# Scalable Understanding of Multilingual MediA (SUMMA)

http://www.summa-project.eu

## H2020 Research and Innovation Action
## Number: 688139
## D6.3 – Intermediate report on platform

| **Nature** | Report | **Work Package** | WP6 |
|---|---|---|---|
| **Due Date** | 31/07/2017 | **Submission Date** | 31/07/2017 |
| **Main authors** | Guntis Barzdins (LETA) | | |
| **Co-authors** | Renars Liepins (LETA), Didzis Gosko (LETA) | | |
| **Reviewers** | Andrew Secker (BBC), Andreas Giefer (DW) | | |
| **Keywords** | platform testing, agile development, media monitoring | | |
| **Version Control** | | | |
| v0.1 | **Status** | Draft | 07/07/2017 |
| v0.2 | **Status** | Draft | 20/07/2017 |
| v1.0 | **Status** | SemiFinal | 25/07/2017 |
| v1.1 | **Status** | Final | 30/07/2017 |

# Contents

**1 Introduction 5**

**2 Added UX functionality 6**

2.1 Entity view . . . 6

2.2 Sorting and Filtering toolbar . . . 8

2.3 Summary and Topic fields . . . 9

2.4 Additional Navigation options . . . 10

2.5 Feed activity monitoring . . . 11

**3 Backend improvements 12**

**4 Conclusion 14**

## List of Figures

1 Entity view for the named entity ”Australia” . . . . . . . . . . . . . . . . . . . . . . 6
2 Entity search view accessible from the Entities tab on the left . . . . . . . . . . . . 7
3 Named entities in MediaItem text are highlighted in different colours according to their Type . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
4 Sorting and Filtering toolbar shows language selection options . . . . . . . . . . . 8
5 MediaItem view enriched with Summary and Prev/Next live-feed chunk navigation buttons . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
6 Story view is enriched with both Summary and Topics fields . . . . . . . . . . . . 9
7 The hourly bars in the Trending graphs now are clickable . . . . . . . . . . . . . . 10
8 Feed activity graph . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
9 Architecture of SUMMA Platform, Version 1.1 . . . . . . . . . . . . . . . . . . . . 13

# Abstract

This report describes the experiences gained after the initial integration of the functionality and interfaces in the SUMMA Platform, Version 1.0 released 3 months earlier and described in ”D6.2 – Release of initial platform”. These 3 months have been spent on intense usability tests by the DW and BBC end-users, weekly JIRA sprints prioritising required bug-fixes and additional functionality, and the actual further development of the SUMMA Platform. The document provides an overview of the key improvements incorporated into the SUMMA Platform, Version 1.1 as the result of this process.

# 1 Introduction

In the 3 months following the initial release of SUMMA Platform, Version 1.0 (described in ”D6.2 – Release of initial platform”), intense usability tests by end-users and weekly JIRA sprints prioritising required bug-fixes and / or additional functionality development have been conducted. Also, several new NLP modules have been developed and integrated into the platform making it truly multilingual, with ASR and MT support currently for 6 input languages: EN, AR, DE, ES, RU, LV, followed by deep NLU processing of the translated English text.

This development cycle resulted in the updated SUMMA Platform Version 1.1, incorporating feedback from both DW and BBC end-users through the process described in deliverables ”D1.2 - Initial Prototype Report”, ”D6.4 - External Media Monitoring Demonstrator betaversion prototype”, ”D7.2 - Initial Evaluation Report”.

Overall, SUMMA Platform Version 1.0 performed well in the user-tests and in some aspects even exceeded expectations. Specifically, SUMMA Platform Version 1.0 was able to handle the full NLP pipeline for 11 simultaneous live video streams on a single server with only the ASR processing distributed in the cloud, which is encouraging news for the scalability tests scheduled in the later stages of the project. Two instances of the SUMMA Platform were used for end-user testing:

- the SUMMA Platform installed at LETA handled multilingual DW news feed and was used for end-user tests at DW.
- the SUMMA Platform installed at UEDIN handled both multilingual BBC and DW news feeds and was used for end-user tests at BBC.

Social media ingestion module developed by BBC NI was added to both SUMMA Platform instances after the end-user testing. The module was tested successfully on RSS feeds from dozens of news outlets from around the world; due to the large volume of incoming MediaItems, it highlighted the importance of the upcoming scalability improvements and testing.

This deliverable documents the improvements incorporated in the SUMMA Platform Version 1.1 since the initial release of Version 1.0.

# 2 Added UX functionality

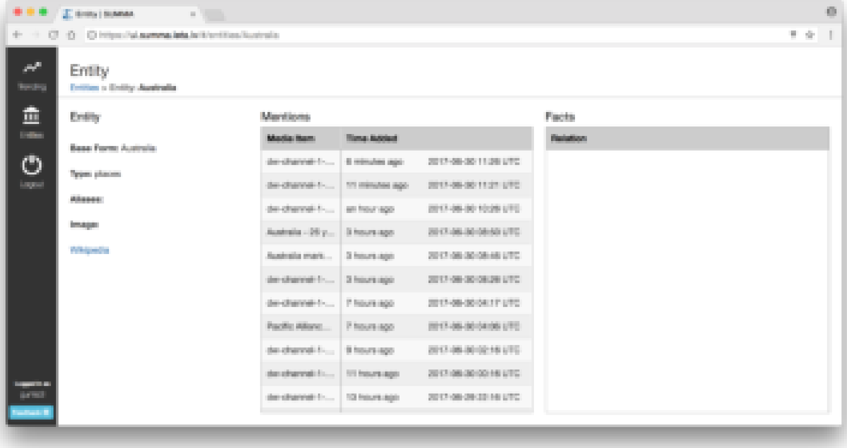
Two factors were driving the need for adding unforeseen functionality to the User eXperience (UX) interface of SUMMA Platform:

- Feedback from end-users about the extra UX functionality needed to make the platform more intuitive and convenient in typical usage scenarios.
- The overall stability and functionality of the platform allowed to accumulate successfully large amounts of MediaItems (11 live TV channels and other news sources from BBC and DW recorded for several weeks) which in turn required to start addressing scalability issues in UX, such as providing more powerful sorting and filtering options. The platform scalability was originally planned to be addressed only later in the project (First Scalability test of the SUMMA Platform is planned on M24).

Below the key UX improvements incorporated in the SUMMA Platform Version 1.1 are described.

## 2.1 Entity view

The lack of the Entities tab in the main menu on the left side of the screen was clearly an omission in Version 1.0, likely caused by the overwhelming complexity of wireframing for the then non-existent SUMMA Platform. The need for explicit visualisation of Entities known to the Named Entity Recognition / Named Entity Linking (NER/NEL) module of the SUMMA Platform became evident as soon as end-users started to use the platform and tried to click on the Named Entities highlighted in the MediaItem text (coloured text in Figure 3). An obvious user expectation was that clicking on a Named Entity should open another view showing more information about this entity along with links to other MediaItems mentioning the same Named Entity. This is now implemented in the new Entity view (Figure 1).



Entity

Entities » Entity: Australia

Entity

Base Form: Australia

Type: places

Aliases:

Image:

Wikipedia

Mentions

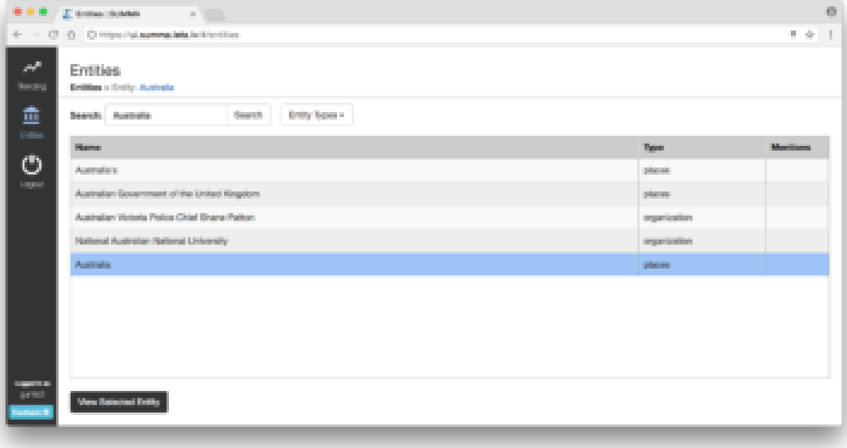
| Media Item | Time Added | |
|---|---|---|
| de-channel-1-... | 6 minutes ago | 2017-06-30 11:26 UTC |
| de-channel-1-... | 11 minutes ago | 2017-06-30 11:21 UTC |
| de-channel-1-... | an hour ago | 2017-06-30 10:26 UTC |
| Australia - 26 y... | 3 hours ago | 2017-06-30 08:50 UTC |
| Australia mark... | 3 hours ago | 2017-06-30 08:46 UTC |
| de-channel-1-... | 3 hours ago | 2017-06-30 08:26 UTC |
| de-channel-1-... | 7 hours ago | 2017-06-30 04:17 UTC |
| Pacific Allianc... | 7 hours ago | 2017-06-30 04:06 UTC |
| de-channel-1-... | 9 hours ago | 2017-06-30 02:16 UTC |
| de-channel-1-... | 11 hours ago | 2017-06-30 00:16 UTC |
| de-channel-1-... | 13 hours ago | 2017-06-29 22:16 UTC |

Facts

| Relation |
|---|

**Figure 1:** Entity view for the named entity "Australia"

This in turn suggested the need to add a more direct access to the Entity view through the dedicated Entities tab on the left side of the screen and Entity search (Figure 2) as an alternative news exploration entry point besides the Trending tab provided originally.



Entities

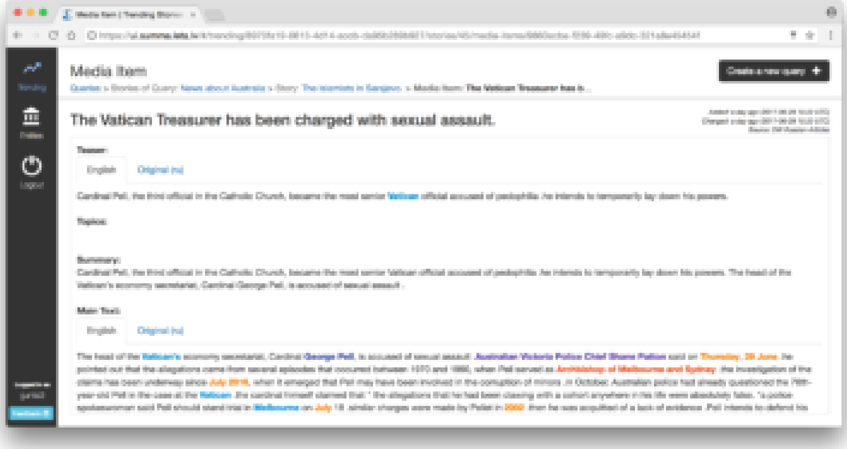
Entities » Entity: Australia

Search: Australia Search Entity Types ▾

| Name | Type | Mentions |
|---|---|---|
| Australia's | places | |
| Australian Government of the United Kingdom | places | |
| Australian Victoria Police Chief Shane Patton | organization | |
| National Australian National University | organization | |
| Australia | places | |

View Selected Entity

Figure 2: Entity search view accessible from the Entities tab on the left



Media Item

Queries » Stories of Query: News about Australia » Story: The Islamists in Sarajevo. » Media Item: The Vatican Treasurer has b...

The Vatican Treasurer has been charged with sexual assault.

Teaser:

English Original (ru)

Cardinal Pell, the third official in the Catholic Church, became the most senior Vatican official accused of pedophilia. he intends to temporarily lay down his powers.

Topics:

Summary:

Cardinal Pell, the third official in the Catholic Church, became the most senior Vatican official accused of pedophilia. he intends to temporarily lay down his powers. The head of the Vatican's economy secretariat, Cardinal George Pell, is accused of sexual assault .

Main Text:

English Original (ru)

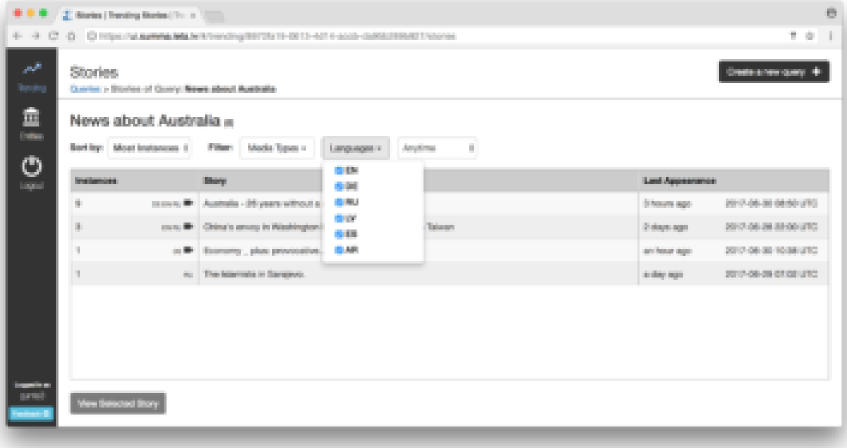
The head of the Vatican's economy secretariat, Cardinal George Pell, is accused of sexual assault. Australian Victoria Police Chief Shane Patton said on Thursday, 29 June. he pointed out that the allegations came from several episodes that occurred between 1970 and 1980, when Pell served as Archbishop of Melbourne and Sydney . the investigation of the claims has been underway since July 2016, when it emerged that Pell may have been involved in the corruption of minors . In October, Australian police had already questioned the 76th-year-old Pell in the case at the Vatican .the cardinal himself claimed that: " the allegations that he had been clearing with a cohort anywhere in his life were absolutely false. "a police spokeswoman said Pell should stand trial in Melbourne on July 18 .similar charges were made by Pellet in 2002 .then he was acquitted of a lack of evidence .Pell intends to defend his

Figure 3: Named entities in MediaItem text are highlighted in different colours according to their Type

Finally, the Entity view (Figure 1) naturally becomes also the placeholder for Facts (Relations, Attributes) about the specific entity. The facts about entities will eventually be populated from the Knowledge Base Population (KBP) task still in progress as part of WP4.

## 2.2 Sorting and Filtering toolbar

The list of identified Stories is the primary entry point into the platform (accessible by selecting All stories in the Trending tab), and this list grows with the amount of MediaItems fed into the SUMMA Platform. Currently the clustering module is grouping all available MediaItems into the Stories and it has a hard limit of max 1000 Stories (under the assumption that no journalist would be able to follow more Stories). Although Queries have been introduced in the UX for this specific purpose of reducing Story clutter, users seemed to have difficulty formulating a Query before looking at All Stories beforehand. Therefore one of the dominating user-feedback requests was to add a Sorting and Filtering toolbar (Figure 4) to the Stories view.



| Instances | Story | Last Appearance | |
|---|---|---|---|
| 6 | Australia - 26 years without a | 3 hours ago | 2017-06-30 08:50 UTC |
| 3 | China's envoy in Washington ... Taiwan | 2 days ago | 2017-06-28 22:00 UTC |
| 1 | Economy _ plus: provocative... | an hour ago | 2017-06-30 10:38 UTC |
| 1 | The Islamists in Sarajevo. | a day ago | 2017-06-29 07:02 UTC |

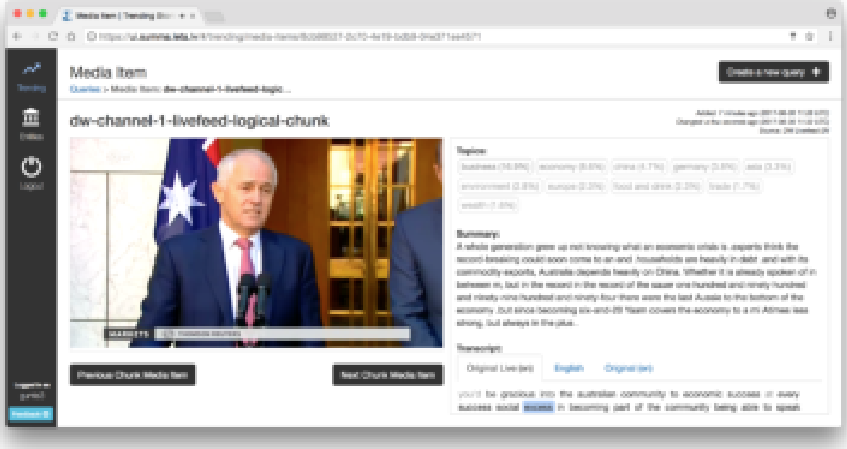
**Figure 4:** Sorting and Filtering toolbar shows language selection options

The added Sorting and Filtering toolbar allows to sort the Stories list by any of the columns, as well as to filter out only stories containing MediaItems in the specific languages, media types, and publishing times.

The Filtering options in the Stories toolbar are complimentary to the Filtering options in the Query definition. Eventually we would like to combine all Filtering options in the same powerful Filtering tool, but have not yet come up with an UX concept for doing so.

### 2.3 Summary and Topic fields

The original design envisioned Topics to be detected only for individual MediaItems, while Summaries were to be generated only for Stories covering multiple MediaItems. In end-user testing both Topic and Summary extraction proved to be highly popular leading to the request for adding a Summary field also to the individual MediaItem view (Figure 5) and adding an aggregated Topics field also to Stories view (Figure 6).



Media Item

Queries > Media Item: dw-channel-1-livefeed-logic...

dw-channel-1-livefeed-logical-chunk

Topics:

business (16.9%) economy (8.6%) china (4.7%) germany (3.8%) asia (3.3%) environment (2.8%) europe (2.3%) food and drink (2.2%) trade (1.7%) wealth (1.6%)

Summary:

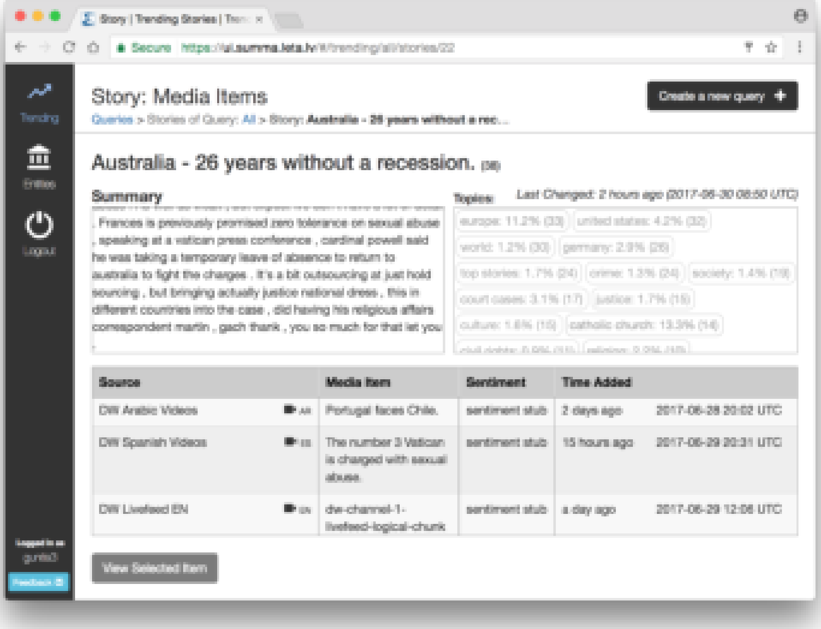
A whole generation grew up not knowing what an economic crisis is. experts think the record-breaking could soon come to an end. households are heavily in debt. and with its commodity exports, Australia depends heavily on China. Whether it is already spoken of in between m, but in the record in the record of the sauer one hundred and ninety hundred and ninety nine hundred and ninety-four there were the last Aussie to the bottom of the economy. but since becoming six-and-26 Years covers the economy to a m Atmes less strong, but always in the plus..

Transcript:

Original Live (en) English Original (en)

you'd be gracious into the australian community to economic success in every success social access in becoming part of the community being able to speak

**Figure 5:** MediaItem view enriched with Summary and Prev/Next live-feed chunk navigation buttons


Story | Trending Stories | Tren
Secure https://ui.summa.leta.lv/#/trending/all/stories/22
Trending
Entities
Logout
Logged in as
gunts3
Feedback
Create a new query
View Selected Item

Story: Media Items

Queries > Stories of Query: All > Story: Australia - 26 years without a rec...

Australia - 26 years without a recession. (38)

Summary

. Frances is previously promised zero tolerance on sexual abuse , speaking at a vatican press conference , cardinal powell said he was taking a temporary leave of absence to return to australia to fight the charges . It's a bit outsourcing at just hold sourcing , but bringing actually justice national dress , this in different countries into the case , did having his religious affairs correspondent martin , gach thank , you so much for that let you .

Topics: Last Changed: 2 hours ago (2017-06-30 08:50 UTC)

europe: 11.2% (33) united states: 4.2% (32) world: 1.2% (30) germany: 2.9% (26) top stories: 1.7% (24) crime: 1.3% (24) society: 1.4% (19) court cases: 3.1% (17) justice: 1.7% (15) culture: 1.6% (15) catholic church: 13.3% (14) civil rights: 0.9% (11) religion: 9.2% (10)

| Source | | Media Item | Sentiment | Time Added | |
|---|---|---|---|---|---|
| DW Arabic Videos | AR | Portugal faces Chile. | sentiment stub | 2 days ago | 2017-06-28 20:02 UTC |
| DW Spanish Videos | ES | The number 3 Vatican is charged with sexual abuse. | sentiment stub | 15 hours ago | 2017-06-29 20:31 UTC |
| DW Livefeed EN | EN | dw-channel-1-livefeed-logical-chunk | sentiment stub | a day ago | 2017-06-29 12:06 UTC |

**Figure 6:** Story view is enriched with both Summary and Topics fields

## 2.4 Additional Navigation options

With end-users actively navigating around the SUMMA Platform UX it has become evident that some additional navigation shortcuts could vastly improve the user experience. Two navigation shortcuts have been added:

- Trending graphs (Figure 7) now are clickable and lead to the list of MediaItems corresponding to the specific bar in the Trending graph.
- Live TV or radio streams are automatically cut into 5 minute chunks each becoming an independent MediaItem (Figure 5). Buttons for navigating to the Previous/Next 5 minute chunk MediaItem have been added to enable linear navigation in the recorded TV and radio streams.

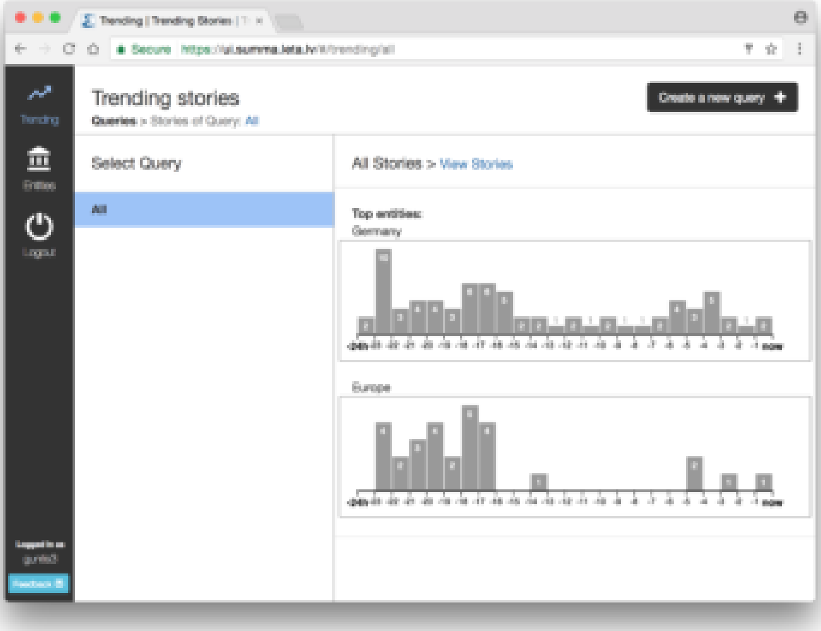
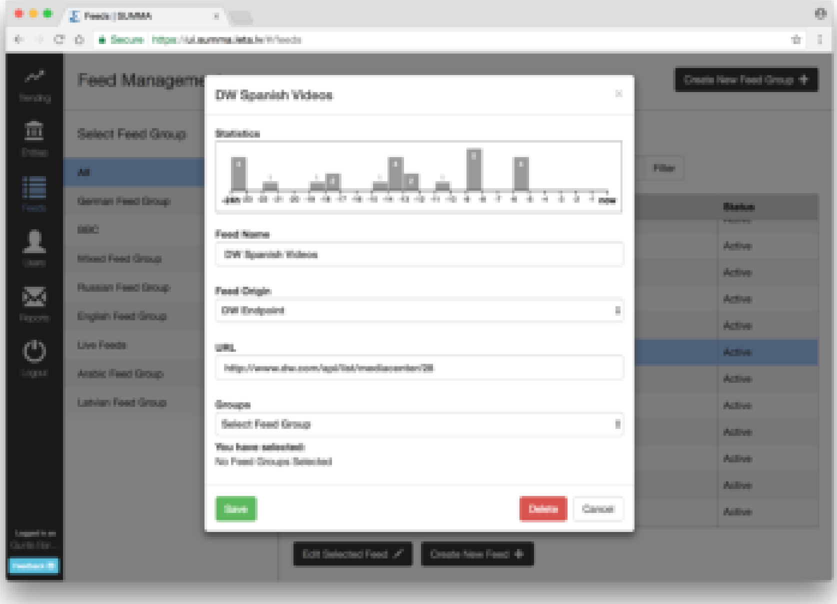


Figure 7: The hourly bars in the Trending graphs now are clickable

An additional new navigation option is the possibility to open several MediaItems simultaneously in separate browser windows - this is achieved by pressing "Shift" on the keyboard while opening a specific MediaItem from the Story view (Figure 6). MediaItems in separate browser windows can be played simultaneously resulting in a mixed audio output; the audio volume for each MediaItem can be adjusted independently. This functionality was added to mimic how media monitoring traditionally is performed at BBC Monitoring, where the monitoring journalist typically follows 4 live TV channels simultaneously.

## 2.5 Feed activity monitoring

Users with administrator privileges eventually would like to see all health-check parameters of the SUMMA Platform. As the first step towards this goal the visualisation of the activity of individual news feeds into the SUMMA Platform has been added (Figure 8).



DW Spanish Videos

Feed Name: DW Spanish Videos

Feed Origin: DW Endpoint

URL: http://www.dw.com/api/list/mediacenter/28

Groups: Select Feed Group

You have selected:
No Feed Groups Selected

Figure 8: Feed activity graph

Feed activity graph is accessed by clicking the "Edit Selected Feed" button in the All feeds list under the Feeds tab. Bars in the activity graph are clickable and thus are currently the most convenient way to watch live feed content in near-real-time (with 5 minute delay, the default live media chunking interval).

# 3 Backend improvements

The SUMMA Platform backend was slightly modified to better handle the large data volume, the multitude of languages, and the occasionally crashing NLP modules.

At the core of the scalability support of the platform is the use of Rabbit Message Queue (RabbitMQ7) for asynchronously distributing NLP jobs to the NLP modules. Until now the NLP modules were implemented as REST servers. And the connection between the Message Queue and the NLP REST APIs was handled by NLP Worker Wrappers (see the deliverable D6.1, section 4 for further details). The use of Worker Wrappers made the scaling of NLP modules cumbersome because the Worker Wrapper and the NLP REST server formed a logical pair and thus needed to be scaled together.

It turned out that Docker-Compose and Docker-Swarm do not support pairwise scaling of containers. In the previous milestone, we experimented with other container management platforms that supported pairwise scaling, but they were not stable enough (see deliverable 6.2, section 3.1). As a result, we resorted to manual pairwise scaling using Ansible deployment scripts (see deliverable 6.2, section 3.2). The use of manual scaling lead to other problems:

- it was hard to monitor which NLP modules are up and which have crashed; and
- it was cumbersome to maintain two places for container description – one in the docker-compose file and the second in Ansible scaling scripts.

To solve these problems, in SUMMA Platform Version 1.1 we decided to change the NLP module integration interface (Figure 9). Instead of using REST APIs that required Worker Wrappers to pass the jobs from the Message Queue, we now make the NLP modules connect to the Message Queue directly and send their results to the Result Queue. This change allows us to use the native docker-compose/docker-swarm scaling support. Thus scaling, e.g., an English ASR module to 5 instances is as easy as writing 'docker-compose scale asr-en=5'. Also, each NLP module can now implement internal multithreading and/or multilinguality by taking from the job queue as many jobs of the appropriate language as it can handle simultaneously, which improves the scalability even further.

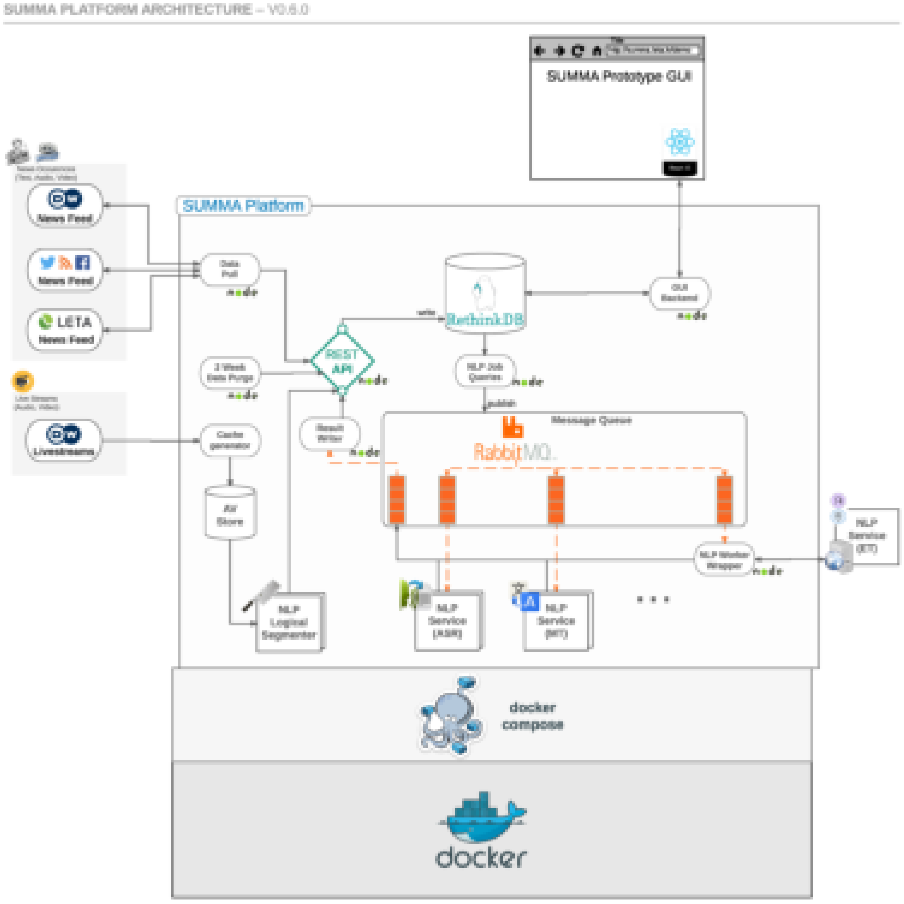


**Figure 9:** Architecture of SUMMA Platform, Version 1.1

# 4 Conclusion

The diversity of added UX functionality and backend improvements highlight the value of the intense end-user testing as the only reliable source of real-life usage scenarios in an environment with realistically large data-sets. We conclude that the overall SUMMA Platform architecture is adequate for the given data volumes and user expectations under realistic usage conditions. At the same time this clearly is an ”intermediate report”, because further improvements to both UX and backend of the SUMMA Platform will inevitably continue to be made throughout the project. For example, UX support for touch-based iPad / Notepad has been requested by DW recently and is being worked upon already.

Also the stability and performance of the individual NLP modules were assessed leading to the preliminary conclusion that Technical Readiness Level (TRL) for the ASR and MT modules is approaching the practically useful level, while TRL for the NLU modules apparently requires further improvements for them to become practically useful.

# ENDPAGE

## SUMMA

## H2020-ICT-2015 688139

D6.3 Intermediate report on platform