twitter. We shall also communicate via the web and social media presence of organisations such as LT-Innovate. Websites of the individual partners will also disseminate progress and key results from the project.

2.3.7 Printed Media

We have prepared a set of printed or customised off-the-shelve dissemination materials, including flyers and large-scale posters designed and produced specifically for different requirements and target groups.

2.3.7.1 Generic Posters A generic poster was produced at the start of the project depicting the overall objectives and a visualised concept, listing consortium partners. This poster can be used for poster sessions or other dissemination events. The poster will be updated in the course of the project to reflect the current status of the results.

2.3.7.2 Scientific Posters In addition to the generic poster, scientific posters will be created by the technical partners, focusing on specific technologies, components, or showing the details of the integrated platform. These will be customised, depending on the technology, event or occasion.

2.3.7.3 Flyer Project flyers will be generated to provide a very simple overview of project objectives and achievements suitable for a non-specialist audience, and to provide overviews of specific project outputs – focusing on the use cases. These will be used at smaller and larger events to raise visibility and remind people of the project after the event.

2.3.7.4 Brochure User brochures, serving as a guideline for use of the SUMMA platform will be created to encourage potential users and to familiarise them with the system. They will contain access details and will also be used during user evaluation sessions and contain a script for the users to follow.

2.3.7.5 Banner A project banner will be created and printed in the second half of the project, to raise visibility at stands during major events.

2.3.8 AV Media

Some AV dissemination material will supplement the printed media, including screencasts and a promotion video. These will be produced in the second half of the project, when the integrated platform is somewhat mature. Online walkthroughs in the form of screencasts will be produced at various stages, showing the workflow as used in a media environment, and will be updated for different versions of the prototypes. Also at least one promotion video is planned, ensuring visibility of the project to key industry sectors and the general public.

2.3.9 Dissemination Events

Described below are the type of dissemination events that will be covered by the project, either through attendance, presentations, publications, panel discussions, poster sessions, etc. SUMMA aims to report its findings in at least 3 conferences/workshops per year. Each attendance is accompanied by a conference report to ensure communication towards the target audience and stakeholders as well as other consortium partners.

The SUMMA website will list all dissemination events that will be covered in the project's lifetime and, in addition, announce major events that are closely related to and of particular interest to SUMMA. Whenever possible, events covered will be announced briefly as an upcoming event in the blog section, then reported on in some more detail, be included in the list of events. Presentations or papers presented will be made available on the project website to the extent possible.

2.3.9.1 Academic Events In the early phase of SUMMA, the consortium will outline which events (conferences, fairs, events, workshops, summer schools) are planned to be attended or organised by consortium partners. Dissemination will be done throughout the project duration. Topics relate to specific project domains, particular research areas and/or work packages, targeting respective research communities as well as industry. All partners will be activated to transfer the knowledge being gained throughout the project to their internal and external communication links. Based on the identified dissemination target groups, each partner will be associated with the task to approach certain target groups and to focus on suitable activities. The following are potential academic events considered for SUMMA dissemination activities: Scientific workshops in NLP e.g EMNLP (emnlp.org), computational linguistics e.g ACL (aclweb.org/website/acl), machine translation e.g. EAMT (eamt2017.org), speech technology e.g. Interspeech (interspeech2017.org), multimedia indexing e.g. MediaEval (multimediaeval.org)

2.3.9.2 Industry Events Industry and innovation-oriented events will be covered to raise visibility of the project in the industrial world and to attract potential users. Examples are:

- media and broadcast conventions e.g. IBC (ibc.org)
- innovation-oriented events, such as LTInnovate (http://www.lt-innovate.org/summit) or STI - Science, Technology and Innovation Indicators (https://sti2017.paris/)
- events focusing on journalism and media, e.g. Scoopcamp (http://scoopcamp.de/) and GMF, DeutscheWelle's Global Media Forum (http://www.gmf.de), EBU and ARD meetings, workshops and conferences.

SUMMA will engage with innovation communities at all levels: local, regional, national, European, and global. At the European level we shall work with CITIA (the Conversational Interaction Technology Innovation Alliance; citia.eu), LT-Innovate (the European Association of the Language Technology Industry, lt-innovate.eu), and META (the Multilingual Europe Technology Alliance, meta-net.eu). We shall engage with these, and other innovation communities, through events such as the annual LT-Innovate Summit and the META forum, as well as through collaboration on various policy initiatives, roadmapping activities, and white paper development. The EU ICT Conference is one of the major European events in the field, and will be an excellent venue to

communicate the achievement of the project to the European public, scientists, technologist, entrepreneurs, and decision makers. We target the 2018 ICT conference for a SUMMA exhibit. We will also work with CITIA, LT-Innovate, and META to organise relevant workshops at the ICT Conference.

More locally, all the partners are involved in innovation activities. In Edinburgh, the annual SICSA DemoFest brings together researchers, developers, and entrepreneurs (http://www.sicsa.ac.uk/events). There are also regular Tech Meetups for startups and potential startups local to most of the partners. UCL will co-organize, on a yearly basis, the Workshop on Automated Knowledge Base Construction (AKBC) (http://www.akbc.ws), which is the premier forum for researchers on knowledge base construction in both academia and industry, and it is perfectly aligned with the research of this proposal, in particular for WP4.

2.3.9.3 Workshops The SUMMA consortium plans at least two major showcases. In Year 2, a project event will be either organised separately or as a workshop within one of the relevant community events. This event will showcase the research results in particular. This could be done within the framework of the LxMLS summer school. In addition, as mentioned above, other workshops will be organised, as the AKBC workshop mentioned above. These may focus on a particular work package, technology, component, or the integrated platform as a whole. In Year 3 of the project, an event will be jointly coordinated, focusing on application scenarios, use cases and demonstrable project results. It will primarily address the media community and other end users rather than technologists. The focus will be on demonstrating the achievements in the SUMMA application context and discuss planned exploitation and further opportunities. Details of concrete plans are described in the activities tables in section 2.4.1.

2.3.9.4 Hackathons We shall organise a number of innovation intensives / hack events in SUMMA, with the explicit aim of diverse exploitation of the SUMMA platform. Each hack event will focus on a particular aspect, such as media monitoring, ASR, language technologies or Data Journalism.

These innovation intensives will (1) inform participants about the outcome of the SUMMA project, and the software made available for exploitation, (2) provide new technology transfer opportunities through gathering and mentoring a selection of motivated young developers and entrepreneurs around the SUMMA output, and (3) provide a testbed for exploring the flexibility of the SUMMA platform.

2.3.9.5 User Days SUMMA will organise user days, Innovation Intensives and other events to establish a SUMMA User Group covering a broad range of stakeholders with interests in the SUMMA platform and use cases. An initial set of User Group members has been established, and WP8 will in part be concerned with extending and strengthening the User Group. Initial members include: Al Jazeera Media Network (aljazeera.com); Brandwatch (brandwatch.com); Cortex Intelligence (cortex-intelligence.com); Imaxin Software (imaxin.com); Media Capital Digital (mediacapital.pt); pressrelations (pressrelations.de); Quorate Technology (quoratetechnology.com).

2.3.9.6 Collaboration Activities Concertation meetings will be facilitated where certain themes relevant to SUMMA and to a number of other European projects are discussed. The aim is for

SUMMA to communicate with other ongoing European projects on a regular basis, to exchange ideas and concepts, to exploit their results and finding, to broaden the potential user base. This includes, but is not limited to, some projects which have already been identified as covering related technologies, such as

- Mixed Emotions: Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets (2015–2017)
- Multisensor: Mining and Understanding of multilinguaL contenT for Intelligent Sentiment (2013–2016; multisensorproject.eu)
- Reveal: REVEALing hidden concepts in Social Media. (2013–2016; revealproject.eu/)
- CODAM: Content-based Digital Asset Management (funded by InnovateUK). (October 2014-September 2016)
- EUMSSI Event Understanding Through Multimodal Stream Interpretation. (2013–2016; eumssi.eu)
- Newsstream (German research project) (2015–2017; newsstreamproject.org/)
- xLime - Cross Lingual Cross Media Knowledge Extraction (2013–2016; xlime.eu)
- MMT - Large-scale commercial online translation infrastructure (2015-2018)
- The Alan Turing Institute (UK) (2015–2020)
- speech.media (Google DNI prototype project) (2017)
- news.bridge (Google DNI prototype project) (2017-2019)

2.3.9.7 Other In addition to the activities mentioned above, other types of events can be included, such as interviews, panel discussions, in-house customised presentations, networking.

2.3.10 Papers and Publications

Both technical and user partners will engage actively in dissemination through the publication of papers to inform the academic and industrial/media world. The goal is to encourage joint publication as much as possible, with collaboration of different project partners and showing joint efforts and results. Some publications will involve all SUMMA project partners.

All publications that will be produced in the project's lifetime will be referred to on the SUMMA website (if possible, they will also be made accessible from there).

2.3.10.1 Conference Papers Conference papers are a major dissemination channel, in particular for academic partners, to report on their research findings. As mentioned above, joint publication involving several partners is encouraged, to stress the enriched outcome through collaboration. Each partner will suggest and select a number of conferences in their specific area, spread over the entire project duration. All technology covered should be presented in conference papers.

2.3.10.2 Journals and Magazines The publication of articles in journals helps to reach a large and focused audience. To get a paper published in an international refereed journal supports the overall goal to have an impact on the topics covered by SUMMA. Other similar publication channels, such as book sections and company magazines, will be pursued. All partners in the project will engage in this effort.

2.3.10.3 Press Releases Over the course of the project we envisage several press releases to be issued by different project partners, announcing major achievements. This depends on internal organisational procedures and the authorised use of press releases. Many organisations have moved from press releases to newsletters, blog and other similar information channels.

2.3.10.4 Blogs The SUMMA blog on the project website provides brief news reports on issues related to SUMMA coverage (technologies, media applications, media monitoring, workflow, related project achievements, etc.), as well as longer technical articles focusing on a specific technology, a component, a research area, user findings, etc. Each partner will contribute to the blog with at least 1 long article, so that it will cover the entirety of the SUMMA technological base.

Blogs hosted on other platforms will also be used to raise visibility and reach a wider audience. This includes for instance the Deutsche Welle's Innovation blog (http://blogs.dw.com/innovation/).

2.4 Initial Dissemination Report

2.4.1 Status of Tools, Materials, Activities

The previous sections described the dissemination activities as they are planned. The current section presents tables indicating the current status and/or results of the planned different tools, materials and activities, using a colour coding showing the status, as well as descriptions of major achievements.

Table 1: Project Identity Status Sheet

Name	Project Identity
Task	T8.1
Due date	M2
Status	Completed
Partner(s) involved	DW, BBC, UEDIN, All
Leading partner	DW/BBC
Description	Logo, uniformity of colours, use of images, language, contact details.
Results/comments	Logo design (see http://summa-project.eu/) Use of a common SUMMA style Templates

Table 2: Project Fact Sheet Status Sheet

Name	Fact Sheet
Task	T8.1
Due date	M1
Status	Completed
Partner(s) involved	DW, UEDIN, BBC
Leading partner	DW
Description	Completing and submitting project fact sheet to EU Commission
Results/comments	Project Fact Sheet was submitted at project start.

Table 3: Project Website Creation Status Sheet

Name	Website creation and launch
Task	T8.1
Due date	M4
Status	Completed
Partner(s) involved	DW, BBC, UEDIN, All
Leading partner	DW
Description	Designing, creating and launching project website
Results/comments	SUMMA website was jointly designed, launched and enhanced in the first half year through a collaboration of Deutsche Welle, BBC and University of Edinburgh. A WordPress template is used. The website can be seen at www.summa-project.eu. It has a slider on the homepage to bring movement to the page. Its interactive pages are aimed at engaging the reader. The top navigation offers five main sections: project info, project partners, project output, related projects, and blog. The aim is to give the user a good, transparent overview and minimize the number of clicks to arrive at a page. The latest blog posts also appear on the home page. Each partner is represented with a logo and a short description of the organisation. The option ”Meet the team” introduces the user to the principal team members. The reader is given an overview of the objectives, the current status of the project, publications (linked to the openAIRE repository), and downloadable dissemination material such as flyers and posters. The blog presents news on the technology and application level, announces new releases, describes the project's participation in events, etc. It also regularly features longer articles, focusing on a particular technology, module or application. The latest SUMMA Tweets also automatically appear on the SUMMA website home page. The link to Github gives a direct line to the software that is made available by the project.

Table 4: Project Website Creation Status Sheet

Name	Website updates
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	DW, All
Leading partner	DW
Description	Updating project website
Results/comments	The website is regularly updated with new content, in particular in the form of blog posts. Also other informative sections, such as prototype descriptions and publication lists are updated as required. See www.summa-project.eu. As dissemination leader, Deutsche Welle is in charge of maintaining and updating the SUMMA website, and works closely with all partners to get their input and contributions to make this an active dissemination channel for the project. In particular new publications, presentations, awards are announced in blog posts and added to the designated pages; the latest blog posts also appear on the home page. A schedule is set up to ensure all partners contribute at least two longer articles on their module or technology, thus enriching the website with substantial and more in-depth content. In the reporting period, it contains for instance articles on neural machine translation and on Timeline Extraction. It is also used as a means of communication with stakeholders, as was true for the first User Day, for instance. The more advanced the technologies and the SUMMA platform are, the more detailed information the ”Prototypes and Technologies” section will contain. The second half of the project will see new regular enhancements of these pages.

Table 5: Project Wiki Creation Status Sheet

Name	Project wiki creation
Task	T8.1
Due date	M2
Status	Completed
Partner(s) involved	UEDIN, All
Leading partner	UEDIN
Description	Creating project wiki
Results/comments	The project wiki is the primary communication channel and repository within the project. It is was set up and is controlled by UEDIN. All partners contribute to the content of the wiki. Also the Jira instance, used for bug and enhancement tracking, is part of the SUMMA wiki. The wiki provides the consortium partners with essential project information means for shared use. It describes the workpackages and tasks, indicates the deadline, provides details on specific technologies or modules, describes progress and achievement. It keeps track of meetings (announcing them and providing meeting summaries). Is serves as common repository for information on Github releases, mailing lists, data repository, etc.

Table 6: Project Wiki Updates Status Sheet

Name	Project wiki updates
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	UEDIN, All
Leading partner	UEDIN
Description	Updating project wiki
Results/comments	As the primary communication channel for the project, the wiki is regularly updated with active contributions from all partners.

Table 7: Project Slide Pack Status Sheet

Name	Project Slide Pack
Task	T8.1
Due date	M4
Status	Completed
Partner(s) involved	UEDIN, DW
Leading partner	DW
Description	Creating standardised project slide pack
Results/comments	A generic presentation on the project has been set up for introducing the project. It has also been added to the web-site. Templates for project presentations using the project identity (common colours and logo) exist.

Table 8: SUMMA Twitter Account Creation Status Sheet

Name	Social Network - Twitter creation
Task	T8.1
Due date	M6
Status	Completed
Partner(s) involved	DW
Leading partner	DW
Description	Creation of project Twitter account
Results/comments	A SUMMA Twitter account has been set up. @SummaEu

Table 9: Twitter Updates Status Sheet

Name	Social Network - Twitter updates
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	DW, All
Leading partner	DW
Description	Updating of project Twitter account
Results/comments	Several partners regularly tweet on the SUMMA Twitter account @SummaEu, relating to research findings, or at events, for instance. It is used to keep related communities informed and learn from them.

Figure 1: SUMMA Twitter

Table 10: LinkedIn Account Creation Status Sheet

Name	Social Network - LinkedIn creation
Task	T8.1
Due date	M18
Status	Completed
Partner(s) involved	DW, All
Leading partner	DW
Description	Creation of project LinkedIn account
Results/comments	A SUMMA LinkedIn account has been set up.
Results/comments	A SUMMA LinkedIn group was created which is growing at a fast pace.

Table 11: LinkedIn Updates Status Sheet

Name	Social Network - LinkedIn updates
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	DW, BBC, UEDIN, All
Leading partner	DW
Description	Updating of project LinkedIn account
Results/comments	A network has been built up using the SUMMA name on LinkedIn. All partners have included SUMMA in their network. Link to the LinkedIn account.

Table 12: Generic Poster Status Sheet

Name	Generic poster
Task	T8.1
Due date	M12
Status	Completed
Partner(s) involved	DW, BBC, UEDIN, All
Leading partner	DW
Description	Creation of a generic project poster
Results/comments	A generic SUMMA poster describing the objectives and the overall process has been created and has been used at several events.

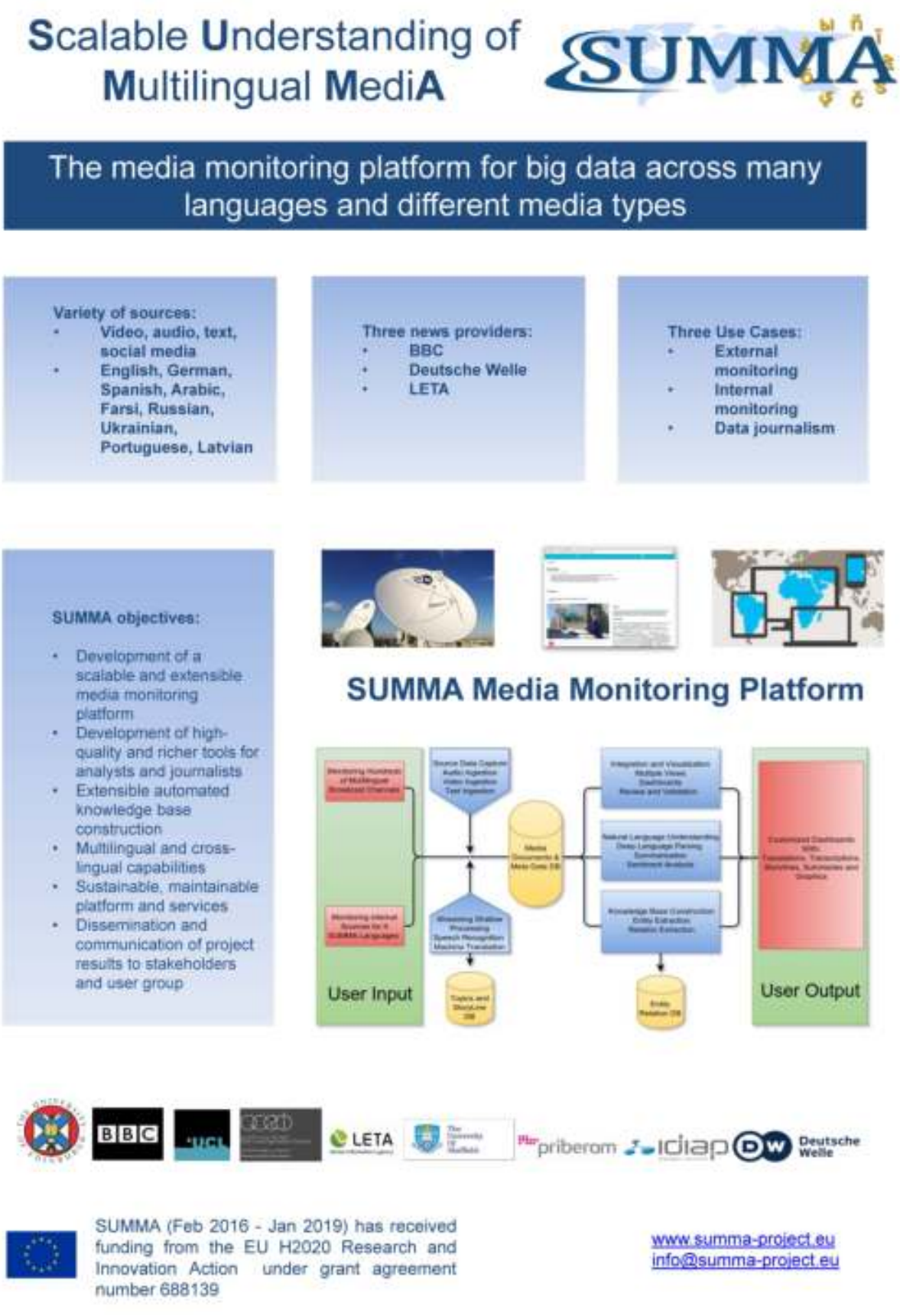


Figure 2: SUMMA Generic Poster

Table 13: Scientific Poster Status Sheet

Name	Scientific posters
Task	T8.1
Due date	M12
Status	Ongoing
Partner(s) involved	UEDIN, All
Leading partner	UEDIN
Description	Creation of several scientific project posters
Results/comments	Scientific posters are created on specific components or technologies or research results. These will be produced by the component owner. A uniform set of 11 SUMMA technical posters was created for the SUMMA user day in July 2017. At least 13 technical or scientific posters were produced for SUMMA in the first 18 months for presentation at several events. For more details, see the list of publications in section 2.4.4. http://summa-project.eu/publications/

Machine Translation

Alexandra Birch, Ulrich Germann, Tomasz Dwojak,
Andrei Popescu-Belis and Hassan M. Sajjad

a.birch@inf.ed.ac.uk, ugermann@inf.ed.ac.uk, t.dwojak@amu.edu.pl,
andrei.popescu-belis@idiap.ch, hsajjad@qf.org.qa

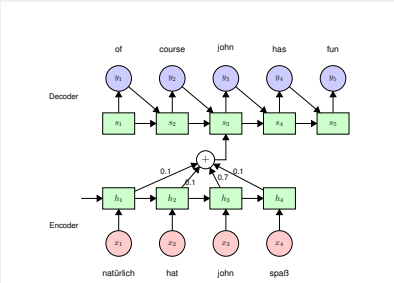
Overview

Translation of media content from across the world into English is a key enabling technology. In the SUMMA platform it allows the monitor to get a broad view of the news and also allows us to apply sophisticated natural language processing tools which have been developed largely for English. Translation for media monitoring faces the following challenges:

- Large volume of incoming text
- High resource (Arabic, German, Spanish, Portuguese, Russian, Latvian) and low resource (Ukrainian, Farsi) language pairs
- Translating output from speech recognition: potential errors, no segmentation, punctuation or capitalization
- Large variety of text styles and registers: speech, newswire, social media
- Constantly changing media landscape

MT meets Neural Networks

Machine translation has recently undergone a paradigm shift from phrase-based statistical models which combined many hand-engineered features, each applied independently, to one large neural network model where features are implicit and global dependencies are captured.



Encoder-decoder model with attention Bahdanau et al. (2015)

NMT toolkits developed for SUMMA

Nematus is implemented in Theano/Python. The toolkit prioritizes high translation accuracy, usability, and extensibility. Nematus has been used to build top-performing submissions to shared translation tasks at WMT 2016/2017 and IWSLT 2016. It is widely used in academic publications.

Marian is an open source neural network toolkit developed specifically for NMT. Marian provides fast, scalable, and efficient production ready software with no external dependencies. It is written in C++/CUDA and offers distributed GPU and CPU capabilities. Marian's CPU translation speed is nearly as good as Nematus' decoding speed on GPU. If GPUs are available, there is a speed up of a factor of ten.

Links

`https://github.com/rsennrich/nematus`
`https://marian-nmt.github.io/`

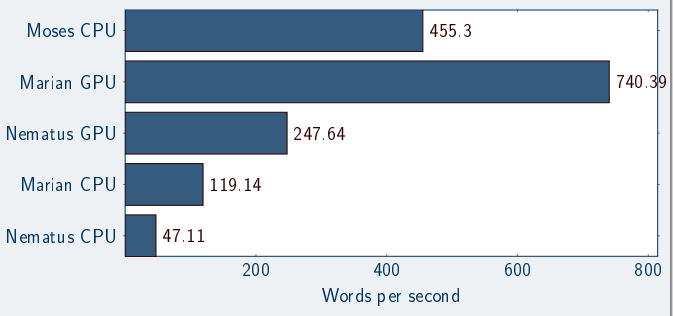
Translation Quality

The main goal of the SUMMA project is to deliver high quality machine translation. We deploy the-state-of-the-art MT models which won numerous tracks at the WMT 2017 shared task "News Translation". Our innovations include:

- Dealing with morphologically rich languages: translating sub-word units (BPE)
- Leveraging in-domain English text: backtranslation
- Deeper models

Translation direction	Shared Task BLEU	Google BLEU
German-English	35.10	32.48
Arabic-English	31.78	30.72
Russian-English	39.14	31.18
Spanish-English	26.83	34.20
Latvian-English	19.00	15.72

Performance



Batch-Decoding Using GPU, Marian can reach performance up to 5,000 wps.

Future Research

Spoken Language Translation

Spoken language translation is a challenge for MT. MT models are trained on written text and they struggle to translate ASR output. We are investigating using further information from the ASR model to improve translation of spoken language. Where ASR models are poor, using potentially more reliable phoneme sequences as additional source information to the MT model, can lead to better translations.

Dialects in Arabic

We want to be able to translate morphologically rich, resource poor Arabic dialects into English. We are investigating language independent tools for segmenting and processing dialects to optimize translation quality.

Low Resource Languages

We will provide translation models for Ukrainian and Farsi. Those languages are "low resourced" as there is very little existing translated corpora for training them. We will develop methods to deal with low resourced languages by leveraging corpora from related languages and applying machine learning techniques such as self-training.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 688139 (SUMMA).

Figure 3: Sample of SUMMA Scientific Poster on Machine Translation

Table 14: Project Flyer Status Sheet

Name	Flyer
Task	T8.1
Due date	M6
Status	Completed
Partner(s) involved	DW, BBC, UEDIN
Leading partner	DW
Description	Creation of a project flyer
Results/comments	In the early stages, a generic flyer was produced, using the SUMMA identity, i.e. colours, logo, et. It is a double-sided A4 flyer printed on quality paper. It has been distributed at several events, including MESA, GMF, LT-Innovate, EBU, and others. The flyer will be used throughout the project for general information on the project. If necessary a new version will be produced at a later stage. The flyer is downloadable from the SUMMA website: http://summa-project.eu/publications/

Figure 4: SUMMA Flyer

Table 15: User Brochure Status Sheet

Name	User Brochure
Task	T8.1
Due date	M30
Status	Not due yet
Partner(s) involved	DW, BBC, UEDIN
Leading partner	DW
Description	Creation of a user brochure
Results/comments	A brochure with user instructions on how to use the platform will be produced in the second half of the year. This will be used for live demos of the system.

Table 16: Project Banner Status Sheet

Name	Project Banner
Task	T8.1
Due date	M24
Status	Not due yet
Partner(s) involved	DW, BBC, UEDIN
Leading partner	DW
Description	Creation of a project banner
Results/comments	In the second half of the project, a banner will be created for dissemination at events.

Table 17: Screencast Status Sheet

Name	Screencast
Task	T8.1
Due date	M20
Status	Not due yet
Partner(s) involved	DW, BBC, LETA
Leading partner	DW
Description	Creation of a screencast of the available prototype
Results/comments	Once the platform has gone through a first set of tests and a stable version of the platform is available, a screencast is made for demonstration and training purposes. New screencasts will be produced after new releases featuring major differences.

Table 18: Promo Video Status Sheet

Name	Promo Video
Task	T8.1
Due date	M24
Status	Not due yet
Partner(s) involved	BBC, DW
Leading partner	BBC/DW
Description	Creation of a promo video
Results/comments	In the final year, a public promo movie will be produced by the broadcasters to be added to the website and for demo use at events. It may be an animated movie and will show the monitoring process and/or platform. It is meant for dissemination in industry and users within the broadcaster organisation.

Table 19: Academic Events Status Sheet

Name	Academic Events
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	All
Leading partner	UEDIN
Description	Attending academic events
Results/comments	In the first half of the project partners represented SUMMA at 20 academic events. See section 2.4.2 for full details.

Table 20: Industry Events Status Sheet

Name	Industry Events
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	BBC, DW, All
Leading partner	BBC/DW
Description	Attending industry events
Results/comments	In the first half of the project partners represented SUMMA at over 20 industry events, such as MESA (the Entertainment Industry), Deutsche Welle's Global Media Forum, and the European Broadcasting Union. Presentations were also held internally, for instance at DW it was introduced at various departmental and managerial meetings. See section 2.3.9 for a description of types of events covered and section 2.4.2 for a more detailed list of events attended in the period M1-18.

Table 21: Workshops Status Sheet

Name	Workshops
Task	T8.1
Due date	M24 and M36
Status	Not due yet
Partner(s) involved	UEDIN, All
Leading partner	UEDIN
Description	Organising Workshops
Results/comments	SUMMA will jointly organise scientific workshops in year 2 and 3.

Table 22: Hackathons Status Sheet

Name	Hackathons
Task	T8.1
Due date	M21 and M33
Status	Not due yet
Partner(s) involved	BBC, DW, All
Leading partner	BBC
Description	Organising Hackathons
Results/comments	BBC is in charge of organising several hackathons within the framework of SUMMA, covering different aspects, e.g. language technologies or data journalism. SUMMA team members participated in a BBC News Labs-organised hackathon on text-to-speech technologies, the Transcriptor NewsHack in January/February 2017.

Table 23: User Days Status Sheet

Name	User Days
Task	T8.1
Due date	M18 and further
Status	Ongoing
Partner(s) involved	UEDIN, BBC, DW, All
Leading partner	UEDIN
Description	Organising User Days
Results/comments	SUMMA organised its first User Day at BBC Monitoring in July 2017. Some 30 external participants attended. The user day was a success in many ways, featuring presentations, demos, poster sessions and panel discussions, and raising high interest from the visitors. A brief report on the user day can be found here: http://summa-project.eu/2017/07/13/post-user-day. The next user day will be held at Priberam's premises, in year 2. A final one will be held at Deutsche Welle in Bonn in year 3.

First SUMMA User Day Shows Potential

Figure 5: SUMMA User Day

Table 24: Collaboration Events Status Sheet

Name	Collaboration events
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	UEDIN, BBC, DW, All
Leading partner	UEDIN
Description	Participating in collaboration events
Results/comments	SUMMA has participated in collaboration events and efforts throughout the project existence, and will continue to do so. Working with and getting feedback from other projects is essential. Deutsche Welle has, for instance, collaborated with the EUMSSI (https://www.eumssi.eu/) project, the Newsstream (https://newsstreamproject.org/) project, and contributed to ARD (the German public broadcaster union, http://www.ard.de) meetings on ASR implementations and EBU (the European Broadcasting Union, https://www.ebu.ch/home) meetings on new technologies implementation in the media community. DW, LETA and Priberam collaborate in the Google-funded speech.media and news.bridge project on a workflow of ASR, MT and voiceover in a large number of languages. The two projects learn from each other on how to implement and improve a combination of ASR and MT output. BBC's ALTO voiceover project (http://bbcnewslabs.co.uk/projects/alto/) also provided expert interchange with SUMMA. UEDIN, QCRI, and BBC collaborate in the organisation of the MGB Challenge (http://mgb-challenge.org).

Table 25: Publications Status Sheet

Name	Publications
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	All
Leading partner	UEDIN
Description	Publication of SUMMA results in conference papers, journals and magazines
Results/comments	SUMMA published about 40 academic papers in the first 18 months of the project. Target publications include conference proceedings, journals and magazines. Publications details are entered and tracked through the openAIRE repository. See section 2.4.4 for a list of the publications.

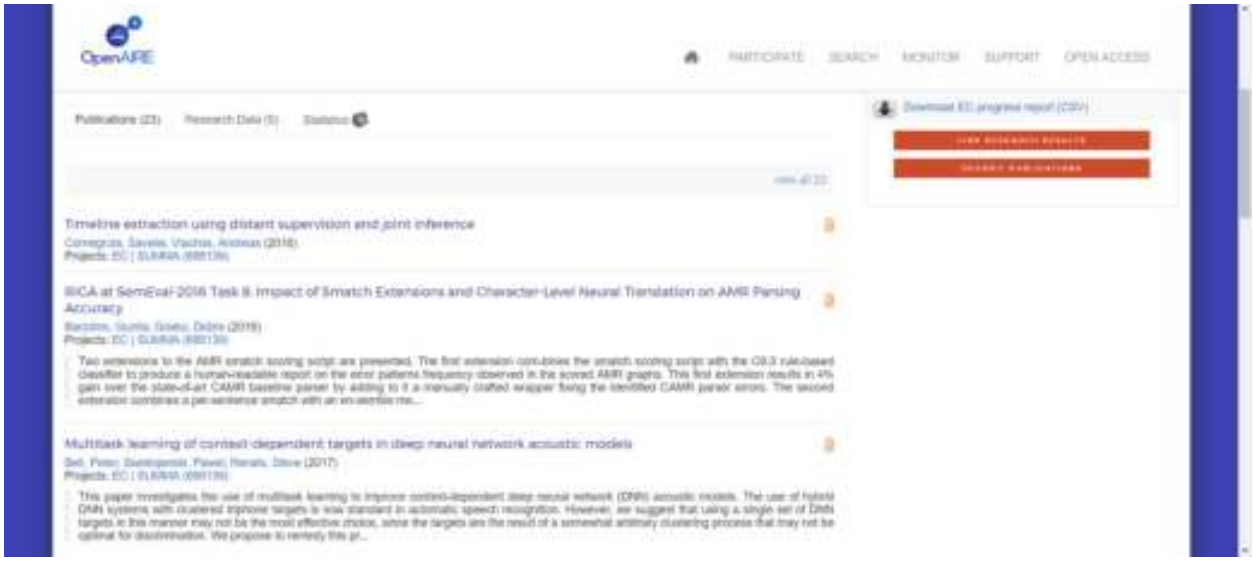


Figure 6: openAIRE Listing SUMMA Publications

Table 26: Press Releases Status Sheet

Name	Press Releases
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	All
Leading partner	UEDIN
Description	Issue of press releases on major SUMMA achievements.
Results/comments	LETA issued a press release during the reporting period. LETA Press Release on EACL Demo

Table 27: Blog Articles Status Sheet

Name	Blog articles
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	All
Leading partner	DW
Description	Partners will disseminate SUMMA results in several blogs, ensuring a wider dissemination of information on the project.
Results/comments	The SUMMA website has its own blog with regular posts on SUMMA activities and achievements, related technology developments, etc. See http://summa-project.eu/blog/. The blog posts included two longer paper articles, one on timeline extraction, the other one on Neural MT developments. SUMMA also featured in other blog posts. DW issued an article announcing the start of the SUMMA project in the Deutsche Welle Innovation Blog in December 2016, and mentioned SUMMA in several other blogposts in the Innovation Blog. More recently, another article was published in the Innovation Blog in July 2017 on the new release of the SUMMA platform. Similarly, the BBC also published two blog posts on SUMMA: introducing SUMMA and on BBC collaboration to build multilingual media monitoring system. See the Publications list below for more details on featuring blog entries.

Table 28: Awards and Prizes Status Sheet

Name	Awards, Prizes, Honourabe Mentions
Task	T8.1
Due date	Continuous
Status	Ongoing
Partner(s) involved	All
Leading partner	DW
Description	Listed here are awards, prizes and honourable mentions bestowed upon the SUMMA project.
Results/comments	LETA Wins AMR Parsing Trophy at SemEval 2016.// Team RIGA (LETA and University of Latvia) achieved the top result in the competition on Meaning Representation Parsing at SemEval-2016 competition in June 2016. See http://summa-project.eu/2016/08/04/leta-wins-amr-parsing-trophy-at-semeval-2016/ for details.

Figure 7: Award AMR Parsing 2017 LETA

2.4.2 List of Dissemination Events M1-M18

Academic Events The consortium participated in some 20 scientifically-oriented events, mostly conferences.

Papers were presented at most events, some the following events. More details are provided in section 2.4.4.:

1. ACL/EACL 2016 and 2017 - Association for Computational Linguistics/European Chapter of the Association for Computational Linguistics
2. ACL/WMT 2016 and 2017 - Association for Computational Linguistics/First and Second Conferences on Machine Translation
3. EMNLP 2016 and 2017 - Conference on Empirical Methods in Natural Language Processing
4. CORBON 2017 - 2nd edition of the workshop on Coreference Resolution Beyond OntoNotes
5. TAC 2016 - Text Analysis Conference
6. Interspeech 2016 and 2017 - Speech Technology Conference
7. IEEE ICASSP 2016 and 2017 - IEEE Conference on Acoustics Speech and Signal Processing
8. SLT 2016 - IEEE Workshop on Spoken Language Technology
9. INLG 2016 - Biannual International Conference on Natural Language Generation
10. IWSLT 2016 - International Workshop on Spoken Language Translation
11. SemEval 2016 - International Workshop on Semantic Evaluation
12. MediaEval 2016 - MediaEval Benchmarking Initiative for Multimedia Evaluation
13. LREC-2016, 10th edition of the Language Resources and Evaluation Conference

    In addition to the conferences and workshops above, partners were active at other academic events, for instance:

14. LXMLS Demo Day 2017, Lisbon, 25 July 2017 - The LXMLS is a summer school about machine learning taking place in Lisbon, Portugal (http://lxmls.it.pt/2017/). Priberam showcased the Entity Tagging and Linking and the Clustering systems currently being developed in the context of SUMMA.
15. IST Phd Open Days 2017, Lisbon, 5 April 2017 The PhD open Days at Técnico (IST, Lisbon) aim to bring together PhD students and experts.
16. EAMT 2016, European Association for Machine Translation, Riga, 30 May - 1 June 2016
17. UK-Speech 2016

Industrial Events

SUMMA was represented at the following industry-focused events during the first 18 months of the project:

1. SUMMA User Day I, Reading, 3 July 2017 - Participants: all consortium partners - Activity: demos, posters, presentations, panel discussion.
2. EBU (European Broadcast Union) technical meeting on media technologies, Bonn, 20 June 2017 - Participants: Deutsche Welle (Peggy van der Kreeft) - Activity: presenting machine translation workflow for DW content.
3. Global Media Forum (GMF) 2017, Bonn, 19-21 June 2017 - Participants: Deutsche Welle (Peggy van der Kreeft, Hina Imran) - Activity: flyers, networking, language technologies for media.
4. Demonstration at Cadena SER, May 2017, Madrid, Spain - Participants: Priberam (Carlos Amaral, Afonso Mendes) - Activity: Demonstration of Priberam's clustering and summarization prototype to gather feedback on its current stage as well as on the applicability to external sources monitoring needs of media organizations.
5. MESA (European branch of the Media & Entertainment Services Alliance) and HITS (Hollywood IT Industry) meeting, London, 2 March 2017 - Participants: Deutsche Welle (Peggy van der Kreeft), UEDIN (Peter Bell, Alexandra Birch), BBC (Chris Hernon) - Activity: Presentations on automated language technology workflow for media companies, machine translation and transcription issues. Discussion on usefulness for such technologies in the media industry.
6. Priberam Machine Learning Seminars 2017, March 2017, Lisbon - Participants: Priberam (Sebastião Miranda) - Activity: Demonstration and talk on "Online news clustering for crosslingual media monitoring" in the context of SUMMA.
7. Demonstration at El País and RTVE February 2017, Madrid, Spain - Participants: Priberam (Carlos Amaral, Afonso Mendes, Sebastião Miranda)- Activity: Demonstration of Priberam's clustering and summarization prototypes to gather feedback on its current stage as well as on the applicability to external sources monitoring needs of media organizations.
8. BBC News Lab Hackaton: newsHACK Transcriptor, London, 31 January - 1 February 2017 Participants: Deutsche Welle (Peggy van der Kreeft, Andreas Giefer), LETA (Renars Liepins), UEDIN (Peter Bell) - Activity: working on interfaces for ASR, including an editorial interface in between ASR and MT.
9. Language Equality in the Digital Age - Towards a Human Language Project, Brussels, 10 January 2017 - Participants: Deutsche Welle (Peggy van der Kreeft) - Activity: flyers, networking, participating in European Parliament committee discussion on the need for a specific language technology programme.
10. VRT Media Fast Forward (MFF) 2016, Brussels, 8 December 2016 - Participants: Deutsche Welle (Peggy van der Kreeft) - Activity: flyers, networking on media applications and discussing VRT automated subtitling tool.

11. LTAccelerate 2016, Brussels, 21-22 November 2016 - Participants: BBC (Susanne Weber), Deutsche Welle (Peggy van der Kreeft), UEDIN (Alexandra Birch) - Activity: panel discussion on the objectives and potential of the SUMMA platform for multilingual media monitoring.

12. Languages and the Media 2016, Berlin, 2-4 November 2016 - Participants: BBC (Susanne Weber) - Activity: Presentation

13. MT Marathon, Prague, 12-17 September 2016 - Participants: BBC (Susanne Weber) - Activity: Introducing SUMMA, Presentation on ALTO BBC MT Tool and language technologies for broadcast media.

14. IBC 2016, Amsterdam, 9-13 September 2016 - Participants: Deutsche Welle (Peggy van der Kreeft) - Activity: Networking, introducing SUMMA, flyer distribution

15. MetadataMadness 2016, MESA (European Chapter of Media and Entertainment Services Alliance), London, 6 September 2016 - Participants: Deutsche Welle (Peggy van der Kreeft) - Activity: Networking and introducing SUMMA.

16. WMT Shared Task on Bilingual Document Alignment - WMT is one of the key Workshops/-Conferences in Statistical Machine Translation, Berlin, 11-12 August 2016 - Participants: UEDIN (Ulrich Germann - Activity: Presentation on bilingual document alignment).

17. META-FORUM 2016, Lisbon, 4-5 July 2016 - Participants: UEDIN (Steve Renals) - Activity: presentation on SUMMA and a Roadmap of Spoken Language Technologies.

18. Global Media Forum (GMF) 2016, Bonn, 13-15 June 2016 - Participants: Deutsche Welle (Peggy van der Kreeft, Hina Imran) - Activity: presentation of SUMMA platform; poster, flyers; networking on media developments.

19. EBU MDN 2016 Metadata workshop, Geneva, 7-8 June 2016 - Participants: BBC (Susanne Weber) - Activity: presentation.

20. LTInnovate 2016, Brussels,17-18 May 2016 - Participants: Deutsche Welle (Peggy van der Kreeft), BBC (Susanne Weber) - Activity: introducing SUMMA.

21. Big Data Info Days 2016, Brussels, 14-15 January 2016 - Participants: UEDIN (Steve Renals), Deutsche Welle (Peggy van der Kreeft) - Activity: presentation of SUMMA project.

2.4.3 List of Planned Dissemination Events M19-M36

Listed below are some of the dissemination events in which SUMMA plans to actively participate during the second half of the project:

1. Interspeech 2017, Stockholm, 20-24 August 2017

2. MTM2017 Machine Translation Marathon, Lisbon 28 August - 2 September 2017

3. WMT2017, Copenhagen, 7-8 September 2017

4. MediaEval 2017, Dublin, 13-15 September 2017

5. IBC 2017, Amsterdam, 14-19 September 2017
6. MT Summit XVI, Technology Showcase, Nagoya, Japan, 18-22 September 2017
7. SICSA DemoFest 2017, Edinburgh, 3 October 2017
8. LTInnovate 2017, Brussels, 9-11 October 2017
9. BBC News Labs Hackaton on language technologies, London, October 2017
10. Text Analysis Conference (TAC-KBP), NIST USA, 13-14 November 2017
11. META-Forum 2017, Brussels, 13-14 November 2017
12. SUMMA User Day II, Lisbon, 2017
13. IEEE ASRU 2017 IEEE Automatic Speech Recognition and Understanding, Okinawa Japan, December 2017 (MGB Challenge 3)
14. EAMT 2018, 21th Annual Conference of the European Association for Machine Translation, Alicante, 28-30 May 2018
15. GMF 2018, Bonn, June 2018
16. ACL 2018, Melbourne, 15-18 July 2018
17. Interspeech 2018, 2018
18. Metadatamadness, London, September 2018
19. IBC 2018, Amsterdam, September 2018
20. International Conference on Language Transfer in Audiovisual Media, Language and the Media 2018, Berlin, November 2018
21. LTInnovate 2018, Brussels, 2018
22. LTAccelerate 2018, Brussels, 2018
23. META-Forum 2018
24. LREC-2018
25. BBC News Labs Hackaton on data journalism, London, 2018
26. SUMMA User Day III, Bonn, 2018

2.4.4 List of Publications M1-M18

Academic Publications

SUMMA academic publications are listed on OpenAIRE.

See the openAIRE SUMMA page for further details on these publications.

1. Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *Proceedings of the Workshop on Spoken Language Technology*, SLT '2016, San Diego, CA, USA, 2016

2. Antonio Valerio Miceli Barone, Jindrich Helcle, Rico Sennrich, Barry Haddow, and Alexandra Birch. Deep Architectures for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, 2017

3. Guntis Barzdins, Steve Renals, and Didzis Gosko. Character-Level Neural Translation for Multilingual Media Monitoring in the SUMMA Project. April 2016. arXiv preprint arXiv:1604.01221, 2016

4. Guntis Barzdins and Didzis Gosko. RIGA at SemEval-2016 Task 8: Impact of Smatch Extensions and Character-Level Neural Translation on AMR Parsing Accuracy. In *Proceedings of SemEval-2016, International Workshop on Semantic Evaluation*, San Diego, CA, USA, June 2016. arXiv preprint arXiv:1604.01278, 2016

5. Peter Bell, Pawel Swietojanski, and Steve Renals. Multitask learning of context-dependent targets in deep neural network acoustic models. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(2):238–247, 2017

6. Savelie Cornegruta and Andreas Vlachos. Timeline extraction using distant supervision and joint inference. In *Proceedings of EMNLP-2016*, pages 1936-1942, Austin, TX, USA, November 2016

7. Marco Damonte, Shay B. Cohen and Giorgio Satta. An Incremental Parser for Abstract Meaning Representation, in *Proc EACL*, 2017.

8. Joachim Fainberg, Steve Renals, and Peter Bell. Factorised representations for neural network adaptation to diverse acoustic environments. In *Proc Interspeech*, 2017

9. Siva Reddy Gangireddy, Pawel Swietojanski, Peter Bell, and Steve Renals. Unsupervised adaptation of recurrent neural network language models. In *Proc Interspeech*, 2016

10. Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for micro-planners. In *Proceedings of ACL*, 2017

11. Ulrich Germann. Bilingual document alignment with latent semantic indexing. In *Proceedings of the First Conference on Machine Translation*, pages 692–696, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2368

12. Normunds Gruzitis and Guntis Barzdins. The Role of CNL and AMR in Scalable Abstractive Summarization for Multilingual Media Monitoring. arXiv preprint arXiv: 1606.05994, 2016

13. Marcin Junczys-Dowmunt, Tomasz Dwojak and Rico Sennrich. . The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, Volume 2: Shared Task Papers, Berlin, Germany, pages 316-322. 2016. Association for Computational Linguistics

14. Marcin Junczys-Dowmunt and Alexandra Birch. Edinburgh Neural Machine Translation Systems for IWSLT 16. In *Proceedings of International Workshop of Spoken Language Translation*, Seattle, USA, December 2016

15. Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is neural machine translation ready for deployment? A case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*, 2016

16. Ondřej Klejch, Peter Bell, and Steve Renals. Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *IEEE Workshop on Spoken Language Technology*, December 2016

17. Ondřej Klejch, Peter Bell, and Steve Renals. Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *IEEE ICASSP*, April 2017

18. Alexandros Lazaridis, Ivan Himawan, Petr Motlicek, Iosif Mporas and Philip N. Garner. Investigating Cross-lingual Multi-level Adaptive Networks: The Importance of the Correlation of Source and Target Languages. In *Proceedings of the International Workshop on Spoken Language Translation*, Seattle, WA, USA, 2016

19. Renars Liepins, Ulrich Germann, Guntis Barzdins, Alexandra Birch, Steve Renals, Susanne Weber, Peggy van der Kreeft, Herve Bourlard, João Prieto, Ondrej Klejch, Peter Bell, Alexandros Lazaridis, Alfonso Mendes, Sebastian Riedel, Mariana S. C. Almeida, Pedro Balage, Shay B. Cohen, Tomasz Dwojak, Philip N. Garner, Andreas Giefer, Marcin Junczys-Dowmunt, Hina Imran, David Nogueira, Ahmed Ali, Sebastião Miranda, Andrei Popescu-Belis, Lesly Miculicich Werlen, Nikos Papasarantopoulos, Abiola Obamuyide, Clive Jones, Fahim Dalvi, Andreas Vlachos, Yang Wang, Sibo Tong, Rico Sennrich, Nikolaos Pappas, Shashi Narayan, Marco Damonte, Nadir Durrani, Sameer Khurana, Ahmed Abdelali, Hassan Sajjad, Stephan Vogel, David Sheppey, Chris Hernon, and Jeff Mitchell. The summa platform prototype. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–119, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://aclweb.org/anthology/E17-3029

20. Liang Lu and Steve Renals. Small-footprint highway deep neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(7):1502–1511, 2017. doi: 10.1109/TASLP.2017.2698723

21. Zita Marinho, André FT Martins, Shay B Cohen, and Noah A Smith. Semi-supervised learning of sequence models with method of moments. In *EMNLP*, pages 287–296, 2016

22. Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Syntax-aware Neural Machine Translation Using CCG. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, 2017

23. Shashi Narayan, Siva Reddy, and Shay Cohen. Paraphrase generation from Latent-Variable PCFGs for semantic parsing. In *Proceedings of INLG*, 2016

24. Shashi Narayan and Claire Gardent. Unsupervised sentence simplification using deep semantics. In *Proceedings of INLG*, 2016

25. Shashi Narayan and Shay B. Cohen. Optimizing spectral learning for parsing. In *Proceedings of ACL*, 2016

26. Dominique Osborne, Shashi Narayan, and Shay Cohen. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430, 2016. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/895

27. Peteris Paikens, Guntis Barzdins, Afonso Mendes, Daniel Ferreira, Samuel Broscheit, Mariana S. C. Almeida, Sebastião Miranda, David Nogueira, Pedro Balage, and André F. T. Martins. Summa at tac knowledge base population task 2016. In *Proceedings of the Text Analysis Conference -TAC*, pages 1–9, Gaithersburg, Maryland USA, 2017

28. Nikolaos Pappas and Andrei Popescu-Belis. Human versus machine attention in document classification: A dataset with crowdsourced annotations. In *Proceedings of the EMNLP 2016 Workshop on Natural Language Processing for Social Media*, Austin, TX, USA, 2016

29. Nikolaos Pappas and Andrei Popescu-Belis. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626, 2017

30. Xiao Pu, Nikolaos Pappas, and Andrei Popescu-Belis. Sense-aware statistical machine translation using adaptive context-dependent clustering. In *Proceedings of Second Conference on Machine Translation (WMT 2017)*, Copenhagen, Denmark, 2017

31. Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April 2017b. Association for Computational Linguistics

32. Rico Sennrich. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, Short Papers, pp. 376-382, Valencia, Spain, 2017. Association for Computational Linguistics

33. Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark, 2017a

34. Pawel Swietojanski, Jinyu Li, and Steve Renals. Learning hidden unit contributions for unsupervised acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(8):1450–1463, 2016. doi: 10.1109/TASLP.2016.2560534

35. Pawel Swietojanski and Steve Renals. Differentiable pooling for unsupervised acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(10):1773–1784, 2016. doi: 10.1109/TASLP.2016.2584700

36. J. Thorne and Andreas Vlachos. An Extensible Framework for Verification of Numerical Claims. In *Proceedings of the Software Demonstrations of EACL 2017, the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain, April 2017

37. Sibo Tong, Philip N. Garner, and Hervé Bourlard. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Proc. of Interspeech*, 2017

38. Emiru Tsunoo, Peter Bell, and Steve Renals. Hierarchical recurrent neural network for story segmentation. In *Interspeech*, August 2017

39. Lesly Miculicich Werlen and Andrei Popescu-Belis. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the 3rd EMNLP Workshop on Discourse in Machine Translation (DiscoMT 2017)*, pages 17–25, Copenhagen, Denmark, 2017b

40. Lesly Miculicich Werlen and Andrei Popescu-Belis. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain, 2017a

Other Publications

1. SUMMA blog article: University of Edinburgh developing attention-based NMT model called Nematus/
2. SUMMA blog article: Timeline Extraction in SUMMA
3. DW Innovation blog article: Scalable Understanding of Multilingual MediA
4. DW Innovation blog article: SUMMA Monitoring Platform Provides Cross-Lingual Overview of DW Content into English
5. BBC Internet blog: BBC collaboration to build multilingual media monitoring system, 1 October 2015
6. BBC NEWSLabs blog: SUMMA - Scalable Understanding of Multilingual Media, 2017
7. Press Release: LETA Press Release on EACL Demo
8. Poster: Speech Recognition - SUMMA User Day July 2017
9. Poster: Punctuation Prediction - SUMMA User Day July 2017
10. Poster: Machine Translation - SUMMA User Day July 2017

11. Poster: Scalable News Clustering and Topic Detection - SUMMA User Day July 2017
12. Poster: Entity Tagging & Linking - SUMMA User Day July 2017
13. Poster: Knowledge Base Population - SUMMA User Day July 2017
14. Poster: Identification and Verification of Simple Claims about Statistical Properties - SUMMA User Day July 2017
15. Poster: Multilingual Semantic Parsing Using Abstract Meaning Representation - SUMMA User Day July 2017
16. Poster: Story Highlight Generation - SUMMA User Day July 2017
17. Poster: Test Data Supplied by BBC - SUMMA User Day July 2017
18. Poster: SUMMA Platform - SUMMA User Day July 2017
19. Poster: The SUMMA Prototype Platform - EACL 2017
20. Poster: Bilingual Document Alignment with Latent Semantic Indexing - WMT 2016

3 Exploitation Plan and Initial Report

This section of the deliverable relates to Task 8.2 (Exploitation), focusing on capturing the information required and exploring options to ensure the outputs of the SUMMA project can be exploited by the partners themselves and others. As the exploitation roadmap, this document sets out the activities required for successful exploitation of SUMMA and reports on progress so far. Two foundation tasks are essential for supporting exploitation activities:

- **Exploitation Committee**
  The role of the Exploitation Committee is to set and execute the exploitation strategy of the consortium as well as to ensure that IPR Management is being carried out appropriately.
- **IPR Management**
  The Exploitation Committee has mechanisms in place to manage the record of IPR being contributed to the project by the consortium members.

Building on this foundation there are three main ways in which SUMMA can be exploited. This document reports on these in order of increasing complexity and in reverse order of preference (i.e. with the simplest outcome described first and with the ideal but most complex outcome described last):

- **Knowledge** (for new applications)
  Through its research, each partner is learning new techniques that can be applied to other projects, and developing ideas for other new services.
- **Component-Based Exploitation** (for other applications)
  The components that power SUMMA have a high potential value as improvements for existing services, or as the basis of other new services.
- **Multi-Strand Exploitation** (SUMMA as a whole)
  Notwithstanding the points above, the focus remains to provide an integrated product. During the project the partners will investigate the options further and will aim to give further consideration to the revenue potential of the best options.

3.1 Exploitation Committee

At the SUMMA project meeting held in Lisbon in November 2016 the project established an Exploitation Committee to coordinate the management of IPR and to set and execute the exploitation strategy of the consortium.

It was agreed there would be one representative from each project partner with initial membership confirmed as follows:

- Alexandra Birch, UEDIN (attending for information)
- Andreas Vlachos, USFD
- Andrei Popescu-Belis, IDIAP (on behalf of Hervé Bourlard)

- Dominic Tinley, BBC (attending for information)
- Guntis Barzdins, LETA
- João Prieto, PRIBERAM
- Peggy van der Kreef, DW
- Sebastian Riedel, UCL
- Steve Renals, UEDIN (representing UEDIN)
- Susanne Weber, BBC (representing BBC)

The Committee agreed to conduct its business via email as far as practicable and an email group was established for this purpose.

The Committee discussed the three tiers of exploitation set out on the previous page and agreed details of each should be captured in this deliverable.

The Committee agreed the first task for the project towards exploitation would be to gather detailed information on all components to ensure as much information as possible is captured on how each component could be reused, either individually or collectively (see 3.2).

The Committee agreed it would convene again at the first full project meeting following completion of this deliverable. Armed with the information contained herein all members will be invited to contribute to a workshop that explores potential use of the SUMMA project outputs by using combinations of components to address known and perceived user needs informed by partners' own knowledge and information gathered from user group meetings.

3.2 IPR Management

The Exploitation Committee agreed each work package should be broken down into an agreed set of components and that the following information should be captured for each component:

- Component name
- Inputs from
- Outputs to
- Component lead partner
- Component contributors
- Brief description
- What it does (more detail)
- How it works (more detail)
- Key innovative aspects

- Potential applications
- Software & IPR status
- Terms & conditions of use
- Performance requirements
- Further documentation
- Alternatives
- Key contact(s)

In particular the Committee saught to address

- Potential applications of any components aside from SUMMA
- The names and details of any libraries used within components and their IPR status
- Answers key questions developers wishing to exploit the component are likely to have such as: How many users can be supported? What spec machines required to run the code?
- The name of and links to alternative open source components if a component will not be part of an open source release

The full results of this exercise are included in section 3.4.

3.3 Knowledge

The first tier of exploitation is through Knowledge. At the very least each partner is learning new techniques through its research that can be applied to other projects and developing ideas for other new services.

This section includes a brief description of the knowledge each partner hopes to gain from taking part in SUMMA and how this relates to its broader activities.

While this is not listed as a requirement of this project deliverable it is nevertheless a first step towards fuller exploitation so the Exploitation Committee felt it was of value to include in this report.

3.3.1 University of Edinburgh (UEDIN)

As a result of SUMMA, UEDIN will further develop its expertise and know-how in automatic speech recognition, machine translation, and natural language processing through improved multilingual coverage, the development of new models and algorithms, and evaluation on data related to the SUMMA use cases. The SUMMA platform will enable us to integrate data-driven language technologies with streaming media provision in a scalable way and will enable the (co-)development of systems for the SUMMA use cases, and for broader applications.

3.3.2 Priberam (PRIBERAM)

In the scope of the SUMMA project, Priberam is enriching its resources, expanding technologies and the coverage of its product offer. Namely it has improved the NLP pipeline, added clustering functionalities to the search engine, implemented new summarization techniques, delivered the new NEL system and expanded the offer on languages covered. On the commercial side we have been actively presenting the SUMMA platform to our media customers in Portugal, Spain and Brazil with very good acceptance and interest. At the end of the Project, Priberam will have a new set of product functionalities and will possibly commercialize SUMMA components or even the complete platform.

3.3.3 University College London (UCL)

Through the SUMMA project, UCL is acquiring expertise in working with multilingual data—previously they mostly focused on English. This involves access to a world leading team of researchers in MT and multilingual speech recognition. Moreover, we are learning about software integration and downstream users of our knowledge-base population (KBP) and fact checking technologies. There is a substantial gap between the research community that develops KBP methods and actual users of this technology.

3.3.4 Idiap Research Institute (IDIAP)

Idiap has a long history of R&D in the component technologies of (multilingual) ASR and NLP/MT. Through SUMMA, Idiap hopes to expand its knowledge of how to apply these technologies in the emerging fields of big data and media monitoring. Idiap will also gain knowledge about how to integrate these technologies in a practical sense, via intermediate technologies such as punctuation prediction. This in turn will enhance technology transfer and future participation in collaborative programmes.

3.3.5 Latvian News Agency (LETA)

Through the SUMMA project LETA is acquiring state-of-art in-house expertise in ASR and MT for the Latvian language - areas where LETA previously relied on external suppliers. Prototype integration, clustering, summarisation, NEL are other SUMMA technologies directly applicable to LETA daily business as a national news agency. If by the end of the project SUMMA Platform will approach a viable product level, LETA could facilitate its commercialisation and support after the project.

3.3.6 British Broadcasting Corporation (BBC)

The BBC has an ambition to use the SUMMA platform as a whole (see 3.5.1). In the meantime work on the SUMMA project is providing valuable insights into automatic speech recognition (ASR), machine translation and topic extraction, all of which are used to some degree within existing BBC systems. While the focus remains on integrating the full SUMMA platform, the knowledge gained in the specific areas mentioned is also being applied to see where improvements can be made to existing systems. As an example, the BBC is liaising with UEDIN on applying

its ASR algorithms developed for SUMMA to other BBC projects. BBC Monitoring (BBCM) specifically is using SUMMA to build relationships in the language-technology community to both gain knowledge and solutions and provide expertise and assistance in developing this kind of technology.

3.3.7 Qatar Computing Research Institute (QCRI)

Qatar Computing Research Institute (QCRI) is a national research institute, established in 2010 by Qatar Foundation for Education, Science and Community Development, a private, non-profit organization that is supporting Qatar's transformation from carbon economy to knowledge economy. As such QCRI performs cutting-edge research in such areas as Arabic language technologies, social computing, data analytics, distributed systems, cyber security and computational science and engineering. SUMMA will both draw on and our expertise in automatic speech recognition and statistical machine translation of Arabic language and its dialects.

3.3.8 Deutsche Welle (DW)

Deutsche Welle improves its expertise in developing dashboards, as well as gaining knowledge in setting up Docker environments with the aim of running a local copy of the SUMMA platform at its premise. It also gains experience with using tablet computers as an ultra-mobile tool to access monitoring information in meetings.

3.3.9 University of Sheffield (USFD)

As a result of SUMMA, USFD will advance its methods in knowledge base population, taking into account the needs of BBC Media Monitoring and DW. Furthermore, USFD hopes that the interaction with these organisations will help it progress its research in fact-checking by informing it through the SUMMA use cases. Finally, the integration with the other technical partners will help gain experience with multiple languages.

3.4 Component-Based Exploitation

As explained in sections 3.1 and 3.2, the Exploitation Committee agreed the first task for the project towards exploitation would be to gather detailed information on all components to ensure as much information as possible is captured on how each component could be reused, either individually or collectively

The Committee agreed each work package should be broken down into an agreed set of components. This section includes a table for each component including all the information known at this stage. This will be amended and appended as the components evolve during the remainder of the project.

3.4.1 WP2 Data Collection and Management

Table 29: AV Transcoder component details

Component name	AV Transcoder
Inputs from	IP streaming video or video files.
Outputs to	HLS video stream
Component lead partner	BBC NI
Component contributors	BBC NI
Brief description	A service that takes a video stream or file and converts it to an HLS media stream.
What it does (more detail)	A service that takes a video stream or file and converts it to an HLS media stream for ingest by any deployments of the SUMMA platform.
How it works (more detail)	Employs FFMpeg on an NVidia based hardware accelerated platform to convert a video stream or file into an HLS media stream.
Key innovative aspects	None
Potential applications	Converting video streams to HLS.
Software & IPR status	Reliant on FFMpeg and NVidia H264 hardware Encoder driver, so subject to their licensing terms.
Terms & conditions of use	FFMpeg is open source, but NVidia software is proprietary. This component does not form part of the SUMMA platform and was created to enable fulfilment of BBC NI's WP2 requirements. It is subject to BBC NI copyright.
Performance requirements	NA
Further documentation	NA
Alternatives	None
Key contact(s)	David Sheppey

Table 30: S3 streaming service component details

Component name	S3 streaming service
Inputs from	HLS streams delivered from AV Transcoder component
Outputs to	S3 objects
Component lead partner	BBC NI
Component contributors	BBC NI, Amazon
Brief description	Redistributes HLS video steams to the Consortium partners
What it does (more detail)	Takes HLS streams from AV transcoder service and stores them within AWS S3 storage. It will then allow partners to connect to the service and subscribe to the media streams.
How it works (more detail)	The component is a service which hosts the AWS S3 file system driver which allows saving of HLS file segments within AWS object storage. It also incorporates a mechanism for retrieval of the segments in the form of an HLS stream via access control.
Key innovative aspects	None
Potential applications	HLS video distribution
Software & IPR status	AWS file system driver is open source.
Terms & conditions of use	This component does not form part of the SUMMA platform and was created to enable fulfilment of BBC NI's WP2 requirements. It is subject to BBC NI copyright. AWS file system driver is open source.
Performance requirements	NA
Further documentation	None
Alternatives	Other S3 file system drivers are available.
Key contact(s)	David Sheppey

Table 31: Text Source Scraper component details

Component name	Text Source Scraper
Inputs from	The Internet
Outputs to	SUMMA platform API or disk dump
Component lead partner	BBC NI
Component contributors	BBC NI
Brief description	Finds and scrapes text from webpages, primarily news websites and blogs
What it does (more detail)	Based on a list of RSS news feeds the tool will look for new articles and 'scrape' the content of the article
How it works (more detail)	Every minute the component checks the HEAD response of each url in sources database and compares its modified time with the database. If it has changed the full page (xml feed page) is requested. The component then reads the feed and checks to see if there is an article that hasn't been scraped. If so the article is added to a job queue. A pool of worker scrapers churn through the queue. When the scrape is complete it is published to Redis pubsub and the component (which is subscribed to this channel) takes the result and either saves it to disk in the Summa format or sends it to the platform API
Key innovative aspects	None
Potential applications	
Software & IPR status	Relies on open source components for the scraping http://www.html-tidy.org/ and http://newspaper.readthedocs.io/en/latest/
Performance requirements	Low because it can be broken into many workers (when finished!). Except the redis database, it will grow over time, so the amount of memory needed will also grow, but prob 32GB of RAM would keep years of articles from 1000s of feeds. Data dumping would require HP server with Xeon and 32GB RAM and 64GB disk space for backup.
Alternatives	Other HTML scrapers are available
Key contact(s)	David Sheppey

Table 32: Twitter Ingester component details

Component name	Twitter Ingester
Inputs from	Twitter
Outputs to	SUMMA platform API or disk dump
Component lead partner	BBC NI
Component contributors	BBC NI
Brief description	Provides tweets to the SUMMA platform
What it does (more detail)	Uses a database of source tags, accounts or lists and sends them to the twitter api so that they can be monitored
How it works (more detail)	The twitter API has a 'streaming' function that allows many tags and accounts to be monitored at the same time. The component uses this part of the API to maintain an open connection to the twitter servers and listens for matching events. When a matching tweet is sent the component either sends it to the summa platform api via HTTP or to disk using the summa data format (or both at the same time)
Key innovative aspects	None
Potential applications	Anything that needs realtime monitoring of twitter where the events represent less than 1% of all tweets at any moment
Software & IPR status	Uses Tweetinvi and the Twitter API which is subject to terms of service
Terms & conditions of use	https://dev.twitter.com/overview/terms/agreement-and-policy
Performance requirements	May require up to 50Mbps of network bandwidth, but other than that low. HP Xeon server with 4GB RAM. Disk space needed is dependant on size of data dump wanted.
Further documentation	
Alternatives	Alternative twitter api connecting libraries are available.
Key contact(s)	David Sheppey

3.4.2 WP3 Stream Processing

Table 33: Speech Recognition Docker component details

Component name	Speech Recognition Docker component
Inputs from	BBC and DW media streams
Outputs to	Punctuation prediction, Machine translation
Component lead partner	UEDIN
Component contributors	UEDIN
Brief description	A docker image which runs an API endpoint for automatic speech recognition.
What it does (more detail)	Transcribes incoming speech to a sequence words.
How it works (more detail)	The API uses pre-trained Kaldi models to transcribe incoming speech.
Key innovative aspects	none
Potential applications	Applications using speech recognition.
Software & IPR status	alex-asr (Apache 2.0) https://github.com/choko/alex-asr
Terms & conditions of use	open source
Performance requirements	One container can handle one stream. Can be run in parallel. Memory and CPU requirements depend on the used Kaldi model.
Further documentation	https://github.com/choko/alex-asr
Alternatives	CloudASR https://github.com/UFAL-DSG/cloud-asr Kaldi GStreamer server https://github.com/alumae/kaldi-gstreamer-server
Key contact(s)	Ondřej Klejch

Table 34: Speech Recognition English Models component details

Component name	Speech Recognition English Models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	UEDIN
Component contributors	UEDIN
Brief description	Deep neural network acoustic models and n-gram language models matched to English broadcast TV
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. Deep neural networks are trained on English TV audio from the BBC using a lightly supervised training procedure, and n-gram models are trained on large quantities of subtitle text.
Key innovative aspects	Lightly supervised training combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of English TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Open source
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be $< 20\%$
Further documentation	http://kaldi-asr.org
Alternatives	Use baseline free models, eg. trained on the TED-LIUM corpus http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus
Key contact(s)	Peter Bell

Table 35: Speech Recognition Arabic Models component details

Component name	Speech Recognition Arabic models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	UEDIN
Component contributors	UEDIN, QCRI
Brief description	Deep neural network acoustic models and n-gram language models matched to Aljazeera Arabic TV programs
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken Arabic words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. Deep neural networks are trained on Arabic TV programs recorded by Aljazeera Arabic TV, the data is manually transcribed, and n-gram language models are trained on Aljazeera web site archive as well as the Arabic giga word corpus.
Key innovative aspects	Lightly supervised training, dialectal modelling, combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of Arabic TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Open source
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be < 20%
Further documentation	http://kaldi-asr.org
Alternatives	none
Key contact(s)	Ahmed Ali

Table 36: Speech Recognition German models component details

Component name	Speech Recognition German models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	UEDIN
Component contributors	UEDIN, IDIAP
Brief description	Deep neural network acoustic models and n-gram language models matched to German broadcast TV and radio programs
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken German words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. Deep neural networks are trained on a combination of manually transcribed broadcasts (TUM BCN data) and data obtained from DW with ad-hoc captioning. Language models trained using common crawl and newswire data.
Key innovative aspects	Lightly supervised training combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of German TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Open source
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be $< 25\%$
Further documentation	http://kaldi-asr.org
Alternatives	none
Key contact(s)	Peter Bell

Table 37: Speech Recognition Spanish models component details

Component name	Speech Recognition Spanish models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	UEDIN
Component contributors	UEDIN, IDIAP
Brief description	Deep neural network acoustic models currently trained on Globalphone and uncaptioned TV broadcasts and n-gram language models trained on Gigaword
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken Spanish words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. Deep neural networks are bootstrapped on the Globalphone corpus and trained further unsupervised on Spanish TV programs; n-gram language models are trained on Spanish gigaword corpus.
Key innovative aspects	Lightly supervised and unsupervised training combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of Spanish TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Open source
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be $< 30\%$
Further documentation	http://kaldi-asr.org
Alternatives	none
Key contact(s)	Peter Bell

Table 38: Speech Recognition Russian Models component details

Component name	Speech Recognition Russian models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	IDIAP
Component contributors	UEDIN, IDIAP
Brief description	Deep neural network acoustic models currently trained on Globalphone, Euronews, and uncaptioned TV broadcasts and n-gram language models trained on large set of Russian news data
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken Russian words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. Deep neural networks are bootstrapped on the Globalphone and Euronews corpora and trained further unsupervised on Russian TV programs; n-gram language models are trained on Russian news data.
Key innovative aspects	Lightly supervised and unsupervised training combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of Russian TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Open source
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be $< 30\%$
Further documentation	http://kaldi-asr.org
Alternatives	none
Key contact(s)	Phil Garner

Table 39: Speech Recognition Latvian Models component details

Component name	Speech Recognition Latvian Models
Inputs from	Speech Recognition Docker component
Outputs to	Speech Recognition Docker component
Component lead partner	LETA
Component contributors	LETA, UEDIN, IDIAP
Brief description	Deep neural network acoustic models trained on large set of Latvian news data
What it does (more detail)	Given input audio provided via the Docker Component, the models are used to compute the most likely sequence of spoken Latvian words.
How it works (more detail)	All models are statistical and are estimated on large quantities of speech and text data. n-gram language models are trained on Latvian news data.
Key innovative aspects	Lightly supervised and unsupervised training combined with state-of-the-art acoustic and language modelling approaches.
Potential applications	Applications requiring speech recognition of Latvian TV broadcasts
Software & IPR status	Kaldi (Apache 2.0) http://kaldi-asr.org
Terms & conditions of use	Licensed
Performance requirements	Specified in terms of expected Word Error Rate (WER), expected to be $< 30\%$
Further documentation	http://kaldi-asr.org
Alternatives	none
Key contact(s)	Roberts Dargis, Arturs Znotins

Table 40: Punctuation prediction component details

Component name	Punctuation prediction
Inputs from	ASR
Outputs to	Machine translation and others
Component lead partner	IDIAP
Component contributors	IDIAP, UEDIN
Brief description	Adds punctuation to ASR output, which is normally just words.
What it does (more detail)	It adds full stops and commas to text from ASR that would not otherwise have such markup. This enahances readability and aids translation.
How it works (more detail)	It combines analysis of pause duration (from the acoustic signal) with a language model that includes knowledge of which words tend to precede and follow punctuation. There are also ad-hoc rules for each language.
Key innovative aspects	It has been shown to generalise across language.
Potential applications	Anything where punctuation needs to be added to ASR output.
Software & IPR status	Mainly written in python. Idiap copyright; available to SUMMA partners.
Terms & conditions of use	Likely to be open-sourced.
Performance requirements	The language models can be large — a few GB — so it's not an embedded thing.
Further documentation	None yet.
Alternatives	In fact there are two solutions: one from UEDIN is based on MT; one from IDIAP is language model plus pause duration.
Key contact(s)	Phil Garner

Table 41: Text normalisation component details

Component name	Text normalisation
Inputs from	(offline)
Outputs to	(offline)
Component lead partner	IDIAP
Component contributors	IDIAP
Brief description	Converts text to a standard form that can be utilised by ASR and MT training.
What it does (more detail)	Amongst other functions, it removes unnecessary punctuation, converts numbers to words and handles abbreviations.
How it works (more detail)	It has a list of rules for each language. It also interfaces with standard python components for number conversion.
Key innovative aspects	Rules for several languages.
Potential applications	Cleaning up text for display or use in HTML. Its use for ASR and MT extends outside the project too.
Software & IPR status	It's in python; open source.
Terms & conditions of use	Open source, 3-clause BSD
Performance requirements	It's an offline scripting sort of thing; no large memory or CPU requirements.
Further documentation	https://github.com/idiap/asrt
Alternatives	There are many text-normalisation solutions including festival (for TTS) and Moses (for NLP). asrt is more modern, but some parts of SUMMA rely on Moses for legacy reasons.
Key contact(s)	Phil Garner

Table 42: Neural Translation Models component details

Component name	Neural Translation Models
Inputs from	BBC and DW text and social media streams, Punctuation prediction model
Outputs to	SUMMA database and all downstream NLP tools
Component lead partner	UEDIN
Component contributors	Alexandra Birch, Tomasz Dwojak, Ulrich Germann, Rico Sennrich, Marcin Junczys-Dowmunt
Brief description	A code base for training and decoding MT models. A docker image which runs an API endpoint for machine translation into English.
What it does (more detail)	It takes SUMMA relevant source language text that has been segmented and punctuated to resemble clean written text, and it translates it into English. It outputs a sequence of words, their alignment to source words, and how high confidence the translated word is.
How it works (more detail)	The API uses a pretrained neural machine translation model, currently created by using Nematus. The efficient CPU/GPU decoder is developed in Marian.
Key innovative aspects	Nematus delivered winning systems in 2016+2017 WMT translation shared task. Marian is a production ready, multi-purpose neural network toolkit with minimal dependencies.
Potential applications	Translation
Software & IPR status	Nematus software with BSD-3 licence is used for training neural models. We use an efficient decoder for generating the translations called Marian which has an MIT licence.
Terms & conditions of use	Open source
Performance requirements	Marian can run on CPU or GPU. It can currently translate at around 5000 WPM on an 8GB GPU. It can translate at about 140 words per minute on one CPU with 16 threads.
Further documentation	https://github.com/rsennrich/nematus https://marian-nmt.github.io/
Alternatives	In the SUMMA project we have developed two open source NMT models which cover all requirements.
Key contact(s)	Alexandra Birch

Table 43: DeepTagger for Topic Labelling component details

Component name	DeepTagger for Topic Labelling
Inputs from	Directly from Text Sources (for written documents) or from Speech Recognition (ASR) for spoken ones, possibly after Text Normalisation (optionally: Machine Translation)
Outputs to	(unspecified for now)
Component lead partner	Idiap
Component contributors	Nikolaos Pappas and Andrei Popescu-Belis
Brief description	Hierarchical neural network for multilingual document labelling with topics or keywords.
What it does (more detail)	The component learns from a set of documents in several languages, with labels (of various granularity) in several languages. The component learns how to label new documents in any language, with labels from all languages available in the training data.
How it works (more detail)	Uses a state-of-the-art architecture for multilingual document modelling, based on hierarchical neural networks with attention. There are three possible options for multilingual training over disjoint label spaces: sharing the parameters at each layer of the network, sharing the attention mechanisms at each layer, and sharing both. The component was tested on a corpus of 600k articles from DW, demonstrating that the architectures are useful for transferring knowledge to low-resource languages, and also improve over monolingual models on the full-scale data.
Key innovative aspects	Use of deep neural networks, sharing of network components across languages.
Potential applications	Topic labelling in multilingual settings, especially when some of the languages are under-resourced.
Software & IPR status	The component makes use of the deep learning library Keras running on top of Theano (open source MIT Licence).
Terms & conditions of use	Open source
Performance requirements	Training requires significant GPU computation. Testing (i.e. running) on new unlabelled data is significantly faster.
Further documentation	https://arxiv.org/pdf/1707.00896.pdf
Alternatives	Unknown
Key contact(s)	Nikolaos Pappas, Andrei Popescu-Belis

Table 44: Storyline Clustering component details

Component name	Storyline Clustering
Inputs from	Text Sources with or without extra metadata from Topic detection or Entity Linking. ASR output already segmented as news articles.
Outputs to	Text Store, Summarization
Component lead partner	Priberam
Component contributors	Priberam
Brief description	Groups stories from the incoming text stream (possibly after ASR or MT) into multilingual storylines
What it does (more detail)	This component finds groups of related documents from an input stream and clusters them into storylines. The module can generate both monolingual clusters and crosslingual clusters, where related documents from different languages are assigned to the same storyline.
How it works (more detail)	The system maintains a pool of active storylines, according to their popularity in the stream. When a new document arrives into the system, it's assigned to the nearest cluster according to a similarity measure, or a new cluster is created if a threshold criterion is not met. The system supports a wide range of configurable features, both sparse and dense, such as Tokens, Lemmas, Named Entities, Ontologies, Timestamps, and Embeddings. Furthermore, the system can learn weights for these features in a training corpus with an optimizer such as SVM-rank. For the crosslingual cluster assignments, the system also offers the option of using any pre-trained crosslingual embeddings.
Key innovative aspects	Scalable crosslingual clustering in an online setting.
Potential applications	Online clustering of multilingual text documents.
Software & IPR status	Entirely written on C++ no external dependencies.
Terms & conditions of use	Licensed
Performance requirements	This component can run on a multi-core CPU host. The current implementation (not final) clusters about 250 news articles per minute.
Further documentation	To be available
Alternatives	Unknown
Key contact(s)	Afonso Mendes and João Prieto

3.4.3 WP4 Automatic Knowledge Base Construction

Table 45: Entity Linking component details

Component name	Entity Linking
Inputs from	Named Entity Recognition
Outputs to	Text Store, Knowledge Base Constructor
Component lead partner	Priberam
Component contributors	Priberam
Brief description	Connects previously recognised mentions mentions to a Knowledge Base. The current implementation uses Wikipedia.
What it does (more detail)	The component receives a text marked with mentions to link and disambiguate the possible knowledge base entities for the current text.
How it works (more detail)	The system calculates a set of features like: similarity of the document and the Wikipedia text for the entities, the coherence of the entities, the concurrent entities etc, and uses a model trained with SVM-rank to rank the possible candidates. It also performs NIL detection.
Key innovative aspects	The system was tested with the main corpora on the literature and performed at the state of the art.
Potential applications	Entity Linking with other Knowledge Bases, like KB for Health, Finance etc..
Software & IPR status	The software written in C++ and is standalone. The knowledge base is open source (Wikipedia, Freebase and DBPedia)
Terms & conditions of use	Licensed
Performance requirements	The current version handles 0.5 news articles per second sequentially. The implementation is multi-threaded and can be deployed over several machines.
Further documentation	
Alternatives	Other publicly available Entity Linkers
Key contact(s)	João Prieto and Afonso Mendes

Table 46: TurboParser component details

Component name	TurboParser
Inputs from	Text Stream
Outputs to	Entity Linking, Knowledge Base Construction, Fact Checking, Summarization
Component lead partner	Priberam
Component contributors	Priberam
Brief description	This component is a full NLP-pipeline that is responsible for all the basic linguistic analysis, namely: Tokenization, sentence splitting, Part of Speech Tagging, Named Entity Recognition, Dependency Parsing, Semantic Parsing and Co-reference Resolution.
What it does (more detail)	TurboParser allows the user to learn a dependency parser / POS tagger / semantic parser / entity tagger / coreference resolver from a treebank, run such parser / tagger / resolver on new data and evaluate the results against a gold-standard.
How it works (more detail)	TurboParser learns to predict tags based on labelled input data using one of its available training algorithms (the Perceptron, MIRA, SVM-MIRA,CRF-MIRA, SVM-SGD and CRF-SGD). Such predictive model is constructed based on a set of features, configurable for each task, such as tokens, lemmas, POS tags, dependency tree tags and gazetteers.
Key innovative aspects	TurboParser implements several step of the NLP pipeline with near state of the art performance both in terms of accuracy and speed.
Potential applications	Most tasks that require linguistic analysis
Software & IPR status	AD3, glog, gflags
Terms & conditions of use	Open source
Performance requirements	470 tokens/sec on single core; 1843 tokens/sec in a multicore CPU (considering the full NLP pipeline).
Further documentation	https://github.com/andre-martins/TurboParser
Alternatives	NLTK and other publicly available
Key contact(s)	Afonso Mendes and André Martins

Table 47: POST/DP/CR/NER Docker component details

Component name	POST/DP/CR/NER Docker
Inputs from	Integration
Outputs to	Knowledge Base Constructor
Component lead partner	Priberam
Component contributors	Priberam
Brief description	Handles pre-processing and executes the NLP pipeline required for the Knowledge Base Constructor module. Responsible to the integration for the NLP pipeline into Docker Compose.
What it does (more detail)	Executes part-of-speech tagging (POST), dependency parsing (DP), coreference (CR) and named entity recognition (NER) on a collection of input documents and outputs in the CoNLL format.
How it works (more detail)	Manages the link between the NLP pipeline tools and the tntegration platform.
Key innovative aspects	
Potential applications	Integration for the NLP pipeline into Docker Compose.
Software & IPR status	Docker
Terms & conditions of use	Open source
Performance requirements	470 tokens/sec on single core; 1843 tokens/sec in a multicore CPU.
Further documentation	Available with the code
Alternatives	
Key contact(s)	Afonso Mendes and Pedro Balage

Table 48: Knowledge Base Constructor component details

Component name	Knowledge Base Constructor
Inputs from	Named Entity recogniser, Entity Linking
Outputs to	Integration
Component lead partner	UCL
Component contributors	UCL, USFD
Brief description	KBC will extract facts (relationships holding between pairs of entities) and populate a knowledge base.
What it does (more detail)	We take a set of documents containing entity mentions and identify the dependency paths linking pairs of entities. From this we populate missing entries in a knowledge base with new facts.
How it works (more detail)	We construct a matrix of entity-pairs versus relations and surface patterns. Factorising this matrix, by projecting down into a lower dimensional space of latent features, allows us to find correlations between surface patterns and knowledge base relations. Thus, given a document containing these surface patterns linking pairs of entities, we can determine the most relevant KB relations.
Key innovative aspects	Our universal schema approach treats knowledge base facts and textual surface patterns in a unified manner.
Potential applications	The approach can be applied to any Knowledge Base combined with Text Corpus.
Software & IPR status	Python (Open Source) TensorFlow (Open Source)
Terms & conditions of use	Open Source
Performance requirements	Our implementation works most efficiently using GPUs. Currently, processing of large datasets could be prohibitive in terms of time and memory usage. Future work will reduce these overheads.
Further documentation	Riedel et al., 2013, Relation Extraction with Matrix Factorization and Universal Schemas, NAACL
Alternatives	Stanford KBP
Key contact(s)	Jeff Mitchell

Table 49: Component Sentence-level Fact Checker component details

Component name	Sentence-level Fact Checker
Inputs from	Automatic Speech Recognition and Clustering and Topic
Outputs to	Integration
Component lead partner	USFD
Component contributors	USFD, UCL
Brief description	It will provide a truth assessment for a sentence combined with evidence from a knowledge base
What it does (more detail)	Given a sentence it attempts to find a table in a knowledge base related to the claim; given a table is found, it attempts to identify an entry that either confirms or denies the claim.
How it works (more detail)	It works by training a classifier to decide whether a column in the table is related to the claim in the sentence. The training data are generated automatically using search queries.
Key innovative aspects	The key innovative aspect is that the automatic generation of training data means that it can be easily extended to new relations beyond the ones we worked on.
Potential applications	It can be used to develop tools to help readers identify false claims in the media they consume.
Software & IPR status	Python (Open Source), Scikit-Learn (Open Source), Stanford CoreNLP (Open Source)
Terms & conditions of use	Open Source
Performance requirements	Runs easily on a laptop
Further documentation	Demo paper accepted for publication at EACL 2017.
Alternatives	Not needed, all open source
Key contact(s)	Andreas Vlachos

Table 50: Multilingual AMR Parser component details

Component name	Multilingual AMR Parser
Inputs from	Incoming news articles after clustering
Outputs to	Storyline generation
Component lead partner	UEDIN
Component contributors	Marco Damonte, Shay Cohen
Brief description	A semantic parser that takes as input text and outputs a canonical representation of the text using abstract meaning representation
What it does (more detail)	Takes as input a sentence (or text), and scans it left-to-right, with the final output being a graph that represents the text in canonical form.
How it works (more detail)	The parser is based on a transition-based system for dependency parsing called ArcEager (hence the name for our parser, AMREager). As mentioned above, it scans the input left to right, and at each point decides how to change a partially-constructed graph according to the new input.
Key innovative aspects	AMR parsing is a young area of research that is now getting increasing attention.
Potential applications	Semantic parsing, question answering, machine translation
Software & IPR status	JAMR aligner (publicly available), Theano (publicly available)
Terms & conditions of use	Open source
Performance requirements	Will run faster on a machine with GPU. Intended to work in server mode, standalone.
Further documentation	http://kinloch.inf.ed.ac.uk/amreager.html
Alternatives	CAMR and JAMR (other publicly available AMR parsers)
Key contact(s)	Shay Cohen

Table 51: English AMR Parser component details

Component name	English AMR Parser
Inputs from	Incoming news articles
Outputs to	Storyline generation or Knowledge Base Population (KBP)
Component lead partner	LETA
Component contributors	Guntis Barzdins, Didzis Gosko
Brief description	A semantic parser that takes as input text and outputs a canonical representation of the text using abstract meaning representation (AMR)
What it does (more detail)	Takes as input a sentence (or text), and scans it left-to-right, with the final output being a graph that represents the text in canonical AMR form.
How it works (more detail)	The parser is based on CAMR parser and applies an input text and output AMR normalisation wrapper to improve the overall smatch score.
Key innovative aspects	Smatch extensions and character-level neural translation improvements to AMR parsing accuracy.
Potential applications	Semantic parsing, question answering, machine translation, KBP
Software & IPR status	CAMR parser (publicly available), JAMR aligner (publicly available), Tensorflow (publicly available)
Terms & conditions of use	Open source
Performance requirements	This parser was developed as part of SemEval-2016, Task 8 where in ensemble with a character-based neural AMR parser it achieved the top result with smatch F1=62% on the SemEval-2016 official scoring set and F1=67% on the LDC2015E86 test set. Docker version can utilise multiple CPUs for faster parsing; neural parsing is excluded due to minor impact on the results.
Further documentation	https://github.com/didzis/CAMR/tree/wrapper
Alternatives	CAMR and JAMR (other publicly available AMR parsers)
Key contact(s)	Guntis Barzdins

Table 52: Multilingual AMR-to-Text Generation component details

Component name	Multilingual AMR-to-Text Generation
Inputs from	Text-to-AMR parsing (incl. Multilingual AMR parsing), AMR-based abstractive summarization
Outputs to	SUMMA Platform integration
Component lead partner	LETA
Component contributors	LETA, University of Latvia, IMCS
Brief description	A system that takes an AMR (Abstract Meaning Representation) graph as input and generates a natural language sentence from it, potentially in multiple languages.
What it does (more detail)	For each input AMR graph, the system applies a sequence of tree pattern-matching and transformation rules, acquiring an abstract syntax tree (AST). The AST is then passed to Grammatical Framework for linearization.
How it works (more detail)	Builds on Grammatical Framework (GF) which provides a wide-coverage resource grammar library (RGL) with a language-independent API – a shared abstract syntax. The idea is to transform the AMR graphs into the GF abstract syntax trees (AST), leaving the surface realization (linearization) of ASTs to the existing English resource grammar. Since GF RGL supports many more languages, this approach can automatically extend to multilingual AMR-to-text generation.
Key innovative aspects	Conversion of AMRs into the intermediate ASTs, instead of language-specific parse trees, or direct AMR-to-text generation. The choice of GF allows for further multilingual text generation, preserving grammatical and semantic accuracy.
Potential applications	Abstractive text summarization, question answering, etc.
Software & IPR status	Open source components: Grammatical Framework and its Resource Grammar Library, the Stanford JavaNLP library, JAMR Generator.
Terms & conditions of use	Open source
Performance requirements	No special hardware requirements.
Further documentation	https://github.com/GrammaticalFramework/gf-contrib/tree/master/AMR/AMR-to-text
Alternatives	Not required, open source.
Key contact(s)	Normunds Gruzitis, Didzis Gosko, Guntis Barzdins

Table 53: CCA-based Extractive Summarization component details

Component name	CCA-based Extractive Summarization
Inputs from	Data collection (textual sources), Machine Translation and Clustering and Topic
Outputs to	Integration
Component lead partner	UEDIN
Component contributors	Nikos Papasarantopoulos, Shashi Narayan, Shay Cohen
Brief description	A system that performs summarization by selecting the most important (for summary purposes) sentences from input documents.
What it does (more detail)	Takes as input a document and outputs a set of sentences from that document that constitute its summary. Output summary consists of up to three sentences.
How it works (more detail)	The main idea behind this system is the joint representation of documents and summaries in one low-dimensional space. Projection to that space is achieved by using the technique of Canonical Correlation Analysis (CCA). Sentences are selected for the final summary according to their proximity to the document in this space.
Key innovative aspects	No need for human generated or (semi-automatically generated) annotations specifically made for extractive summarization.
Potential applications	Any application that needs to distil large documents to a small summary that will give readers the gist of the document without them having to read the whole document.
Software & IPR status	Mainly written in Python, with some components in Matlab; Matlab components can easily be ported to Python.
Terms & conditions of use	Open source
Performance requirements	Depending on the amount of training data, it may require substantial amount of memory to train, as it may operate on large matrices.
Further documentation	NA
Alternatives	Neural Network Extractive Summarizer
Key contact(s)	Nikos Papasarantopoulos, Shay Cohen

Table 54: Neural Network Extractive Summarizer details

Component name	Neural Network Extractive Summarizer
Inputs from	WP2 data collection (textual sources) and WP4 (clustering and topic)
Outputs to	WP6 integration
Component lead partner	University of Edinburgh
Component contributors	University of Edinburgh
Brief description	Neural network extractive summarizer is one of the rear component of the SUMMA pipeline. It generates extractive summaries of news documents.
What it does (more detail)	Given a document, neural extractive summarizer generates a summary of the document by identifying and concatenating (2–3) salient sentences.
How it works (more detail)	The core components are a neural network-based hierarchical document encoder and a hierarchical attention-based sentence extractor. Our hierarchical document encoder combines convolution neural networks and recurrent neural networks to derive the document meaning representation. Our sentence extractor then labels each sentence in the document for relevance in the summary.
Key innovative aspects	Because of its neural-based architecture, it circumvents human-engineered features using continuous sentence features. The proposed architecture uses this side information such as title and image captions along with the document to identify salient sentences in the document.
Potential applications	We model document summarization as a sequence labelling task. Hence, it could be easily adapted to any sequence labelling task such as sentence or document compression, and query-based summarization.
Software & IPR status	It uses TensorFlow.
Terms & conditions of use	Open source
Performance requirements	Standalone mode, ideally with GPU system
Further documentation	We have submitted our paper to ACL 2017 for blind peer-reviewing.
Alternatives	NA
Key contact(s)	Shashi Narayan, Shay Cohen

Table 55: Extractive Summarization ILP based component details

Component name	Extractive Summarization ILP based
Inputs from	Storyline clustering
Outputs to	WP6 integration
Component lead partner	Priberam
Component contributors	Priberam
Brief description	The ILP based extractive summarizer identifies and extracts as summary the most relevant sentences from the given texts.
What it does (more detail)	Given a set of documents, the summarizer identifies and extracts the key sentences from the source documents which cover the most information possible from source texts constrained to a maximum number of words.
How it works (more detail)	The summarizer uses an Integer Linear Programming (ILP) formulation to solve the optimization problem of maximising the coverage of the terms present in the document constrained to a maximum number of words (TSP problem). This model ensures the summary keep the main information from the source texts. Before the ILP, the text is processed by a natural language processing pipeline that extracts the main terms for each document and measure their informativeness. A machine learning algorithm calculate scores for features computed over each term.
Key innovative aspects	ILP based coverage maximization algorithm
Potential applications	The approach could be used in applications where is necessary to identify the gist of the documents. Examples of such applications are: terminology extraction; text compression; headline generation; among others.
Software & IPR status	No external dependencies.
Terms & conditions of use	Licensed
Performance requirements	Current API is able to handle an average of 23 documents per second
Further documentation	NA
Alternatives	NA
Key contact(s)	Afonso Mendes and João Prieto

Table 56: Extractive Summarization Semantic ILP Based component details

Component name	Extractive Summarization Semantic ILP Based
Inputs from	Storyline clustering, TurboParser
Outputs to	WP6 Integration
Component lead partner	Priberam
Component contributors	Priberam
Brief description	From a set of documents clustered as a storyline builds an extractive summary by selection a set of relevant sentences from the documents.
What it does (more detail)	The component use an ILP formulation to solve the extractive summarization problem. Semantic features are used to determine the main concepts.
How it works (more detail)	The summarizer uses an Integer Linear Programming (ILP) formulation to solve the optimization problem of maximising the coverage of the terms present in the document constrained to a maximum number of words (TSP problem). This model ensures the summary keep the main information from the source texts. Before the ILP, the text is processed by a natural language processing pipeline that extracts the main concepts for each document and measure their informativeness. A semantic parser (SRL or AMR) is used to determine the concept while a machine learning algorithm scores they.
Key innovative aspects	Use of semantic level of analysis with ILP for automatic summarization
Potential applications	The approach could be used in applications where is necessary to identify the gist of the documents. Examples of such applications are: terminology extraction; text compression; headline generation; among others.
Software & IPR status	No external dependencies.
Terms & conditions of use	Licensed
Performance requirements	NA
Further documentation	NA
Alternatives	NA
Key contact(s)	Afonso Mendes and João Prieto

Table 57: Abstractive Summarization component details

Component name	Abstractive Summarization
Inputs from	Storyline clustering, TurboParser, AMR Parser
Outputs to	WP6 Integration
Component lead partner	Priberam
Component contributors	Priberam
Brief description	From a set of documents clustered as a storyline builds a summary by selecting the most relevant information from the documents. The text of this summary is generated by the system that may or may not have words in common with the source texts.
What it does (more detail)	The component first identifies and select the most relevant concepts used in the source texts and their relation. Second, the component formulate sentences connecting the concepts and relations for summary production.
How it works (more detail)	The summarizer may use ILP, semantic or neural approaches in order to select the most relevant concepts and relations available in the source texts. Then, an abstractive step possibly based on grammar framework or neutral network approaches transforms these concepts and relations into well structured sentences that summarizes the key concepts in the source documents.
Key innovative aspects	Abstractive summarization remains as a challenge in automatic summarization. This component may innovate by the design of new architectures for this task.
Potential applications	The approach could be used in applications where is necessary to generate short texts. Examples of such applications are: terminology extraction; text compression; headline generation; among others.
Software & IPR status	No external dependencies.
Terms & conditions of use	Licensed
Performance requirements	NA
Further documentation	NA
Alternatives	NA
Key contact(s)	Afonso Mendes and João Prieto

Table 58: Twitter Sentiment Analyser component details

Component name	Twitter Sentiment Analyser
Inputs from	WP2 data collection
Outputs to	WP6 Integration
Component lead partner	Priberam
Component contributors	Priberam
Brief description	Story Sentiment Analysis provides an overview of the sentiment polarity in respect to a collection of tweets from the storylines.
What it does (more detail)	The component receives a collection of tweets and calculate the overall sentiment transmitted by the posts.
How it works (more detail)	The sentiment analysis component uses a SVM classifier with lexicon and linguistic features to detect where a text may be classified as positive, negative or neutral.
Key innovative aspects	During the development of the component, some innovative aspects may be further investigated: possible new architectures; specific sentiment lexicons for the domain; particularities of short texts.
Potential applications	NA
Software & IPR status	No external dependencies
Terms & conditions of use	Licensed
Performance requirements	NA
Further documentation	NA
Alternatives	NA
Key contact(s)	Afonso Mendes and João Prieto

3.4.4 WP6 Integration

Table 59: SUMMA Platform Integration component details

Component name	SUMMA Platform Integration
Inputs from	All SUMMA components
Outputs to	All SUMMA components
Component lead partner	LETA
Component contributors	All SUMMA partners
Brief description	SUMMA Platform integration links all SUMMA components into a unified end-user application.
What it does (more detail)	SUMMA Platform allows to ingest video/audio/text news items in any of the supported languages and to transcribe/translate these news to English. Additional services of the SUMMA platform is rich semantic markup of the translated news (NER/NEL, entity linking, clustering, summarisation, sentiment analysis) providing additional tools for news filtering and processing.
How it works (more detail)	The core components of SUMMA Platform integration are data feed modules tailored to video/audio/text data ingestion from BBC and DW existing systems, JSON database, NLP job queue, and UX user interface. The integration platform is built around the central JSON database. Newly ingested stories are converted into the single JSON file containing title, body, URL to audio/video, metadata (timestamp, source channel, author etc.). This JSON file is subsequently enriched by NLP modules (other SUMMA components) with additional fields: ASR transcript, translation, NER/NEL markup, clustering and summarisation outputs. Finally the web-based user interface module queries JSON database for information to be displayed in the UX. To ensure BigData scalability of the SUMMA platform, it is able to fire multiple copies of the NLP modules in the cloud to handle the load in near-realtime.
Key innovative aspects	Platform integration and scaling is based entirely on Docker containers. This allows unprecedented independence for component developers. It also ensures seamless portability and scaling of the system through Docker Compose. Human-readable REST API and JSON interconnects used to implement industry-standard Microservices architecture.

Potential applications	Other scalable NLP pipelines. The integration is agnostic to actual NLP modules or services used in the pipeline.
Software & IPR status	SUMMA Platform integration is based entirely on industry-standard open-source components: Docker, Docker Compose, Rethink DB, Elastic Search, RabbitMQ.
Terms & conditions of use	The IP added by LETA concerns the specific configuration files, data ingestion and format conversion modules, overall system structure, and UX user interface. The integration platform as such will be open-sourced after cleaning from SUMMA-specific URLs and similar sensitive information.
Performance requirements	Cloud-based deployment is required for BigData applications and scaling is dependent on the actual mix of video un text feeds. The system is tested to handle 200 live streams simultaneously in the cloud environment with ASR component operating at 1/4 real-time being the key constraint. The minimal requirement for testing purposes is a PC with 32GB of RAM and 200GB of HDD.
Further documentation	http://summa-project.eu
Alternatives	NA
Key contact(s)	Guntis Barzdins, Renars Liepins, Didzis Gosko

Table 60: Balsamiq Wireframes component details

Component name	Balsamiq Wireframes
Inputs from	NA
Outputs to	NA
Component lead partner	BBC NI
Component contributors	BBC NI, DW
Brief description	Wireframes mockups for interface for WP6.
What it does (more detail)	Represents the web-based interface so that this might be commented upon.
How it works (more detail)	NA
Key innovative aspects	None
Potential applications	Discussion of web-based interface
Software & IPR status	Created using Balsamiq
Terms & conditions of use	
Performance requirements	Can be viewed within the Balsamiq software or compiled as a PDF for viewing.
Further documentation	
Alternatives	None
Key contact(s)	David Sheppey

Table 61: HTML Mockups component details

Component name	HTML Mockups
Inputs from	NA
Outputs to	NA
Component lead partner	BBC NI
Component contributors	BBC NI
Brief description	HTML template created from Balsamiq mockup.
What it does (more detail)	Acts as a starting point for the prototypes to be created by Leta for WP6.
How it works (more detail)	NA
Key innovative aspects	UX design
Potential applications	None
Software & IPR status	Contains JQuery library and will possibly contain other libraries as it is developed further.
Terms & conditions of use	JQuery is open source.
Performance requirements	NA
Further documentation	None
Alternatives	None
Key contact(s)	David Sheppey

3.5 Multi-Strand Exploitation

There are two main use cases for SUMMA as an integrated product. Further use cases will be explored at the next full project meeting where the Exploitation Committee will convene to explore potential use of the SUMMA project outputs by using combinations of components to address known and perceived user needs informed by partners' own knowledge and information gathered from user group meetings. The two use cases currently under consideration are as follows:

3.5.1 BBC integrated tool

3.5.1.1 What BBC does The British Broadcasting Corporation (BBC) is a UK-based international public service broadcaster with a worldwide audience of over 250 million. Its use case for SUMMA relates to the activities of its BBC Monitoring division.

BBC Monitoring scours openly available media sources around the world to provide news, information and insight to BBC journalists, UK government customers and commercial subscribers, allowing them to make better, more informed decisions.

It keeps track of broadcast, press and social media sources in multiple languages, from over 150 countries. It sifts, translates and reports breaking news, media behaviour and emerging trends, with direct insight from local sources. Services include:

- Round-the-clock monitoring of media sources in multiple languages, alerting subscribers to key stories and themes as they emerge
- Accurate, quick of foreign-language media reports, important speeches, statements and social media content
- Analysis of media and social media behaviour, based on expert understanding of the local context
- Direct access to specialists in specific areas of interest
- Personalised feeds, notifying subscribers of relevant content as it is published
- A full reference section with biographies, guides to local media environments, and key organisations

BBC Monitoring provides access to these services to internal BBC customers, government and commercial subscribers through a recently launched web portal. The level of access depends on the level of subscription

3.5.1.2 How BBC does it now The business model for BBC Monitoring is based firmly on having a team of experts in place to analyse and make sense of the media data that is collected from around the world.

BBC Monitoring team has a team of around 220 people who translate information from radio, television, press, news agencies and the internet from 150 countries in more than 70 languages. Staffing levels were recently cut from around 450 people, which makes the case for new technical solutions all the more urgent.

Many companies offer media alerts but the unique selling point of BBC Monitoring is to take the additional step of providing context and analysis as well as manually checking the facts are accurate. Alerts are used to point to the areas on which to focus the further work.

BBC Monitoring is available via a web interface www.bbc.co.uk/monitoring to all BBC staff and subscribers. Features include:

- News updates
- Country insights
- Thematic insights
- Names in the news (i.e. highlighting people of whom clients may not be aware)
- Government lists (e.g. senior office holders in all countries updated as they change)
- Media review summaries (SUMMA summarisation could help with this but to have something well written requires human intervention)
- Election guides – lists of parties with analysis
- Words as spoken – including historical context

In all cases the raw information could be gathered by a machine but the data will only ever be broadly reliable.

Despite the huge advances in machine learning and artificial intelligence there is as yet no suggestion that computers could replace what BBC Monitoring staff achieve day to day. However, machine tools can greatly enhance the breadth, depth and quality of output.

At the moment each specialist will monitor up to four audio visual sources at any one time, switching to one source if a noteworthy event occurs on one of those channels, and using recently recorded footage to catch up where significant events have happened on multiple channels simultaneously during quieter periods.

The amount of analysis that can take place is finite, which is where SUMMA offers some great opportunities .

3.5.1.3 How BBC plans to change this with SUMMA

The challenge for BBC Monitoring is that significant staff time is spent on basic data gathering and analysis in an environment where the number of sources is constantly increasing. This allows less time for staff to spend working on higher value products.

BBC Monitoring is in the process of transforming into a digital journalism organisation. The idea is not to replace any one person with a machine but to use the existing team more efficiently. By automating the first stage in the monitoring process time will be freed up to work on the manual interventions that offer subscribers greater value.

The unique opportunity from SUMMA platform is in providing analysis of non-English audio visual sources. The tool can take much of the legwork and stress of monitoring multiple channels simultaneously.

Individuals in the BBC Monitoring work in different ways depending on the areas they cover and the events taking place at that time. The SUMMA prototype as it now stands enables each user to customise alerts based on multiple facets (e.g. looking for key terms like a person or concept).

Because BBC Monitoring has great confidence in the ability and value of its specialists it does not foresee an issue with the SUMMA product being made commercially available to other organisations in parallel.

3.5.2 DW integrated tool

The SUMMA platform addresses several of DW's major challenges by speeding up existing workflows considerably.

First, it speeds up the exchange of news stories between eight of its 30 languages. One or several tailored feed groups can be set up to monitor the news output of other DW language sections (or, indeed, the entire range of DW's output). As every incoming media item is translated into DW's lingua franca, English, language departments can be informed about new DW content within seconds of the latter's publication. Currently, this process, including s, can take several hours.

Second, the ability to monitor trending topics which emerge on external news sources across several languages.

Third, the ability to compare emerging news trends on eternal sources with DW's current output is an ideal opportunity to spot blind spots in DW's news coverage. As a consequence, journalist can spot at a glance what stories they are missing.

Finally, SUMMA's ability to summarize storylines and visualize trends via its dashboard components makes it an ideal tool for inter-departmental meetings, where fast access to information on what stories are currently being published by the organisation is vital.

4 Conclusion

With 27 dissemination events attended and 39 publications produced, the SUMMA project has placed significant emphasis on dissemination activities during this reporting period. The work on Dissemination and Exploitation during M1-M18 provides a solid foundation on which to build.

Within the project the partners have consulted internally as to how output of the project may be best exploited to meet some real business needs, with an initial focus on the BBC and DW use cases.

During M19-M36 the Exploitation Committee will continue to refine exploitation plans for the SUMMA platform while wider dissemination activities will continue to ensure SUMMA technology is socialised with the wider community.

ENDPAGE

SUMMA

H2020-ICT-2015 688139

D8.3 Dissemination and Exploitation Plan and Initial Report