Scalable Understanding of Multilingual MediA (SUMMA)

http://www.summa-project.eu

H2020 Research and Innovation Action
Number: 688139
D2.2 – Initial Data Provision report

Nature	Report	**Work Package**	WP2
Due Date	31/07/2017	**Submission Date**	31/07/2017
Main authors	David Sheppey (BBC) and Lauren Boyes (BBC)		
Co-authors	Andrew Secker (BBC), Robert Cobain (BBC), Guntis Barzdins (Leta), Peggy van der Kreeft (DW), Pedro Balage (Priberam)		
Reviewers	Steve Renals (UEdin)		
Keywords	data, provision		
Version Control			
v0.1	**Status**	Draft	27/07/2017
v0.2	**Status**	Reviewed	28/07/2017
v1.0	**Status**	Final	31/07/2017

Contents

List of Figures

Abstract

The purpose of this document is to set out the technical specification and proposed solution for the data provision for the SUMMA project - specifically for work packages 2 & 6. This only includes the system that is currently under development by the consortium and not the final deployable solution.

This document will be used by the development teams for the building of the solution and will cover the following aspects of this work:-

- Details of the provided data dumps
- Details of the content provided via API
- Provision of Live AV feeds
- Provision of text-based media
- Details of customised test and training data not already covered by the above items

This document is aimed at the development teams working on this project in BBC, Deutsche Welle, LETA and the Chief Technical Architect of the overall SUMMA system.

It may also be considered to form part of a technical reference for other partners.

1 Data Dumps Provided by BBC

The data dumps were primarily created in order to provide data to the consortium. This was so that partners could see the kind of media files that they would need to process. A very important use of the data dumps was that later from these a subset of this data was used to create the SUMMA streaming test sets which is being used to test the speech recognition, segmentation and machine translation technologies. The SUMMA streaming test set is described in more detail in Deliverable 3.1.

1.1 Initial Data Dump

January 2016
The first data dump was provided by BBC Monitoring via the BBC's corporate Box account. This consisted of sample text files, Tweets and articles provided in text files as well as AV samples recorded locally by BBC Monitoring.

1.2 Data Dump 2

31st July 2016
This was delivered to the BBC's Corporate Box account at month 6 as scheduled. It consisted of both AV and text-based media.

The AV media in the data dump originated from BBC Monitoring and included the BBC's Arabic and Persian services. It consisted of 8 video streams covering the full 24 hour period of 12th July 2016, each supplied in two forms, an audio only version and an audio + video version.

The audio was presented as audio encapsulated MPEG2 .TS files in AAC-HE 64 kb/s format, chunked into 60 minute segments to allow for easy transfer and handling of the files. As well as the audio files there were also video files delivered in their original format as MPEG2 .TS files with bitrates between approximately 1mb/s and 12mb/s depending on the original source material. These were also chunked at 60 minute intervals to time-align with the audio files described above.

In the case of material originating from the BBC's Arabic and Persian services, there were also accompanying TEXT sidecar files containing some scripts of the original material wherever these were available.

The text-based data in the data dump was supplied by BBC NI and contained material derived from Twitter, Facebook and websites etc. embedded within JSON files which also contain any relevant metadata pertaining to the source.

From this point forward no textual data was provided by the BBC following review from our legal team. It was deemed unlawful for the BBC to provide text-based media to a third party. The consortium agreed that as a solution the BBC would develop a text-based scraping component for inclusion within the SUMMA platform. This would allow individual partners to gather their own text-based media without involvement from the BBC.

1.3 Data Dump 3

31st October 2016
This data dump consists of week-long recording of feeds in English, German, Russian and Arabic, each in one hour chunks, supplied to University of Edinburgh on a USB hard drive. These feeds were:-

- Arabic
  - Al Jazeera
  - BBC Arabic
  - LBC Int Lebanon
  - Oman TV
  - 2M Monde
- English
  - Al Jazeera English
  - BBC News Channel DSAT
  - CNN International
- German
  - IRIB 1 (N-TV)
  - DW Deutsch
- Russian
  - Rossiya 24

1.4 Data Dump 4

25th December 2016
This data dump consisted of week-long recording of feeds in English, Spanish, German, Russian, Persian (Farsi) and Arabic, each in one hour chunks. These feeds were:-

- Arabic
  - Al Jazeera
  - BBC Arabic
  - LBC Int Lebanon
  - Oman TV
  - 2M Monde
- English
  - Al Jazeera English

  - BBC News Channel DSAT
  - CNN International
- Spanish
  - Canal 24h
- Russian
  - Rossiya 24
- German
  - IRIB 1 (N-TV)
- Persian
  - IRINN - Iranian News Network

1.5 Data Dump 5

21st February 2017
This data dump consisted of 7 days of data from:-

- Russian
  - Life TV
  - Rossiya 24
- Spanish
  - 2M Monde
- Persian
  - IRINN - Iranian News Network

1.6 Data Dump 6

3rd March 2017
This data dump consisted of 7 days of data from:-

- Arabic
  - 2M Monde
  - Al Jazeera
  - LBC Int Lebanon
  - Oman TV
  - BBC Arabic

- English
  - Al Jazeera English
  - BBC News Channel DSAT
  - CNN International
- Spanish
  - Canal 24h
- Russian
  - Life TV
  - Rossiya 24
- German
  - IRIB 1 (N-TV)
  - DW Deutsche
- Persian
  - IRINN - Iranian News Network
- Ukrainian
  - UKRTV5

1.7 Additional Data Dumps

AV material is currently being recorded into a local store in Belfast. If additional data dumps are required these can be created on demand from the local storage (noting that material is only cached for a period of 14 days) and shared with project partners on an external hard drive.

2 Data Dumps Provided by DW

The Deutsche Welle data dumps were provided at an early stage of the project as sample test sets, so partners could familiarise themselves with and try out the specific type of content and formats provided by DW. Partners had access to these data collections for processing. These dumps included AV and text data in all eight DW SUMMA languages.

It was agreed that future data dumps on specific languages or topics could be provided upon request. This was done in the first half of the project.

Data dumps are provided via BBC's corporate Box account, which was the agreed repository for SUMMA data provision at that time.

2.1 Initial Data Dump

January 2016
The first data dump consisted of sample text and AV files in all eight DW SUMMA languages (all except Latvian) in the format used for API access. That allowed partners to process DW content at an early stage.

2.2 Data Dump 2

15 March 2016
A specific data collection on the topic of the 5th anniversary of the Syrian uprising – 24-hour coverage of the news.

2.3 Data Dump 3

12 July 2016
A specific data collection on the topic of the 10th anniversary of the Second Lebanon War – 24-hour coverage of the news.

2.4 Additional Data Dumps

July 2016 - July 2017

No more specific data dumps were requested from DW after the first six months of the project. There was no need for regular data dumps, as DW content was being continuously supplied through the DW API in all eight DW SUMMA languages.

The remaining data dumps related to customised data sets for training and testing. These are described in section 8.

3 Data Dumps and other Latvian Language Resources Supplied by LETA

LETA has supplied following Data Dumps and other Latvian language resources to the consortium:

- 5 hours of Latvian TV broadcast data along with subtitles in Latvian.
  This data dump is identified as ”LV Broadcast sample” and is supplied to SUMMA consortium members through BBC’s Corporate Box account.

- Latvian TEXT news archive produced by LETA and covering the entire year 2015 and the first 2 months of 2016. This large (160MB compressed) archive is suitable for contemporary Latvian news language modeling in ASR and MT. This data dump is identified as ”LETA_Latvian_Article_Export” and is supplied to SUMMA consortium members through BBC’s Corporate Box account.

Three corpora of the transcribed Latvian SPEECH:

- LV_test_data.zip (500 MB) - ASR test data for SUMMA (for SUMMA project use only)
- larko_20160210.tar (2.76 GB) - Latvian transcribed speech corpus, 8 hours 44 minutes. Public, can be shared without constraints. Described at http://larko.ailab.lv/index.php/info
- teleruna_20160121.tar.gz (7.85 GB) - Latvian transcribed noisy speech corpus (phone etc.), 27 hours. Created by LETA (for SUMMA project use only). This data dump is identified as ”Latvian_Speech_Corpora_(Transcribed)” and is supplied to SUMMA consortium members through BBC’s Corporate Box account.

Latvian-English parallel corpora for use in SUMMA project and in WMT-2017 competition, where Latvian was included for the first time in the ”News translation task”. This data dump is identified as ”Latvian-English_Parallel_Corpora” and is supplied to SUMMA consortium members through BBC’s Corporate Box account.

Besides that from May, 2017 LETA is supplying semi-live Latvian TV news feed into the SUMMA Platform for end-user evaluation and Latvian ASR and MT testing purposes. Technically the Latvian HLS video feed is gathered from the Latvian Public Broadcasting web portal http://www.lsm.lv/ by techniques used to gather video feeds from DW portal http://www.dw.com.

4 DW Content Provided via API

Deutsche Welle has its API access in place and has made the content of its mobile site (m.dw.com) available to the consortium, together with instructions, from the early stages of the project. It contributes material in English, German, Arabic, Spanish, Russian, Persian, Portuguese and Ukrainian.

This ensures a continuous flow of audiovisual (audio in MP3 and video items in MP4) and textual data (text articles in JSON) in eight of the SUMMA languages. Accompanying metadata is also provided in JSON format. Deutsche Welle provides two levels of JSON files for articles: a teaser format providing a list, overview of items, and a detailed format, with a full-text version containing all details of the assets.

This content provided via API is ingested into the SUMMA integrated platform, but is also downloaded separately by component owners for processing in their components or modules.

In July 2017, the API recorded the following numbers of available audio and video content assets in the SUMMA languages

- English: 14299
- German: 10428
- Spanish: 9705
- Russian: 2194
- Ukrainian: 2136
- Portuguese: 1997
- Farsi: 1656
- Arabic: 338

For online text articles, the total numbers are:

- English: 166760
- German: 256967
- Spanish: 139209
- Russian: 181122
- Ukrainian: 70151
- Portuguese: 60393
- Farsi: 24989
- Arabic: 107230

5 Live AV Feeds Provided by BBC

To facilitate testing and training, during the SUMMA development phase, the BBC has made available some live AV sources. Since the beginning of 2017 we have been streaming 11 sources onto the Internet for this purpose and these are available to any partner wishing to subscribe - subject to acceptance of the BBC's licensing terms. Details of these feeds can be supplied upon request.

These feeds are in HLS format using MP4 encoding within an MPEG2 transport stream. The bitrates are variable and are quality equivalent to the original source material.

Recorded material can also be accessed via the same mechanism, subject to a purging period, currently 14 days.

So far University of Edinburgh and Leta have subscribed to these feeds.

The following diagram illustrates the flow of AV and text-based media into the SUMMA platform. The AV is processed by LETA's main system and the text-based media by a software component supplied by the BBC. This has been integrated into the test system in Belfast and is undergoing testing. This additional component will ultimately form part of the SUMMA platform and will be made available to LETA for full integration into the platform in due course.

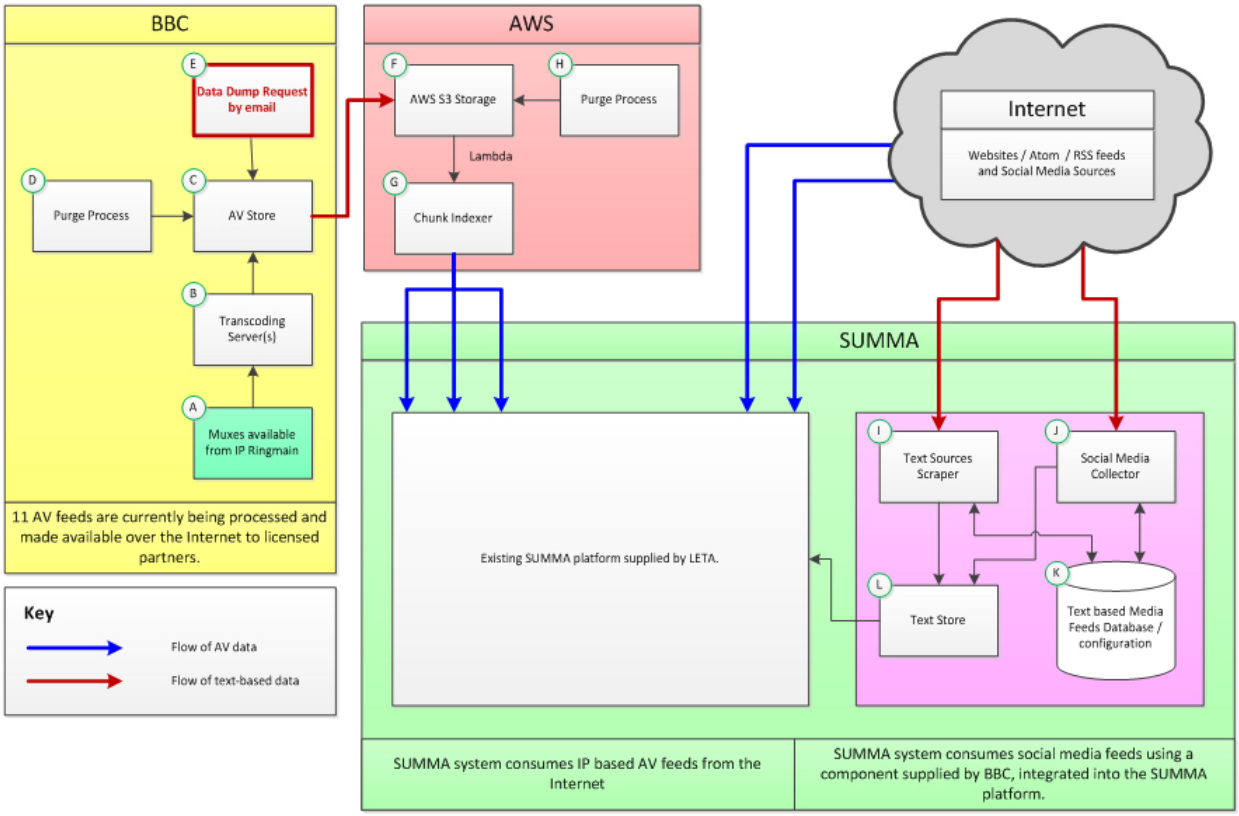


Figure 1: Diagram showing how BBC has supplied training and testing data to the SUMMA platform

The BBC's responsibility under WP2 is to supply live video feeds to the consortium partners for the duration of the project. This is achieved using a small cache in AWS to temporarily store the video segments. This cache is simply to act as a buffer for incoming HLS segments.

The storage of video for later retrieval in WP6 has been integrated into the SUMMA platform by LETA and is not included in the AWS architecture for the following reasons:-

- All SUMMA Platform registered end-users need access for review to the relevant video cache fragments through the web-based SUMMA Platform UX (User eXperience) interface. Username and password rather than IP address validate the SUMMA Platform end-users making their direct access to AWS video cache problematic - it would require accessibility from any IP address rather than only from the statically subscribed IP addresses of LETA and UEDIN.
- Video cache needs to be preserved for the duration of the project and possibly also beyond. Meanwhile the AWS based solutions caches video for only a period of 14 days. Once the project ends, AWS video cache will be withdrawn altogether.
- DW live video feeds do not have any video cache whatsoever - HLS video chunks are gone within the minutes after the live video feed has been transmitted on Internet.
- All live video content needs to be downloaded and briefly cached by SUMMA Platform anyway to pass it further to the NLP modules such as Automatic Speech Recognition, Machine Translation and others.

The SUMMA Platform internal video cache is described in detail in the deliverable ”D6.1 – Platform architecture and API Modelling tools selection”, Section 2.3 ”Intermediate layer: HLS storage”.

6 Live AV feeds Provided by DW

Even though for DW the on-demand content is more important for its primary use case, DW has also provided access to some of its streaming content.

These live streams (HLS streams) are currently part of the content processed in the platform on the LETA server. The stream is cut into chunks of manageable blocks and then transcribed and translated. This is a major achievement, as we can now get streaming content from different DW languages into a joint English-language searchable repository.

The SUMMA platform currently ingests the DW English live feed. It is foreseen that also the other live streams, i.e., German, Spanish and Arabic are included.

7 Text-Based Media

7.1 Introduction

The BBC has developed a software component for inclusion within the SUMMA platform. This component will enable each consortium partner to gather text-based media for their own use. Following a legal review it was determined that providing text-based content directly to the partners was unlawful and broke copyright restrictions. This solution not only avoids this issue, but allows the individual partners to select from their own preferred list of sources.

7.2 Sources

Facebook's terms and condition prohibit the programmatic retrieval of Facebook content, so this source is now considered out of scope. Instead we have concentrated on Twitter and RSS/Atom/RDF feeds, which include web pages referenced within.

Currently tweets and articles/pages are gathered and sent into SUMMA as they are made available. However, it is likely that this behaviour will be altered to prevent the SUMMA platform having to contend with unnecessary load. Instead, Twitter messages will be gathered over a defined period of time, during which they will be prepared as a single digest and pushed into SUMMA as one.

The monitoring of Twitter accounts is achieved via the Twitter streaming API.

This component is ready for integration with the main SUMMA Platform. During the development phase it was monitoring over 1100 RSS feeds and 137 Twitter accounts.

Deutsche Welle has supplied a list of DW Twitter feeds which can be used in the project.

	dw_travel	Tweeting the best of DW's multimedia coverage on travel! Let the journey begin! dw.com/travel
	dw_learngerman	Improve your German – and have fun! We'll deliver tons of exercises straight to your newsfeed.
	dw_futurenow	
	dw_politics	The official account for political news from the DW's political team based in Berlin. Political news, campaign stories and coverage of German politics.
German	dw_deutsch	Wir schreiben, was wichtig ist. Ihr sagt uns, was ihr denkt: Diskutiert mit uns Nachrichten aus Deutschland und der Welt.
	dw_politik	
	dw_sport	
	dw_wirtschaft	Wirtschaftsnachrichten. Berichte. Reportagen. Aus Deutschland und der Welt. Alles, was wichtig ist. Lesen, sehen, hören, mitmachen.
	dw_wissenschaft	Du findest Forschung, Medizin und Technik richtig gut? Wir auch!
	dw_kultur	
	dw_reporter	
	dw_reise	
Arabic	dw_arabic	DWإعربية تقدم من ألمانيا تغطية إعلامية محايده لأهم الأحداث في العالم العربي وأوروبا والعالم. شاركونا في الحوار وتفاعلوا معنا!
Russian	dw_russian	Подробно о самом главном. От независимого источника. Вам тоже есть, что сказать! Присоединяйтесь!
Portuguese	dw_brasil	A DW oferece um olhar independente para você formar sua opinião. Discuta conosco os fatos mais importantes do Brasil e do mundo.
Spanish	dw_espanol	Le ofrecemos noticias y trasfondo. Usted nos da su opinión. ¡Dialogue con nosotros acerca de la actualidad alemana y global!
Ukrainian	dw_ukrainian	Ми повідомляємо про те, що важливо для Німеччини та України. Дискутуйте про це разом з нами!
Persian	dw_persian	.با ما همراه شوید، در جریان مهمترین تحولات ایران و جهان قرار بگیرید و در بحث‌ها شرکت کنید

Figure 2: Some DW Twitter Feeds

7.3 Software Component

There is no GUI for the configuration of the feeds to be monitored. Instead these can be configured by importing OPML files which are dropped into a folder. OPML is the most common way to share lists of feeds, https://en.wikipedia.org/wiki/OPML. The software can also import CSV files in the format used by BBC NewsLab's Juicer system.

For testing we are using 4 OPML files, 2 Arabic, 1 Russian and 1 General, plus 2 large CSV files containing English language and Spanish news sources. During testing we have approximately 1,100 feeds defined in these files and being consumed.

The SUMMA GUI allows for RSS feeds to be added, edited and deleted individually through the interface.

Thus, the participating broadcasters may add RSS feeds and podcasting to the channels to be provided. Deutsche Welle has granted SUMMA access to all of its RSS feeds and podcasting channels.

RSS Feed Job Producer
This creates jobs that check an RSS/ Atom/ RDF feed for new articles. This is a worker that executes every 10 minutes and establishes which feeds need to be checked for updates. In the SUMMA database, every feed has a 'check frequency' which indicates how often the feed should be polled for changes. Currently the check frequency is 1 minute for all feeds but the intention is that this value would be set appropriately for each individual feed, so feeds that relatively inactive feeds are checked less often and vice versa. If it is determined that a feed is to be checked, a new job is created in the SUMMA message queue for the feed worker(s).

Feed Worker
The Feed Worker connects to the SUMMA message queue and awaits a new feed job from the task producer. Once a job is available it retreives the 'last modified tag' from the message message in the job queue and compares this to the 'last modified' tag taken from the HEAD request of the feed url.

HTTP headers have certain key/value pairs (last modified and ETag) that indicated whether the page has changed or not. Only the HEAD is read at this stage rather than the entire body to save bandwidth. Asking the HTTP server for the HEAD only will potentially help with scalabilty. In the case where the HTTP server does not have either of these key/values, the feed is assumed to have changed.

If the feed has changed a second HTTP request is made for the whole page instead of just the HEAD.

Once the page has downloaded the worker attempts to parse the XML (if it is XML!). The first step is to check the encoding, this should be UTF-8 but sometimes (with Russian mostly it seems) the encoding is something else. Other encodings are currently not supported by the worker so at this point the feed is rejected. Based on the current feed list the following unsupported encodings have been found:-

- ISO-8859-15
- Windows-1250
- Windows-1251

- Windows-1252

Next the markup is validated to check if it is HTML. Sometimes what was once an XML feed is now being redirected to an HTML page which is either an error message or a list of new locations for the feeds etc.

If the data is HTML the parser attempts to find feeds in the <head> of the HTML by looking for a <link> of types RSS and Atom. If new feeds are found they are added to the SUMMA database for the Feed Task Producer to pick up later.

If the feed is XML it is run through an open source library called 'Tidy' (http://www.html-tidy.org/). This fixes minor problems with the markup that would normally prevent the XML reader from parsing the file. The tidied XML is then checked to see whether it is of RSS, RDF or Atom type, because each of these types has a different structure. The XML is then parsed and every item/ article URL is checked against the SUMMA database to establish if it has already been added previously. If not then a new article job is created in the message queue.

Any errors are saved to the database but currently they are not displayed in the GUI.

Article worker
This worker connects to the SUMMA message queue and waits for 'article' jobs from the feed worker. When a new job arrives, the web page is downloaded and scraped using the python module 'Newspaper3k'. The scraped article text along with language and published date is then inserted into the SUMMA database.

Twitter worker
This uses a standard developer key that allows access to the Twitter steaming API. This only allows 1% of Tweets at most, however, as long as the number of tweets generated by all the feeds combined is less than 1% of Twitter, this should not present a realistic limitation. The Twitter RESTful API has much less generous rate limits and so is not used at the moment other than to query for the ID of usernames entered into the GUI by end users.

Currently users are not filtered by whether they are 'public' (verified) persons or not. Also, this worker is currently not complete as it loads the Twitter usernames to follow from a static list, i.e. there is no bulk import tool as yet. Ultimately it will query the SUMMA database for a list of all of the Twitter feeds to follow.

The Twitter feeds and users are loaded and a request is made to the streaming API to start a stream. When a new tweet is received it is immediately added to the SUMMA database.

Regarding scalability, one Twitter worker is enough to handle 1% of Twitter but there is a potential bottleneck regarding the SUMMA pipeline. To combat this LETA have suggested collating a number of Tweets into a digest which might be processed as one larger document. This has not yet been agreed or implemented but will likely be the way forward.

100% of Twitter could amount to 100s megabits per second and would probably need 10s of workers, although the bottleneck would mainly be network bandwidth at this point, the SUMMA database might present another bottleneck if this was to be the case.

Metadata
Only basic metadata is recorded for text-based sources at present. The following data is passed into SUMMA.

Twitter:

- Tweet Id
- Tweet CreatedAt
- Tweet Text
- Tweet FullText
- Tweet CreatedBy ScreenName
- Tweet Language

RSS:

- Article URL
- Article Title
- Article Scraped Text
- Feed Title
- Article Language
- Article Publish Date

7.4 Social Media and Sentiment Analysis

In order to test storyline sentiment analysis and summarization, a new dataset was built in the context of the SUMMA project. This dataset consists of storyline highlights and sentiment analysis in tweets.

The creation of this dataset had the participation of Priberam and UEDIN, in colaboration with BBC and DW. Priberam provided a collection of semi-automatically clustered news articles and tweets as well as a web system for the annotation process. UEDIN hired students to perform the annotation and helped with the annotation management.

The annotation task was performed in three steps: database creation; highlights annotation; and tweets sentiment annotation.

For clustering the news, news articles from BBC were collected and automatically clustered into storylines using the existent news clusters SUMMA system from component T3.4, WP3. After this, the clusters were validated by Priberam, BBC, and DW.

For each storyline, the annotators annotated highlights and entities. To that end, it was defined that each highlight should represent one and only one main aspect or event addressed by the storyline and that each highlight should be self-contained, i.e., it must be fully comprehensible despite of the other highlights. The highlights summary shall be composed from 3 to up 6 highlights. The entities might be interpreted as any important real-world object, such as persons, locations, organisations, products, etc.

The annotation of tweets for sentiment was performed using all the highlights and entities previously annotated. For each target (highlights and entities), the annotator might mark if the sentiment of the tweet towards it is 'not related', 'positive', 'neutral' or 'negative'.

The final dataset will provide a testing set for the task covered by WP5, in particular the tasks T5.2 and T5.3, which provide the summarization and sentiment analysis components. Such dataset is the first of its kind and it will guide research and development in the context of SUMMA as well as foster future academic research.

8 Customised Test and Training Data Provided by DW

DW has collected and provided data to train and test the specific modules and technologies based on the requirements from the technical partners. As explained in the Data Management Plan D2.1, the provision will happen gradually and the collection will be built up, depending on the availability of the sources, the feasibility of processing and preparing data. Scripts have been written to automate this process as much as possible, converting the original transcript into a machine-readable format. Nevertheless, the internal editorial formats need to be prepared and checked to make them suitable for wider dissemination, which makes the effort quite resource-intensive.

Deutsche Welle has provided items with teasers which can be used to train the summarisation tool.

Deutsche Welle has also supplied datasets for ASR and MT training and testing specifically. These come from internal DW datasets with editorial manuscripts, which can be used to train transcription tools. Multilingual datasets in well-resourced languages such as English, German, Spanish and to a certain extent Arabic, are collected for evaluating machine transcribed and translated output.

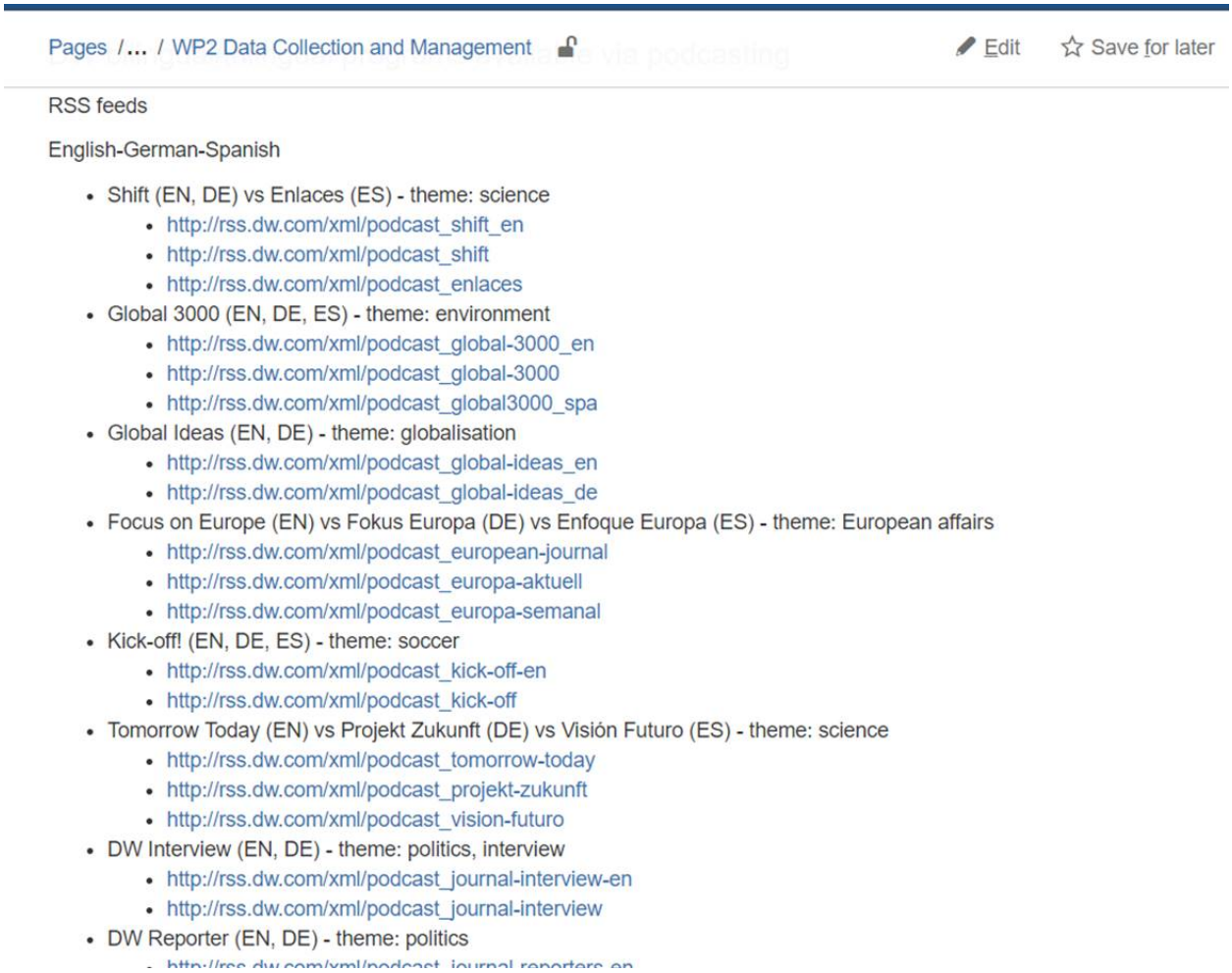


Figure 3: Some of DW's Multilingual Sources

In this first half of the project, the Deutsche Welle team has worked closely with the ASR teams from the University of Edinburgh and IDIAP to set the requirements and procedure for supplying ASR training and test sets for different languages. A prioritisation and time schedule was estab-

lished. In the first year, the focus was on supplying German and Spanish video and audio content with corresponding transcripts/manuscripts.

At the end of Y1/beginning of Y2, DW provided over 10 hours of AV material with manuscripts for both German and Spanish, plus approximately 60 hours of original German audio content for ASR training. German data consisted mostly of DW audio programmes on spoken news. A particular programme used for this is DW's Langsam Gesprochene Nachrichten, which provides audio files of news in German at both regular speed and slowly spoken for language learning. The regular speed audio version is used for SUMMA ASR training purposes.

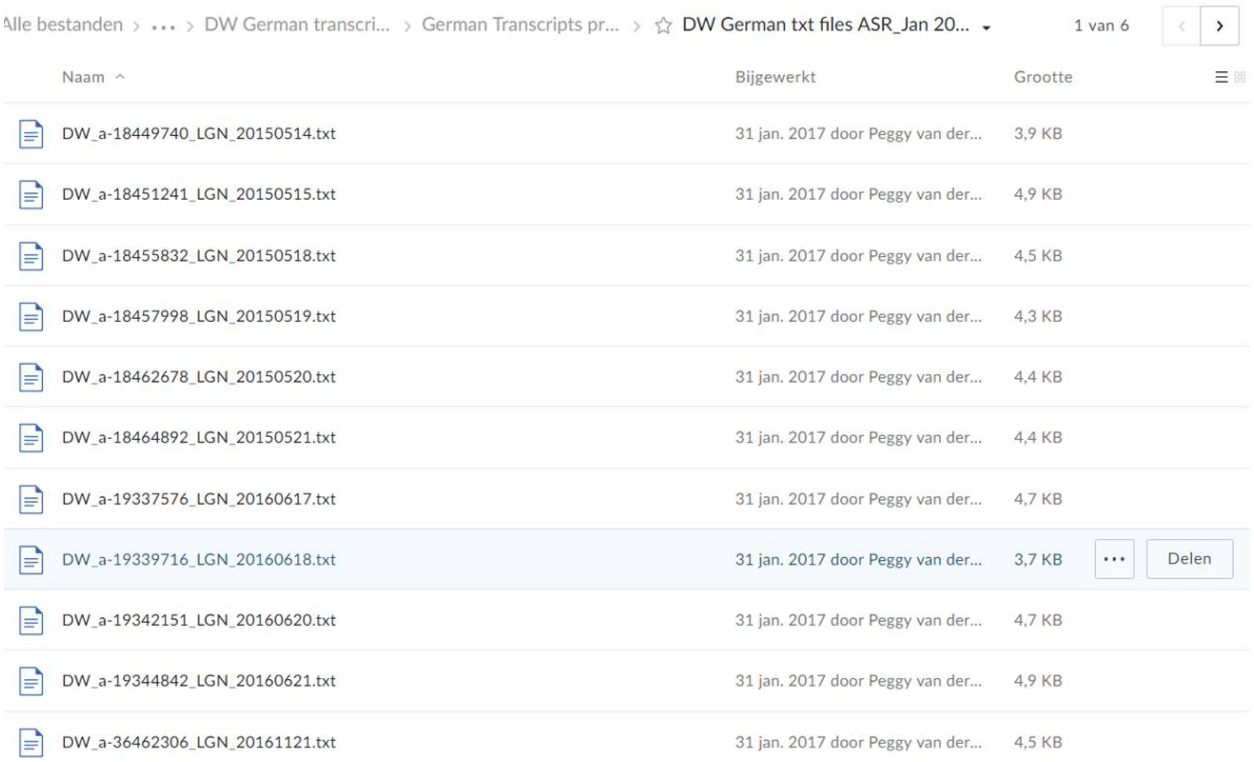


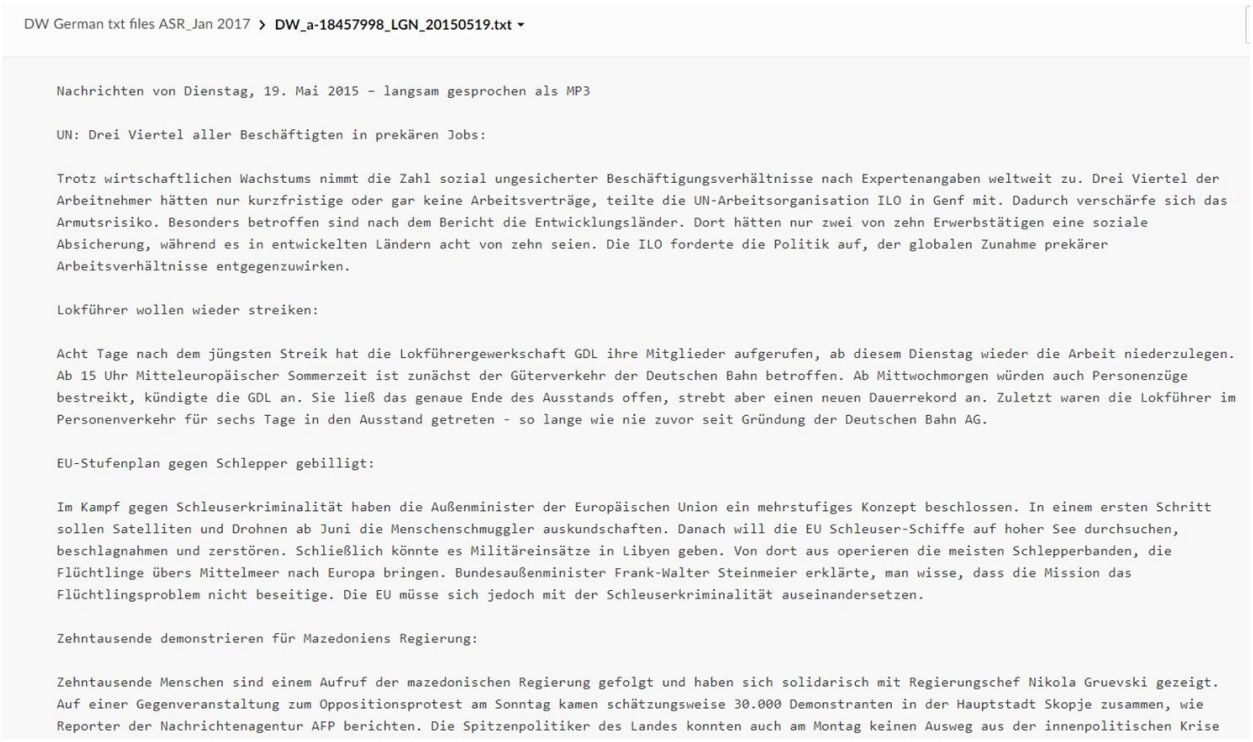
Figure 4: Provision of German transcripts

Russian, Ukrainian and Farsi manuscripts will be collected for training and testing data in the second half of the project.

English and Arabic manuscripts were available to the technical partners from other sources.

Improvement of the transcription output is in the interest of ASR developers as well as content partners to arrive at better results. Thus, customised training data is important.

Multilingual source material will be primarily collected for evaluation and testing purposes.



DW German txt files ASR_Jan 2017 › **DW_a-18457998_LGN_20150519.txt**

Nachrichten von Dienstag, 19. Mai 2015 – langsam gesprochen als MP3

UN: Drei Viertel aller Beschäftigten in prekären Jobs:

Trotz wirtschaftlichen Wachstums nimmt die Zahl sozial ungesicherter Beschäftigungsverhältnisse nach Expertenangaben weltweit zu. Drei Viertel der Arbeitnehmer hätten nur kurzfristige oder gar keine Arbeitsverträge, teilte die UN-Arbeitsorganisation ILO in Genf mit. Dadurch verschärfe sich das Armutsrisiko. Besonders betroffen sind nach dem Bericht die Entwicklungsländer. Dort hätten nur zwei von zehn Erwerbstätigen eine soziale Absicherung, während es in entwickelten Ländern acht von zehn seien. Die ILO forderte die Politik auf, der globalen Zunahme prekärer Arbeitsverhältnisse entgegenzuwirken.

Lokführer wollen wieder streiken:

Acht Tage nach dem jüngsten Streik hat die Lokführergewerkschaft GDL ihre Mitglieder aufgerufen, ab diesem Dienstag wieder die Arbeit niederzulegen. Ab 15 Uhr Mitteleuropäischer Sommerzeit ist zunächst der Güterverkehr der Deutschen Bahn betroffen. Ab Mittwochmorgen würden auch Personenzüge bestreikt, kündigte die GDL an. Sie ließ das genaue Ende des Ausstands offen, strebt aber einen neuen Dauerrekord an. Zuletzt waren die Lokführer im Personenverkehr für sechs Tage in den Ausstand getreten - so lange wie nie zuvor seit Gründung der Deutschen Bahn AG.

EU-Stufenplan gegen Schlepper gebilligt:

Im Kampf gegen Schleuserkriminalität haben die Außenminister der Europäischen Union ein mehrstufiges Konzept beschlossen. In einem ersten Schritt sollen Satelliten und Drohnen ab Juni die Menschenschmuggler auskundschaften. Danach will die EU Schleuser-Schiffe auf hoher See durchsuchen, beschlagnahmen und zerstören. Schließlich könnte es Militäreinsätze in Libyen geben. Von dort aus operieren die meisten Schlepperbanden, die Flüchtlinge übers Mittelmeer nach Europa bringen. Bundesaußenminister Frank-Walter Steinmeier erklärte, man wisse, dass die Mission das Flüchtlingsproblem nicht beseitige. Die EU müsse sich jedoch mit der Schleuserkriminalität auseinandersetzen.

Zehntausende demonstrieren für Mazedoniens Regierung:

Zehntausende Menschen sind einem Aufruf der mazedonischen Regierung gefolgt und haben sich solidarisch mit Regierungschef Nikola Gruevski gezeigt. Auf einer Gegenveranstaltung zum Oppositionsprotest am Sonntag kamen schätzungsweise 30.000 Demonstranten in der Hauptstadt Skopje zusammen, wie Reporter der Nachrichtenagentur AFP berichten. Die Spitzenpolitiker des Landes konnten auch am Montag keinen Ausweg aus der innenpolitischen Krise

Figure 5: Example of German transcript

9 Conclusion

We expect the data and the methods by which they are fed into the SUMMA system to change as incremental improvements are applied to the platform. A further and final data provision report (D2.4) will document these changes at the end of project month 36.

ENDPAGE

SUMMA

H2020-ICT-2015 688139

D2.2 Initial Data Provision report