

D2.3 – Data Collection Report Y2

Document Number	D2.3
Document Title	Data Collection Report Y2
Version	2.0
Status	Final
Work Package	WP2
Deliverable Type	Report
Contractual Date of Delivery	31.10.2015
Actual Date of Delivery	31.10.2015
Responsible Unit	Websays
Keyword List	Data Collection
Dissemination level	PU



Editor

Marc Poch (Websays SL, Websays)

Contributors

Vincenzo Lanzolla (Teleperformance Italy, TP)

Letizia Molinari (Teleperformance Italy, TP)

Hugo Zaragoza (Websays SL, Websays)

SENSEI Coordinator

Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science

University of Trento, Italy

giuseppe.riccardi@unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	15/07/2015	Draft	Marc Poch (Websays)	Table of Content
0.2	14/08/2015	Draft	Hugo Zaragoza (Websays)	Section "Web Data collection" added
0.3	21/08/2015	Draft	Marc Poch (Websays)	Section "Overview"
0.4	21/08/2015	Draft	Vincenzo Lanzolla(TP)	Sections "Data Access", "Collection of Call-Center", "Appendix A,B,C" added
0.5	15/08/2015	Draft	Marc Poch (Websays)	Minor corrections
0.6	15/09/2015	Draft	Marc Poch (Websays)	Executive Summary
0.7	22/09/2015	Draft	Hugo Zaragoza (Websays)	Approach section and revision
0.8	28/09/2015	Draft	Letizia Molinari(TP)	Section Annotations & Statistics added
0.9	30/09/2015	Draft	Marc Poch (Websays)	Statistics update
0.10	1/10/2015	Draft	Vincenzo Lanzolla(TP)	Contributions to "Overview" and "Conclusion"
1.0	04/10/2015	Draft	Marc Poch (Websays)	Consolidated version ready for scientific review and quality check
1.1	12/10/2015	Draft	Elisa Chiarani (UNITN)	Quality check completed. Minor remarks
1.2	12/10/2015	Draft	Fabio Celli (UNITN)	Review completed. Minor remarks
1.3	16/10/2015	Draft	Marc Poch (Websays)	Improvements after Quality check
2.0	27/10/2015	Final	Giuseppe Riccardi (UNITN)	Final Review



Executive summary

Deliverable 2.3 describes the data collected during Period 2 of the project, the annotation efforts and developed tools.

The document presents the call center data collection, together with its annotation and indexation tasks.

The deliverable describes the web data collection and the work carried out for data extraction, pre-processing and data indexing.

Finally document reports information about data publication and sharing beyond the consortium, and the methods to obtain copy-righted free materials.

Table of contents

EXECUTIVE SUMMARY	4
LIST OF ACRONYMS AND NAMES	7
1. OVERVIEW	8
1.1. FOLLOW-UP TO PERIOD 1 ACTIVITIES	8
1.2. APPROACH	9
1.3. DATA ACCESS	10
1.3.1. Public Data Access	10
1.3.2. Partner's Data Access	10
2. PERIOD 2 DATA COLLECTION	12
2.1. COLLECTION OF CALL-CENTRE DATA	12
2.1.1. Previous work	12
2.1.2. Elastic Search - Kibana	12
2.1.3. SENSEI ACOF tool improvements	12
2.1.4. Annotations	13
2.1.5. Statistics	15
2.2. WEB DATA COLLECTION	16
2.2.1. Previous work	16
2.2.2. Data sources	16
2.2.3. Topics and Use Cases	17
2.2.4. Content Extraction	18
2.2.5. Pre-processing	19
2.2.6. Data Statistics	20
2.2.6.1. General News and Newspaper Social Media Publications	20
2.2.6.2. RATP	22
2.2.7. Data storage and navigation improvements	25
2.2.7.1. Time based index partitioning (sharding)	25
2.2.7.2. Sorting data by Conversation Size	25
3. CONCLUSIONS	28
REFERENCES	29
APPENDIX A: ELASTIC SEARCH	31
Elastic Search JSON Element Description	31
Elastic Search JSON Format	32
ES report Comparison	37
APPENDIX B: SENSEI ACOF TOOL V2.0	40
My SQL Data Model	40
New View: Monitoring prefilled	40



<i>New View: Report Elastic Search</i>	41
<i>Data Export to Excel</i>	43
APPENDIX C: STATE OF DATA COLLECTIONS	44

List of Acronyms and names

Acronym	Meaning
ACOF	Agent Conversation Observation Form
API	Application Program Interface
CRF	Conditional Random Field
QA	Quality Assurance
RATP	Régie Autonome des Transports Parisiens
REST	Representational State Transfer
TRS	TRanScription

1. Overview

1.1. Follow-up to Period 1 activities

During Period 1 (P1) of the project Speech and Social media data have been collected and annotated. Regarding Speech data, DECODA, LUNA and call center data were collected and annotated. During Period 2 (P2) of the project such collections have been indexed in Elastic Search and can be queried using Kibana. The SENSEI ACOF tool has been updated and improved to its version 2. Social Media data has been crawled constantly during the second year of the project and new case studies have been developed. Parsers have been improved and better indexing techniques have been used. All these updates will be presented in detail in this deliverable.

The D2.3 data collection is composed of data files of speech and social media, pre-processed and annotated. Following the D2.2 approach, the D2.3 data collection is composed of several sub-collections:

- Speech:
 - DECODA
 - LUNA
- Social Media v2
 - General News Topics
 - Newspaper Publications
 - RATP
 - Orange

Some tasks necessary to achieve this deliverable are briefly described here:

- Split Social Media data collection index for performance and data growth handling
- Topics validation and definition
- Adaptation of the Websays parsers to the required sources
- Update and bug fixing of crawling and parsing tools
- Evaluation and validation of the obtained data
- Import of machine annotated data and human annotated data in Elastic Search

D2.3 data collection is a continuation of the D2.2 collection.

The main work carried out since D2.3 is listed here:

- Speech:
 - Internal (TP) and external (UNITN) Calibration
 - Review of TP annotation work on the LUNA and DECODA corpus
 - Definition of the new monitoring form with pre-filled synopsis and annotations
 - Upgrade of the sensei ACOF tool and implementation of Elastic Search/Kibana

- Annotation Agent oriented summarizes
- Annotation of conversation caller/requester-oriented (synopsis) on SENSEI Web Annotation tool
- Analysis of machine generated annotations
- Social Data
 - New parser added when needed
 - Updated and corrected already existing parsers
 - Development of new dashboard topics for new study cases
 - Manual evaluation to detect errors
 - Data corrections when needed
 - Data index division to be able to handle growth and better performance
 - New sorting system for better accessing large conversations in the collection

1.2. Approach

One of the goals of WP2 is to provide a unified view of “conversations” for both speech and text dialogues. The xml representation used in the SENSEI repository is the result of a complex task of abstraction to find common mappings between such different scenarios.

The designed data schema represents tokens following the next description:

TOKEN:

- Features
 - category: pos tag
 - kind: word
 - length: length of word in characters
 - root: lemma of word
 - string: text of word
 - turn_id: identifier of turn
 - word_id: word identifier in current turn
 - disfluency: disfluency marker
 - named_entity: named entity label
 - dep_label: dependency label
 - dep_gov: word id of governor in turn
 - id_text: marker for synchronization with TRS
 - morpho: morphological features
 - speaker: speaker id from TRS
 - start_time: start time from beginning of conversation
 - end_time: end time from beginning of conversation

- eos: whether or not this word ends a sentence type
- type: Token
- id: unique hash of all features of word
- start: character offset
- end: character offset

1.3. Data Access

1.3.1. *Public Data Access*

The data publication and sharing beyond the consortium has been prepared following the survey and conclusions extracted from deliverable D8.4 Second Ethical Issues Report.

The initial data set contains three parts and over 1M items. A small sample of all the collections are provided for public online access from the SENSEI web site, together with this document, which provides an overview of the data and instructions about how to request the entire data sets. The method of data acquisition and usage is discussed in D8.2 Ethical Issues Plan. Here we provide a summary, mainly repeating the same information, recalling the most relevant information fully contained in deliverable D8.2.

For the Social Media collection, the website provides a data bundle for D2.1: a small sample of 1000 social media items from the Social Media collection, together with the entire list of public URLs of the data crawled for this collection. The entire collection (as well as individual parts of the collection) can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it.

For LUNA data we provide a small complete sample; the entire collection is distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).

For DECODA data we provide a small complete sample. The entire collection is distributed by SLDR/Ortolang (<http://crdo.up.univ-aix.fr>, ID: <http://sldr.org/sldr000847>). Researchers or practitioners may get access to the annotated corpus of human conversations free of charge by accepting the SLDR/ORTOLANG license.

For the Teleperformance data (limited to the annotations produced by QA Supervisors during the filling of AOFs), is available to the partners internally since D2.1 and D2.2 constitute the first public installment of the data. Similar to the social media data, the Teleperformance data can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it.

1.3.2. *Partner's Data Access*

For partners, a SVN data repository has been setup on one of the SENSEI servers containing all the data for easy access. In the case of the LUNA collection, the data will be distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).



The Websays Dashboard has also been made available to all partners in order to provide a rich visual interface to browse the Social Media portion of the data.

All partner have web access, upon authentication, to the SENSEI ACOF Annotation tool developed by Teleperformance, where they can find LUNA and DECODA selected conversations with integrated the relatives machine annotations and human annotations.



2. Period 2 data collection

2.1. Collection of Call-Centre Data

2.1.1. *Previous work*

Deliverable D2.1 described the LUNA and DECODA collections as well as the data model and specifications of the data to be acquired for the Call Center Quality Assurance process.

Deliverable D2.2 described the selected set of data collected during the first year of the project, the call center annotation efforts and the developed tools to annotate the conversations in Italian and French language.

P1 annotation dataset contains human annotations for more than 300 different conversations.

2.1.2. *Elastic Search - Kibana*

In P2, an instance of Elastic Search and Kibana platform have been installed on the sensei servers to collect data from multiple sources and have a near real-time search system to run complex query and analyze data.

All Teleperformance annotated data of P1 and P2 have been converted to the JSON format presented in Appendix A and loaded and indexed in Elastic Search.

The synopsis provided by AMU in P2 and the answers to questions provided by UNITN, have been loaded in Elastic Search on the same index of TP annotated data in order to increase the performance of the complex query that involve data from multiple sources.

For reporting and analysis purpose, Elastic Search offers many ways and methods such as Marvel plugin, Kibana platform or calls to Elastic Search's REST API, the choice depends on the skill of the user.

As it is not easy to create query on Elastic Search, it has been developed a new report in ACOF tool v2 that allows users, with minimal IT skills, to create complex query and analyze the result in a customizable tabular output.

Data indexed in Elastic Search can be extracted and analyzed making calls to Elastic Search's REST API or using the Kibana platform, for that purpose in Appendix A are compared three different way to interact with ES and Kibana, based on the experience and the skills of the users.

2.1.3. *SENSEI ACOF tool improvements*

A full description of the tool and its main feature are described in Section 2.3 of D2.2, here we present the major changes applied in P2 and released in version 2.0.

The first version of the tool was based on a relational database MY SQL and worked only with data annotated by TP Quality Assurance.

After the installation of Elastic Search, the tool has been upgraded in order to call Elastic Search's REST API, anyway the MY SQL Database is not ceased but is maintained as a backup of data: when a new Monitoring form is created, the tool save the data on both Elastic Search and My SQL database.



This integration of annotated data coming from multiple partners and the development of new view and reports that show these multiple data, let the SENSEI ACOF tool to well support the Evaluation Tasks and scenarios, fully defined in D1.3.

The JSON format of annotated data is reported in Appendix A as well as the new view and reports of the latter version of SENSEI ACOF are reported in Appendix B.

2.1.4. Annotations

The Annotations Activities during the P2 follow four macro areas:

- Evaluation of the reliability of the speech annotated corpus (test/retest and inter-annotator agreement);
- Complete annotation of the data sets and study binding of language to metadata annotations (Italian & French);
- Refinement of the evaluation metrics, and identification of baselines;
- Prepare spoken conversation collection for prototype evaluation.

To continue go on through the project it has been necessary to modify some tools and ACOF Guide Line. Main review has been done on the Synopses, defined after meeting, focus group and call conference. New Synopses guide line has been provided by and the way we need to review the Synopses criteria is because Synopses of call-center conversations are collected in order to evaluate the quality of systems that generate such summaries automatically. The evaluation is performed by comparing the content of system-generated synopses with a set of human-written synopses. Since every person will write different synopses when asked to summarize a conversation, guidelines should limit the variability of human-written summaries, and therefore limit the number of handwritten synopses required for modeling a “good” summary of a conversation. The new target length of a synopsis is 7% of the number of words of the conversation.

ACOFs are grids that the QA professionals use during call center conversation assessment. For SENSEI we have defined a specific ACOF.

The ACOF items are scored, the following indicators can be reported:

- Overall score of the call quality;
- Overall score of the call quality for each main area of interest;
- Overall score for each single behavior within each area of interest;
- Identification and extraction of problematic calls, i.e. the ones scored below a given threshold;
- Identification and extraction of calls managed by a given agent.

The ACOF was modified according on the need and in order to optimize selection of speech segments that can support the QA professional evaluation.

In order to prepare spoken conversation collection for prototype evaluation was provided a list of what is considered “positive word” that can influence the judgments about Pass/Fail/NA of ACOF items. Was provided details about AOF items with respect their being evaluated at the turn level/conversation level, and motivated on the basis of lexical choices of the Agent and/or

QA professional perception of intonation. Based on the updating Guide Line existing Synopses annotation has been updated. Everything based LUNA and DECODA conversational file audio.

In P2 Quality Assurance annotator has been involved in the Evaluation Phases. The Evaluation scenario has been defined follow different analysis:

- Planning phases, lead during a meeting between Teleperformance and UNITN representative (November 2014);
- Definition of the details on ACOFs human annotation in view of automatic generation;
- Planning of Corpus Annotation including ACOFs and Synopses;
- Evaluation scenario, where the main objective was the refinement of the evaluation methodology based on evaluation results, refined metrics, and user acceptance criteria.

Task Evaluation Model (speech), is based on:

- Intrinsic evaluation:
 - **Synopses:** comparison of automatic generated synopses with gold standard assessing readability, fluency, and content (Pyramid, ROUGE)
 - **ACOF:** assessed under different profiles:
 - classification task;
 - annotation task, including inter-annotator agreement.
- Extrinsic evaluation:
 - Conversation retrieval with a complex user need;
 - Comparison between two conditions: Baseline and SENSEI-enabled ACOF filling.
- Insight-oriented evaluation.

Annotation and evaluation workflow

The process of annotation and evaluation moves through 3 steps:

1. Gold standard creation
2. Intrinsic evaluation
3. Extrinsic evaluation

The **gold standard** creation step consists in the annotation of existing dialogs by expert annotators, to obtain a complete and well annotated corpus to be compared with the results of automatic algorithms and other manual annotation. Teleperformance has developed the prototype to create this gold standard. The prototype gives annotators the possibility to listen the dialog, read the transcription, answer precise questions, mark turns as evidence of their answers, create two type of synopsis (brief and extended) and add free text comments and notes. Follow statistics in the dashboard:

- annotated dialogs grouped by service (LUNA or DECODA);
- annotated dialogs grouped by service and annotator;
- annotation per file;

- annotated scores per file.

The **Intrinsic Evaluation** consists in the automatic comparison of the algorithms results with the gold standard of the different services. The extrinsic evaluation consists in measuring the increment of performances (as # of task/time, or time per task) using the SENSEI additional information during the annotation task in comparison with the traditional annotation methods.

Extrinsic Evaluation Baseline, the baseline annotation methods permits annotators to annotate dialogs using the usual methodologies and instruments.

The details of the annotation activities are fully described in D1.3.

2.1.5. Statistics

The work of the remarks has had its continuation with subsequent monitoring with different purposes. From 09/12/2014 to 30/06/2015, Teleperformance Quality Assurance professionals produced 1.856 Agent Observation Form, of which 1.032 for LUNA (audio recordings in Italian language) and 824 for DECODA (audio recordings in French language).

In P2 for LUNA (audio recordings in Italian language) 1032 Agent Observation Form have been annotated for 359 distinct dialogs.

Each conversation has been listened and evaluated from different evaluators. The average qualitative score annotated is 59,4%.

Table 1: LUNA Annotators Scores

Annotator	Service	Score Average	Num
Annotator1	LUNA	63,93	188
Annotator2	LUNA	48,51	160
Annotator5	LUNA	68,13	102
Annotator6	LUNA	56,69	235
Annotator3	LUNA	68,29	101
Annotator7	LUNA	58,65	246
TOTAL		59,4	1032

In P2 for DECODA (audio recordings in French language) 824 Agent Observation Form have been annotated for 308 distinct dialogs.

Each conversation has been listened and evaluated from different evaluators. The average qualitative score annotated is 84,45%.

Table 2: DECODA Annotators Scores

Annotator	Service	Score Average	Num
Annotator1	DECODA	82,52	246
Annotator2	DECODA	70,77	142

Annotator3	DECODA	88,71	274
Annotator4	DECODA	92,25	162
TOTAL		84,45	824

2.2. Web Data Collection

2.2.1. Previous work

Deliverable D2.1 presented the definition of a reach data schema for the collection of data and metadata from social media. Many different social media sources were taken into account (blogs, Twitter, Facebook, Youtube, etc.) and newspaper forums were targeted as the main source of data because of their complex dialogue structure.

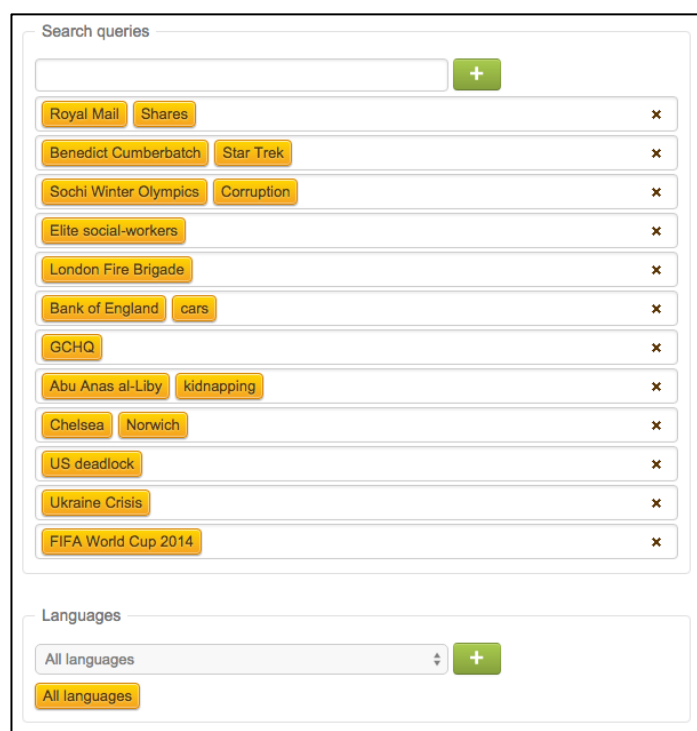
In D2.2 we presented the first year data collection with all the efforts done in data crawling, manual data curation by Websays analysts, bug fixing, parsers improvements, etc. Meanwhile partners started using the data and reported feedback to fix inconsistencies, make improvements, crawl other data, etc.

Period 1 data collection contained over 4 million posts and over a 1.5 million conversations.

2.2.2. Data sources

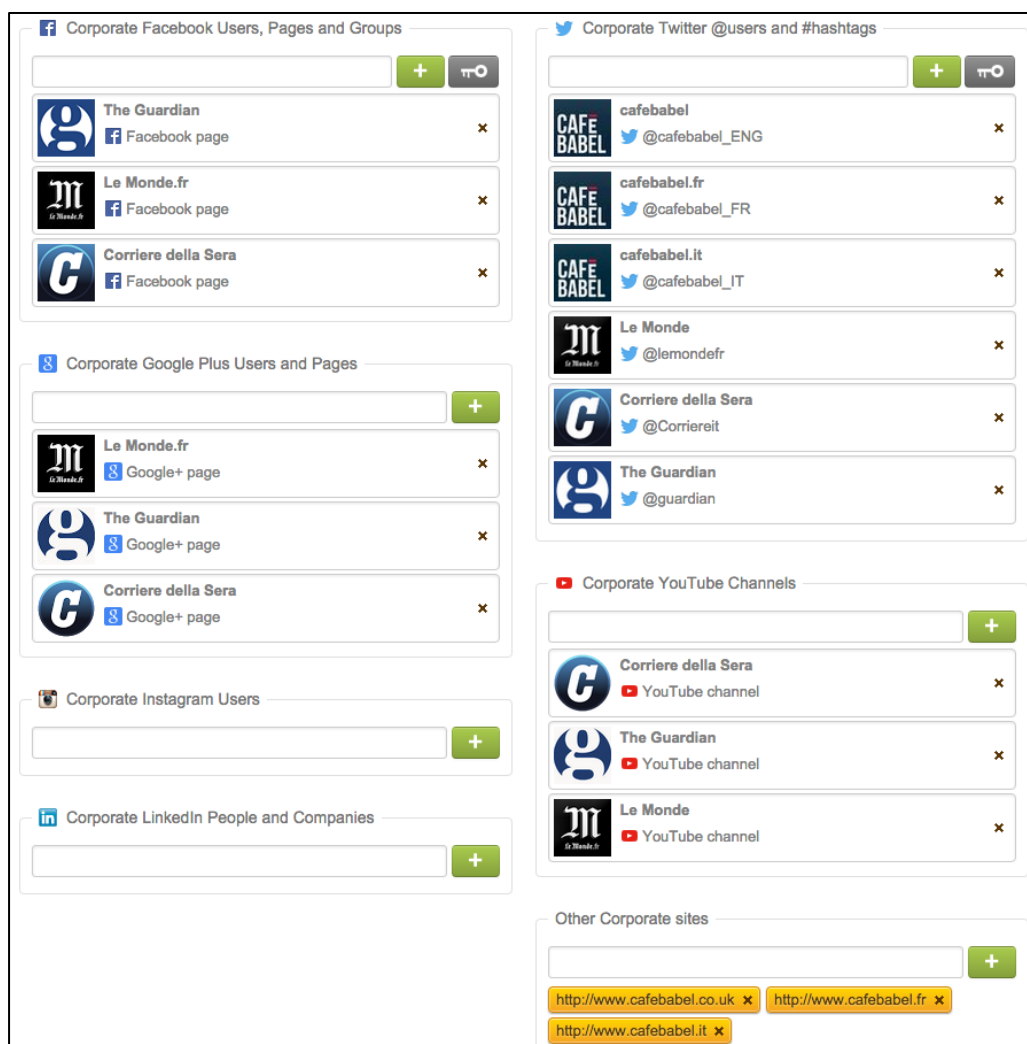
As showed in deliverable D2.2 the system is configured to crawl by terms and from specific domains. A list of data sources for which special parsers have been developed can be found on Table 6 of D2.2.

For illustration, term queries for a SENSEI's crawl profile can be seen in Figure 1, on the other hand, channels being crawled for specific newspapers can be found on Figure 2.



The screenshot shows the SENSEI search queries interface. It features a search bar at the top with a green '+' button. Below the search bar, there is a list of search queries, each with a green '+' button and a red 'x' button. The queries are: Royal Mail, Shares, Benedict Cumberbatch, Star Trek, Sochi Winter Olympics, Corruption, Elite social-workers, London Fire Brigade, Bank of England, cars, GCHQ, Abu Anas al-Liby, kidnapping, Chelsea, Norwich, US deadlock, Ukraine Crisis, and FIFA World Cup 2014. At the bottom, there is a section for Languages, with a dropdown menu set to 'All languages' and a green '+' button.

Figure 1: term queries



The interface displays several sections for monitoring corporate channels:

- Corporate Facebook Users, Pages and Groups:** Lists The Guardian, Le Monde.fr, and Corriere della Sera.
- Corporate Google Plus Users and Pages:** Lists Le Monde.fr, The Guardian, and Corriere della Sera.
- Corporate Instagram Users:** Empty search bar.
- Corporate LinkedIn People and Companies:** Empty search bar.
- Corporate Twitter @users and #hashtags:** Lists @cafebabel_ENG, @cafebabel_FR, @cafebabel_IT, @lemondefr, @Corriereit, and @guardian.
- Corporate YouTube Channels:** Lists Corriere della Sera, The Guardian, and Le Monde.
- Other Corporate sites:** Lists URLs for cafebabel.co.uk, cafebabel.fr, and cafebabel.it.

Figure 2: corporate channels monitored

2.2.3. Topics and Use Cases

Following the same approach presented in D2.2, for Period 2 data some topics have been defined as to get more relevant and accurate data depending on the project needs.

Some topics presented in D2.2 are:

- RATP (Paris public transportation system)
- Orange (Telephone company)

Some other topics have been created for Period 2:

- “Charlie Hebdo”: about the terrorist attacks of 7th of January 2015. There are 95k posts about this topic on the data collection. Next figure (Figure 3) shows the volume during the days of the attack.

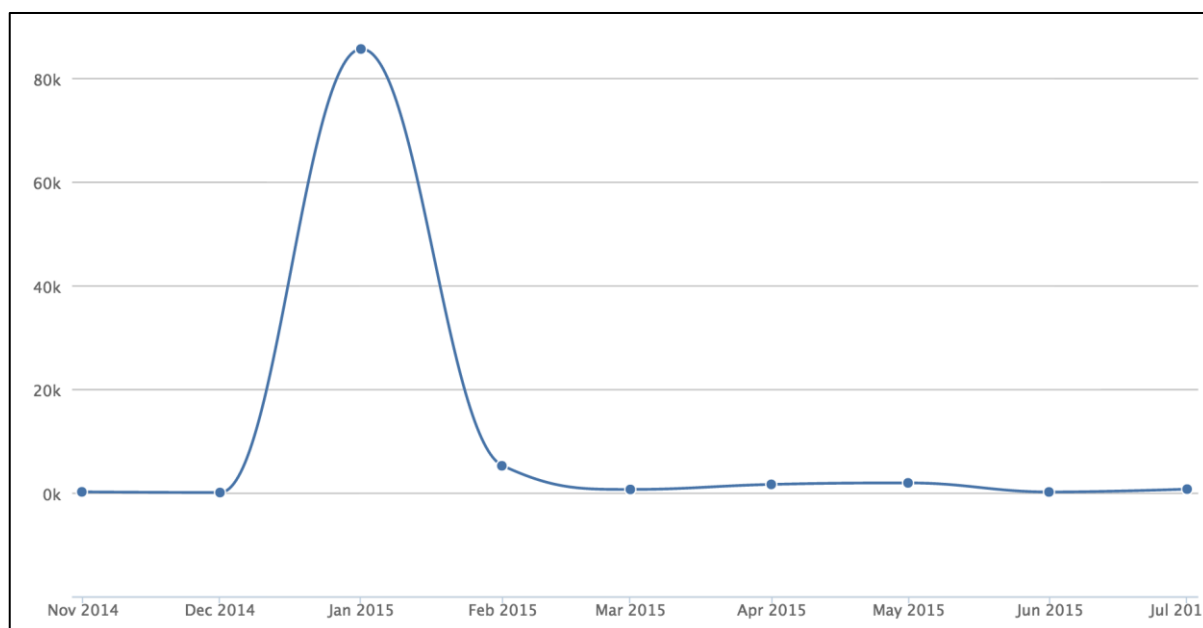


Figure 3: Charlie Hebdo topic volume

- “Germanwings plane crash”: Germanwings plane crash of 24th of March 2015. There are 6k posts about this topic on the data collection. Figure 4 shows how the data sources being monitored in SENSEI had no data at all about “Germanwings” company but suddenly there is a huge growth in volume during the plane crash dates and after.

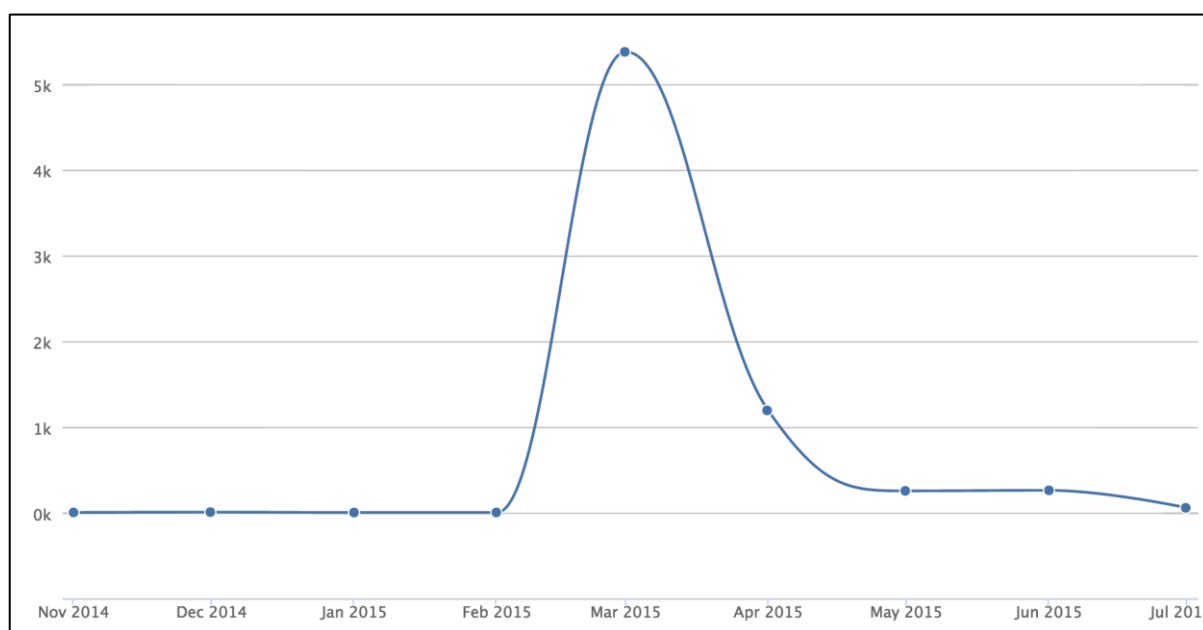


Figure 4: Germanwings volume

2.2.4. Content Extraction

As presented in D2.2 content extraction task is composed of three steps. Each step requires a specially designed and developed module adapted to each of the sources aimed by SENSEI.

- Boiler Plate Detection



- Content Extraction
- Structure Parsing

As data sources continually evolve and make changes to their respective web pages the SENSEI specialized parsers have been updated when necessary. When analysts, partners or the system have detected parsing problems for a given source the involved parser has been updated and fixed.

For Period 2 new parser have been developed for Critizen¹, el Confidencial Digital², Corriere³ (to get the “mood” on comments) and Zucconi’s blog⁴.

2.2.5. Pre-processing

All documents crawled for the SENSEI data collection are pre-processed using the Websays pipeline.

The main components for pre-processing documents are:

- **Language Detection:** as mentioned in D2.2 it is very challenging for short texts (especially if they contain brands, acronyms, etc.). The method used to detect the language of a post is:
 - Fast look-up for similar texts with language label corrected by a human
 - Remove terms that can mislead the automatic classifier
 - Character heuristics for alphabet-specific languages (e.g. Japanese, Russian)
 - Dictionary based frequent expressions
 - A HMM based on character n-gram is used to detect the most likely languages
 - A topic-specific error cost-matrix is used to correct biases (or boost specific languages) for each specific topic
- **Online-Terms Detection:** a set of regular expressions are used to identity URLs, smileys, @authors, hashtags, retweet and forward notations, etc.
- **URL normalization:** URLs in text are typically expressed as relative or partially specified paths, and they can use URL shorteners. In this step URLs are normalized and resolved so that they lead to their full unique URL. During 2015 there have been an special effort to improve URL normalization taking into account new trends in parameterization in newspaper content URLs.
- **Named Entity Detection:** a combined approach is used to named entity detection:
 - Dictionary lookup method. Human analysts built the dictionaries.
 - A CRF model trained on a standard generic named entity corpus is used to detect named entities in English, French, Italian, Spanish and Portuguese
- **Sentiment Detection:** a combined approach is used for sentiment detection:

¹ www.critizen.es

² www.elconfidencial.com

³ www.corriere.it

⁴ zucconi.blogautore.repubblica.it

- A weighted-dictionary method is used to detect clearly positive and negative expressions for Spanish, Catalan, English, Italian, French and German.
- A proprietary nearest-neighbour based method is used to detect similar posts

2.2.6. Data Statistics

Year 2 data collection contains over 10 million posts and over 1.3 million conversations. Data is considered from 2014-11-01 to 2015-07-31. We will present detailed information about “General News and Newspaper Social Media Publications” and “RATP”.

2.2.6.1. General News and Newspaper Social Media Publications

Figure 5 shows the monthly crawled posts and it shows that more than 1 M posts is crawled per month. There are over 9 million posts and 1.3 million conversations.

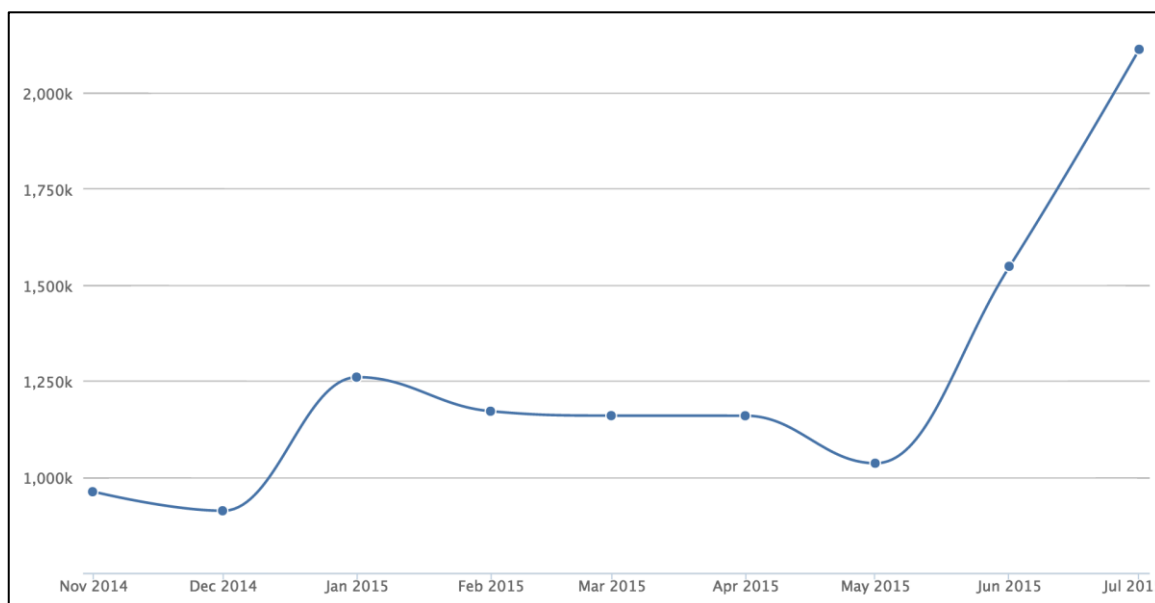


Figure 5: monthly volume

Most frequent domains (and number of posts per domain):

www.youtube.com	93440
plus.google.com	92439
www.corriere.it	34063
www.theguardian.com	17551
www.lemonde.fr	16119
www.independent.co.uk	6323
xml2.temporeale.corriereobjects.it	4639
www.reuters.com	1894
instagram.com	1872
video.corriere.it	1821
vimeo.com	1532
www.thetimes.co.uk	1090
roma.corriere.it	957

milano.corriere.it	893
www.nytimes.com	892
www.istanbulfinanshaber.com	856
zucconi.blogautore.repubblica.it	709
www.chron.com	693
uk.reuters.com	684
www.reddit.com	647
linkis.com	644
www.ft.com	615
www.dailymotion.com	575
timesofindia.indiatimes.com	502
www.newslocker.com	477
in.reuters.com	475
campus.lemonde.fr	425
abonnes.lemonde.fr	420

Most used languages are English, Italian and French as shown in Figure 6.

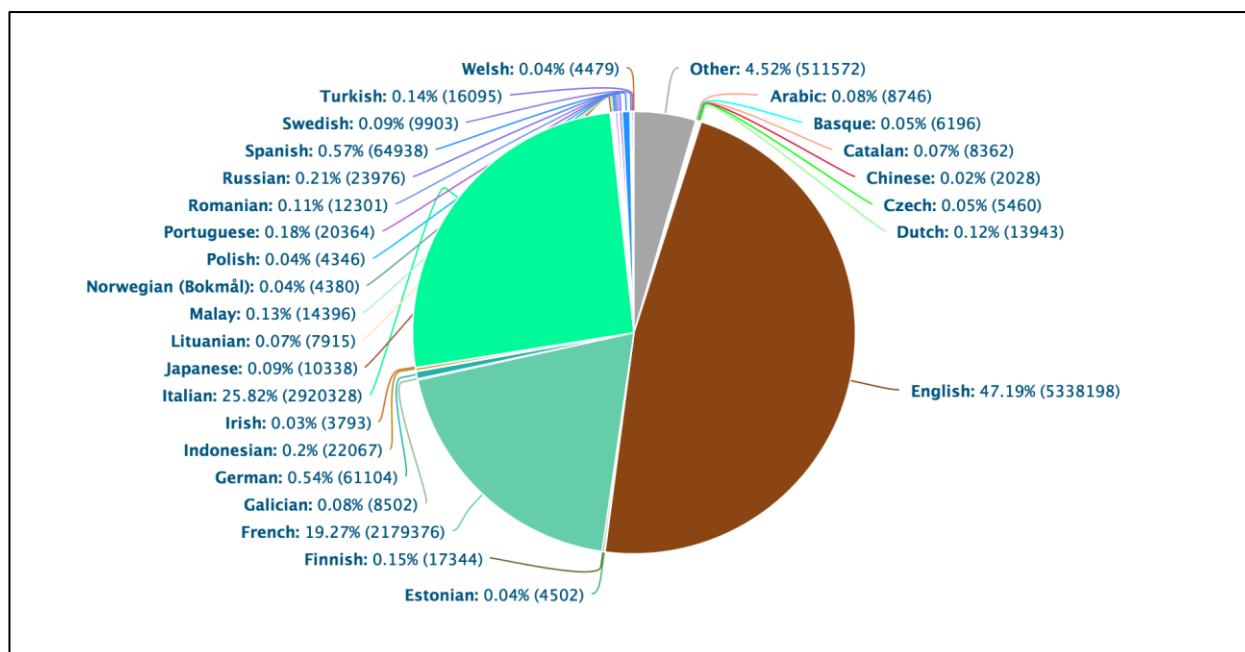


Figure 6: used languages

Most frequent author location strings are presented in Figure 7.



Figure 7: author locations

2.2.6.2. RATP

There are 0.4 million posts with 163k conversations.

The monthly crawled data is presented in Figure 8.

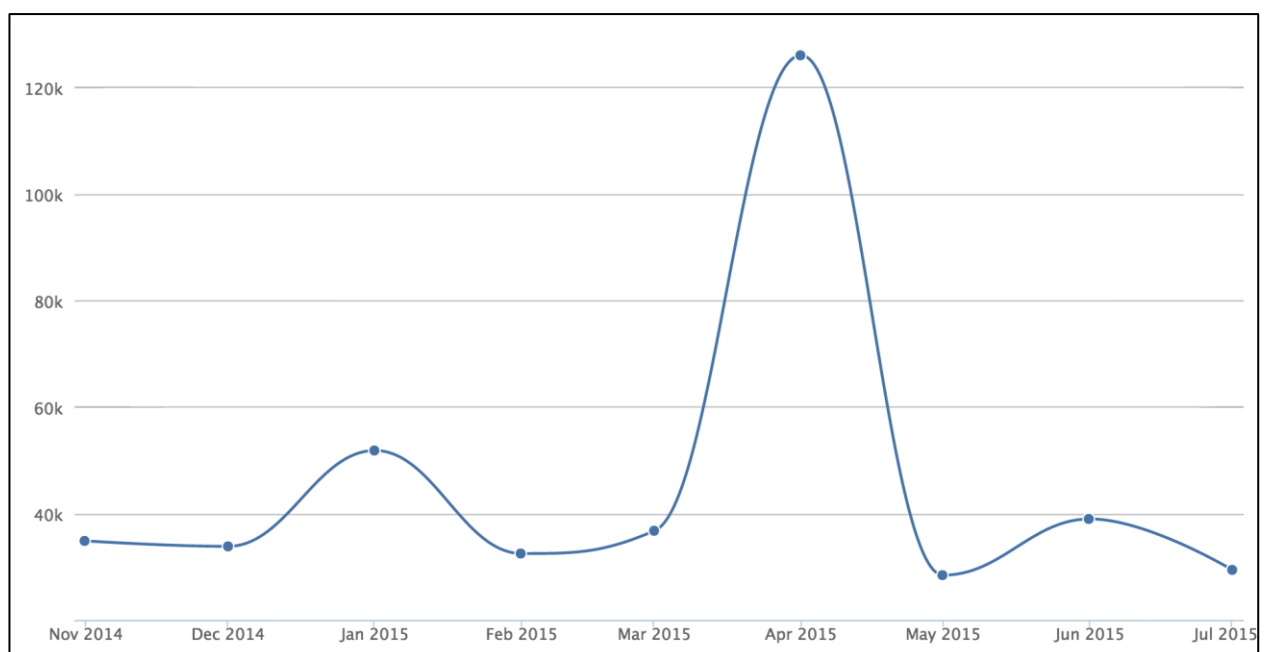


Figure 8: RATP volume

Most frequent domains (and number of posts per domain):

instagram.com	7430
www.youtube.com	3966
plus.google.com	1142
www.lefigaro.fr	826
bootstrap.liberation.fyre.co	478
www.liberation.fr	366
www.lemonde.fr	328
www.wizbii.com	327
vimeo.com	165
www.leparisien.fr	140
www.20minutes.fr	103
www.ratp.fr	76
ile-de-france.infosreg.fr	75
cyber-actu.com	59
www.vianavigo.com	57
karepmudhewe.com	54
www.lepoint.fr	54
www.mobilicites.com	53
www.ghazli.com	48
fr.news.yahoo.com	47
www.rtl.fr	45
www.francetvinfo.fr	41
www.lesechos.fr	37
soundcloud.com	35
france3-regions.francetvinfo.fr	32
tomelbezphotography.tumblr.com	31
www.metronews.fr	31
lactualite24.com	28

Most used language is French as shown in Figure 9.

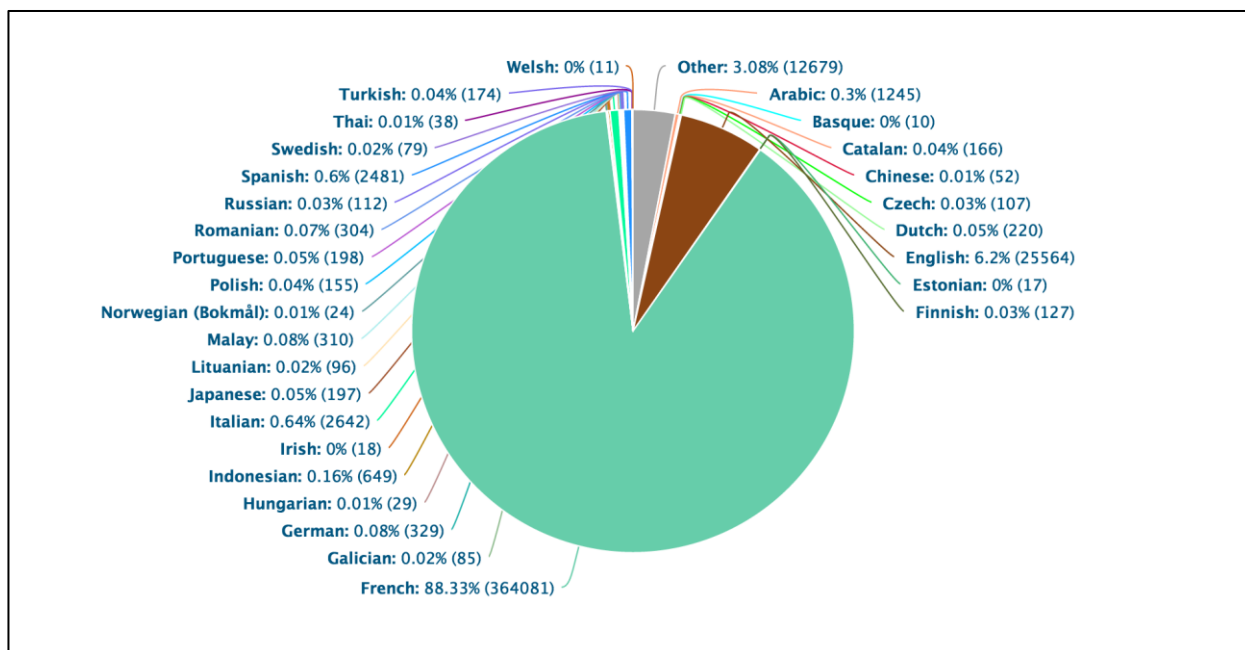


Figure 9: RATP languages

Most frequent author location strings are presented in Figure 10.



Figure 10: RATP author locations

2.2.7. Data storage and navigation improvements

2.2.7.1. Time based index partitioning (sharding)

As seen in Figure 5 data on SENSEI account has been growing constantly. SENSEI data is stored on a Websays dedicated server. To make the SENSEI account faster and able to handle to constant high volume growth the SENSEI data has been split and older data has been moved to another dedicated server.

Solr is the technology used to index and query for the SENSEI data using the Websays Dashbord. When the amount of data to be indexed is large, sometimes it s good to split it in parts and distribute those parts in different machines or nodes. Each of those Solr parts are called shards (Figure 11).

SENSEI data has been split and two shards have been created. A new server has been assigned to SENSEI to store the newly created Solr shard. Any time a query to the SENSEI account is done in the Websays dashboard the system automatically queries both machines and its respective shards and combines the results as if only one Solr server had been involved.

To be able to take these technologies into production a set of tools have been developed which allow WEBSAYS administrators to split, dump and index data in the different servers.

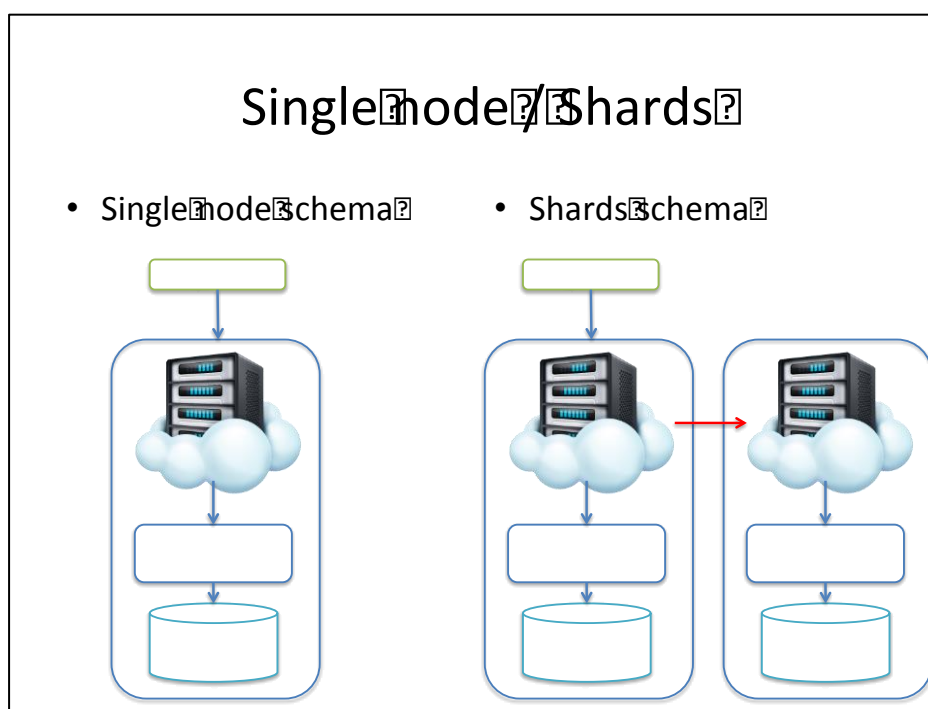


Figure 11: Sharding

2.2.7.2. Sorting data by Conversation Size

Sometimes, the large variety of data in the SENSEI account, as shown in 2.2.6 (Data Statistics), makes it difficult for dashboard users to easily select only posts with comments (not single posts). To help partners interested in such kind of content a new “sort” system has been added

to the Websays dashboard. The “Conversation Size” sort system let users to sort their selected posts by the size of their conversations.

Figure 12 shows a capture of the Websays dashboard showing that there are 26k posts for 13th of August in the SENSEI account. Using the Conversation Size sort system the user can see on top of the search the posts with highest number of comments.

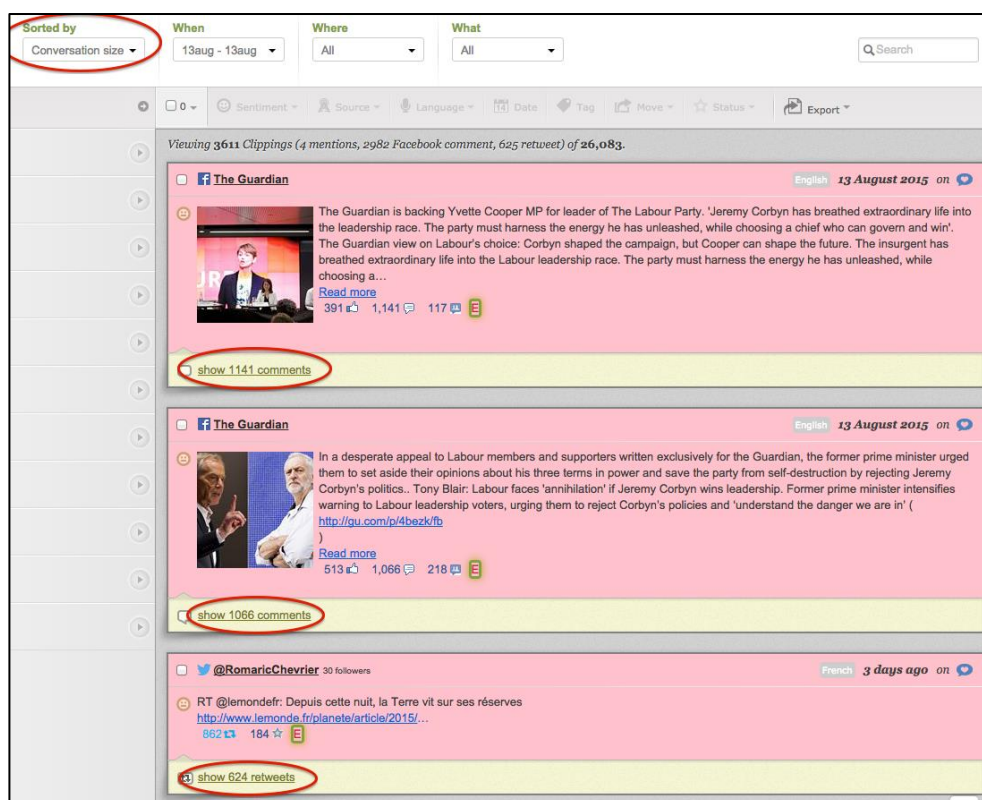


Figure 12: Websays Dashboard sort by Conversation Size

Sorting by Conversation Size is more complex than the already existing sort types in the WEBSAYS dashboard. While other systems require a single query to the corresponding node the conversation Size sort system requires multiple queries: first of all a facet query is done to get of the conversations with the highest amount of comments. Afterwards, a query is performed for each Id to be shown in the dashboard to obtain the desired posts and its comments. This sorting system produces a heavy load on the system. On the other hand, it is very useful for users. Feedback from partners working in WP3 and WP4 has been very positive.

Moreover, in the second year of the project, the development of the “conversation size” metric for SENSEI (D2.2) opened a new venue in ranking content in Websays. Again, this was introduced in the main Websays pipeline, and today 50% of clients benefit from this development. Due to performance constrains, we cannot yet bring this feature to our largest costumers (in terms of volume of data), so R&D continues on this front.

This is a feature improvement which finds its way in several Websays tools:

- New “Sort By” option sorting by conversation size allows client to quickly view the social media items of most current impact and engagement.



- New “Top-10” report shows to our clients the top-10 items (sorted by conversation size) of each of their social media channels
- New Alerts: when then conversation-size metric grows beyond a certain threshold (increasing over time) the client receives an SMS and/or email alert.

3. Conclusions

With respect to social media data collection and preprocessing, work started in Period 1 continued throughout Period 2, adding new parsers, new topics of conversation. In order to improve user access and browsing of the data we had to develop new index mechanisms (time based index partitioning) and new ranking algorithms (conversation size) which successfully allowed consortium partners to query and browse the data in order to find the most relevant information.

Some of these developments have also been exploited commercially: by integrating them into the Websays pipeline, Websays clients are already benefitting.

With respect to speech, the main activities for Period 2 have been focused on the review of Global Guide line related to the ACOF and Synopses, defined during meeting, workshops and internal (TP) and external (UNITN) Calibration. All the ACOF have been reviewed and re-annotated on the LUNA and DECODA corpus. At the end of Period 2 started the planning and also the deployment of Evaluation phases, based on intrinsic and extrinsic methodology. To support the Annotation and Evaluation phases the ACOF tools has been enriched of new features and the Elastic Search/Kibana solution has been adopted.

We have also worked on the resolution of IPR issues on the collected data (in synch with WP8), use case development with WP1 and pre-processing and data preparation issues with the rest of the packages.

In Period 3, we will continue data collection and pre-processing, with special focus on the needs of the prototypes developed. The coexistence of machine annotated data and human annotated data in Elastic Search, in conjunction with the reporting features of Kibana, will better support data analysis activities and the prototype Evaluation phases. As the prototype takes shape, Teleperformance will start proposing sensei technology to its clients and partners.

References

- [Artiles et al., 2007] Artiles, J., Gonzalo, J., and Sekine, S. (2007) "The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task", In Proc. of SemEval.
- [Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008) "Intercoder agreement for Computational Linguistics", Computational Linguistics, 34(4).
- [Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003) The Logic of Conversation. Cambridge University Press.
- [Bagga and Baldwin, 1998] Bagga, A., Baldwin, B. (1998) "Entity-based cross-document coreferencing using the vector space model", In Proc. of COLING/ACL.
- [Baker et al., 1998] Baker, F.C., Fillmore, J.C., and Lowe, B.J. (1998) "The Berkeley FrameNet project", In Proc. of COLING/ACL.
- [Barzilay and Lee, 2004] Barzilay, R. and Lee, L. (2004) "Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization", In Proc. of NAACL-HLT.
- [Bazillon et al., 2012] Bazillon, T., Deplano, M., Bechet, F., Nasr, A., and Favre, B. (2012) "Syntactic annotation of spontaneous speech: application to call-center conversation data", In Proc. of LREC.
- [Bechet and Nasr, 2009] Bechet, F. and Nasr, A. (2009) "Robust dependency parsing for Spoken Language Understanding of spontaneous speech", In Proc. of INTERSPEECH.
- [Bechet et al., 2012] Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Bèze, M., De Mori, R., and Arbillot, E. (2012) "DECODA: a call-center human-human spoken conversation corpus", In Proc. of LREC.
- [Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006) "Domain Adaptation with Structural Correspondence Learning", In Proc. of EMNLP.
- [Byrd et al., 2008] Byrd, R.J., Neff, M.S., Teiken, W., Park, Y., Cheng, K.S.F, Gates, S.C., and Visweswariah, K. (2008) "Semi-automated logging of contact center telephone calls", In Proc. of CIKM.
- [Carlson et al., 2001] Carlson, L., Marcu, D., and Okurowski, M.E. (2003) "Building a discourse-tagged corpus in the framework of rhetorical structure theory". In J. Kuppevelt and R. Smith (eds) Current Directions in Discourse and Dialogue. Kluwer.
- [Chambers and Jurafsky, 2011] Chambers, N. and Jurafsky, D. (2011) "Template-based information extraction without the templates", In Proc. of ACL.
- [Chen and Martin, 2007] Chen, Y. and Martin, J. (2007) "Towards robust unsupervised personal name disambiguation", In Proc. of EMNLP.
- [Coppola et al., 2009] Coppola, B., Moschitti, A., and Riccardi, G. (2009) "Shallow Semantic Parsing for Spoken Language Understanding", In Proc. of NAACL.
- [Csomai and Mihalcea, 2008] Csomai, A. and R. Mihalcea (2008) "Linking documents to encyclopedic knowledge", IEEE Intelligent Systems.



[Daumé, 2007] Daumé III H. (2007) "Frustratingly Easy Domain Adaptation", In Proc. of ACL.

[Dinarelli et al., 2009] Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., and Riccardi, G. (2009) "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics", In Proc. of EACL Workshop on Semantic Representation of Spoken Language.

Appendix A: Elastic Search

Elastic Search JSON Element Description

Element	Type	Data Field Description
FileName	String	Name of the Conversation file
Service	String	Value in (DECODA,LUNA)
ScoreQuestion<i>	String	<i> i=1....12 ScoreQuestion <i> contains the Score(PASS,FAIL,N/A) for question_id<i>
ScoreQuestionValue<i>	Float	<i> i=1....12 ScoreQuestionValue <i> contains the Score for question_id<i> calculated taking into account the weight assigned to question_id<i> and the value selected for ScoreQuestion <i>
Turn<i>	String	<i> i=1....12 Turn<i> contains the Segment turns selected by Quality Assurance Professional answering question_id<i>
FlagGeneral<i>	String	<i> i=1....12 FlagGeneral<i> contains the value Y or N of FlagGeneral for question_id<i> FlagGeneral<i> equal 'Y' indicate that there is no relevant speech turn for question_id<i>
Note<i>	String	<i> i=1....12 Note<i> contains the Note related to question_id<i>
Synopsis	String	Synopsis filled by Quality Assurance professional
SynopsisPredicted	String	Synopsis machine generated and provided by University partners
Scorelisten	Float	Overall score of the Monitoring Form automatically calculated on the basis of weighted average
Comment	String	Overall Comment of the Monitoring Form
StartDate	Date	Date & Time when the Monitoring Form has been loaded by Quality Assurance Professional to start the evaluation of the conversation. With the EndDate define the time spent to compile the monitoring form.
EndDate	Date	Date & Time when the Monitoring Form has been Completed and saved by Quality Assurance professional. With the StartDate define the time spent to compile the monitoring form.
idListen	integer	Identifier of the Monitoring form saved in the MySQL database
idUser	integer	Identifier of the User
sysdate	date	System insert date
sysdatemod	date	Last Modified date
LenSynopsis	integer	Length as number of the characters contained in Synopsis
IpAddress	String	IP Address of the user machine

Elastic Search JSON Format

```
{
  "mappings":{
    "_default_":{
      "properties":{
        "Comment":{
          "type":"string"
        },
        "EndDate":{
          "type":"date",
          "format":"dateOptionalTime"
        },
        "FileName":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral11":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral10":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral11":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral12":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral2":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral3":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral4":{
          "type":"string",
          "index":"not_analyzed"
        },
        "FlagGeneral5":{
          "type":"string",
```



```

    "index": "not_analyzed"
  },
  "FlagGeneral16": {
    "type": "string",
    "index": "not_analyzed"
  },
  "FlagGeneral17": {
    "type": "string",
    "index": "not_analyzed"
  },
  "FlagGeneral18": {
    "type": "string",
    "index": "not_analyzed"
  },
  "FlagGeneral19": {
    "type": "string",
    "index": "not_analyzed"
  },
  "IpAddress": {
    "type": "string",
    "index": "not_analyzed"
  },
  "LenSynopsis": {
    "type": "integer"
  },
  "Note1": {
    "type": "string"
  },
  "Note10": {
    "type": "string"
  },
  "Note11": {
    "type": "string"
  },
  "Note12": {
    "type": "string"
  },
  "Note2": {
    "type": "string"
  },
  "Note3": {
    "type": "string"
  },
  "Note4": {
    "type": "string"
  },

```

```

    "Note5":{
      "type":"string"
    },
    "Note6":{
      "type":"string"
    },
    "Note7":{
      "type":"string"
    },
    "Note8":{
      "type":"string"
    },
    "Note9":{
      "type":"string"
    },
    "ScoreQuestion1":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion10":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion11":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion12":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion2":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion3":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion4":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion5":{
      "type":"string",
      "index":"not_analyzed"
    }

```

```

    },
    "ScoreQuestion6":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion7":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion8":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestion9":{
      "type":"string",
      "index":"not_analyzed"
    },
    "ScoreQuestionValue1":{
      "type":"float"
    },
    "ScoreQuestionValue10":{
      "type":"float"
    },
    "ScoreQuestionValue11":{
      "type":"float"
    },
    "ScoreQuestionValue12":{
      "type":"float"
    },
    "ScoreQuestionValue2":{
      "type":"float"
    },
    "ScoreQuestionValue3":{
      "type":"float"
    },
    "ScoreQuestionValue4":{
      "type":"float"
    },
    "ScoreQuestionValue5":{
      "type":"float"
    },
    "ScoreQuestionValue6":{
      "type":"float"
    },
    "ScoreQuestionValue7":{
      "type":"float"
    }

```

```

    },
    "ScoreQuestionValue8":{
      "type":"float"
    },
    "ScoreQuestionValue9":{
      "type":"float"
    },
    "Scorelisten":{
      "type":"float"
    },
    "Service":{
      "type":"string",
      "index":"not_analyzed"
    },
    "StartDate":{
      "type":"date",
      "format":"dateOptionalTime"
    },
    "Synopsis":{
      "type":"string"
    },
    "SynopsisPredicted":{
      "type":"string"
    },
    "Turn1":{
      "type":"string"
    },
    "Turn10":{
      "type":"string"
    },
    "Turn11":{
      "type":"string"
    },
    "Turn12":{
      "type":"string"
    },
    "Turn2":{
      "type":"string"
    },
    "Turn3":{
      "type":"string"
    },
    "Turn4":{
      "type":"string"
    },
    "Turn5":{

```

```

        "type": "string"
      },
      "Turn6": {
        "type": "string"
      },
      "Turn7": {
        "type": "string"
      },
      "Turn8": {
        "type": "string"
      },
      "Turn9": {
        "type": "string"
      },
      "idListen": {
        "type": "integer"
      },
      "idUser": {
        "type": "integer"
      },
      "sysdate": {
        "type": "date",
        "format": "dateOptionalTime"
      },
      "sysdatemod": {
        "type": "date",
        "format": "dateOptionalTime"
      }
    }
  }
}

```

ES report Comparison

The Report function of the ACOF tool v2.0 is a user friendly interface developed to run query on Elastic Search by simply selecting some filtering criteria and have the result formatted in a tabular output as shown in Figure 1.

Visualizza il report degli ascolti effettuati.

Data o periodo:

Data Inserimento/Modifica

Insert Date

User

Servizio

Filename

Note

Considerazioni

Synopsis

Synopsis Predicted

☐ Considerazioni
 ☐ Synopsis
 ☐ Synopsis Predicted
 ☐ Monit. Start Date
 ☐ Monit. End Date

Subitem	Score	Mostra colonne
1 - Rispetta la procedura di apertura	PAS	<input checked="" type="checkbox"/>
2 - Ascolta in maniera attiva e pone domande pertinenti	FAIL	<input checked="" type="checkbox"/>
3 - Espone le informazioni in maniera chiara, completa ed essenziale		<input type="checkbox"/>
4 - Gestisce le obiezioni tranquillizzando il cliente e puntando sempre sulla sua soddisfazione		<input type="checkbox"/>
5 - Gestisce con sicurezza la telefonata		<input type="checkbox"/>
6 - Utilizza parole positive		<input type="checkbox"/>
7 - Segue lo script di Chiusura		<input type="checkbox"/>
8 - E' educato e propositivo con il cliente		<input type="checkbox"/>
9 - E' in grado di adeguarsi allo stile di comunicazione dell'interlocutore mantenendo sempre professionalità		<input type="checkbox"/>
10 - Gestione dei tempi: negozia l'attesa motivandone sempre la ragione		<input type="checkbox"/>
11 - Capacità di ascolto		<input type="checkbox"/>
12 - Prende in carico la problematica con ricontatto successivo		<input type="checkbox"/>

Cerca ☐ Elasticsearch query

Reset

Risultati: 73 - 7	ID	Transcription file	Service	Score %	Insert Date	Modify Date	Inserted By	Subitem 1	Subitem 2
Excel	699	20091112_RATP_SCD_0122.trs	DECODA	45	2014-09-26 18:28:37	2015-03-05 12:27:00	cosima.caramia	PASS	FAIL

Figure 13: ACOF Report view

The Kibana user interface is more complicated than ACOF report, because user has to write the query using logical operator “AND” / “OR” and the format of the result is not customizable.

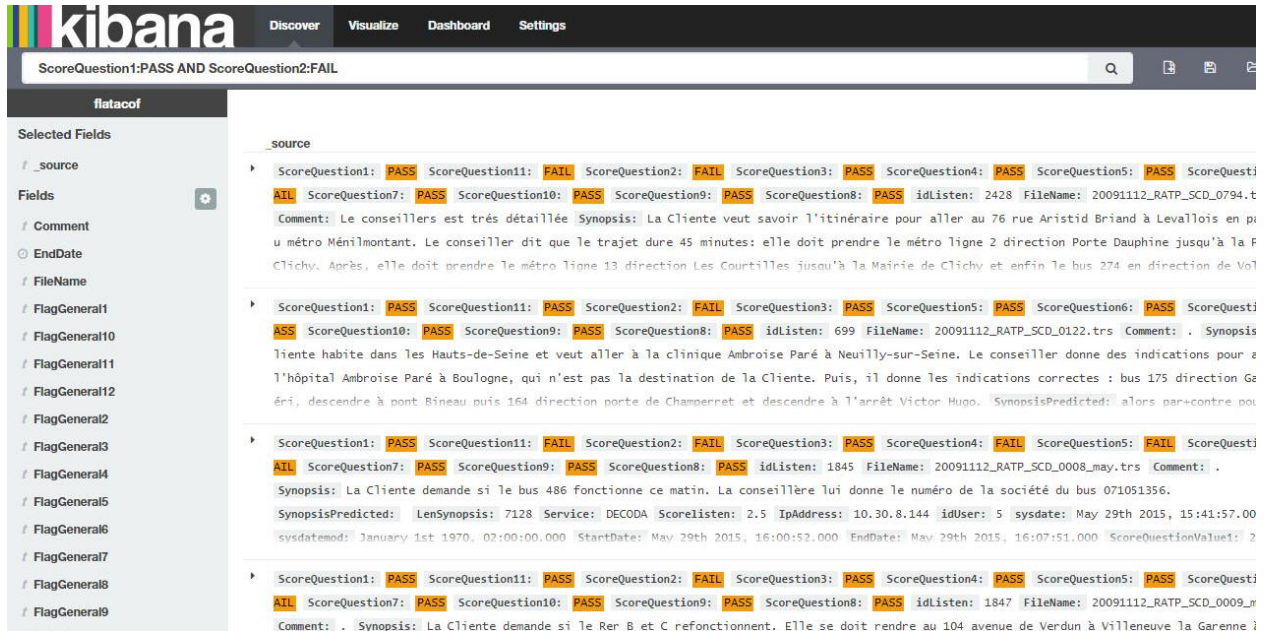


Figure 2: Kibana Report view

Another interface to run query on Elastic Search is to use the Marvel native plugin, that let user to write the query and read the result in JSON format, as shown in figure 3 below.

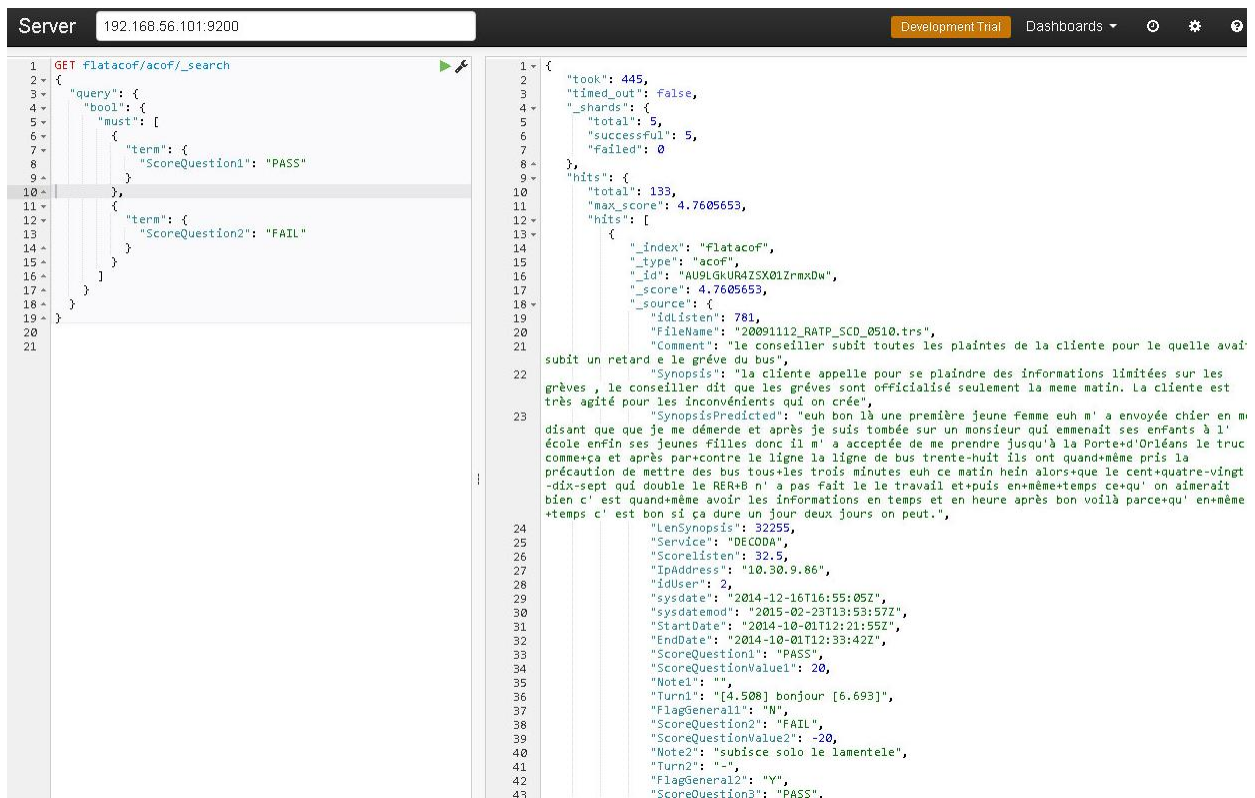


Figure 3: Marvel Elastic Search view

Appendix B: SENSEI ACOF tool v2.0

My SQL Data Model

In version 2.0 has been added the table tblSynopsis which contains the synopsis automatically generated by machine systems.

The data model of the application v2.0 is composed of eight tables, as illustrated in figure 9 below.

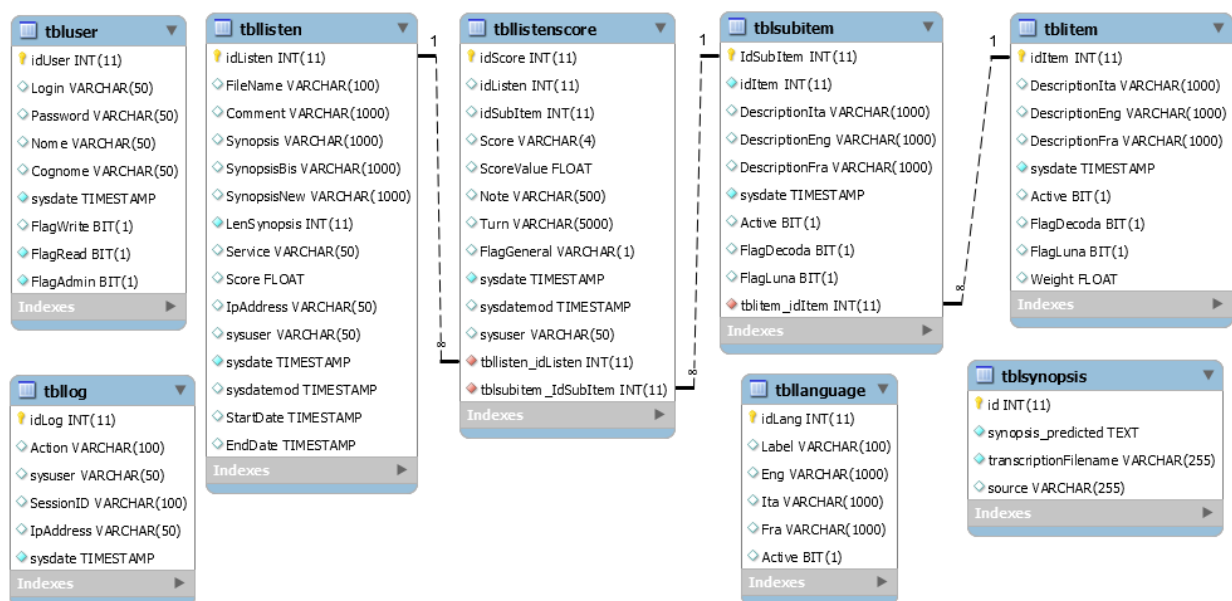


Figure 14: Data model of SENSEI ACOF tool v2.0

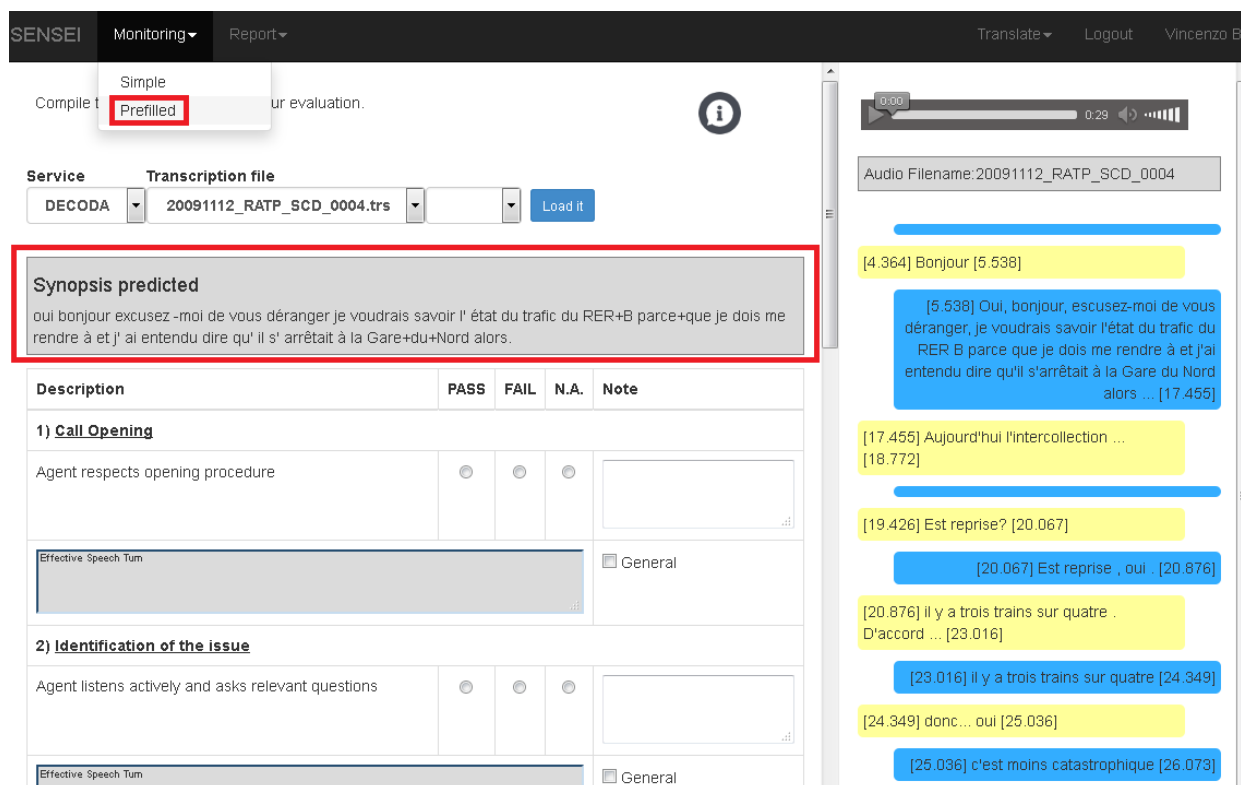
New View: Monitoring prefilled

The new view Monitoring prefilled has the same structure and functions of the Monitoring view described in D2.2, with the following additional features:

- At the top of the page user visualizes the new field “predicted synopsis” which contains the synopsis automatically generated and provided by University partners.
- The answer to the question are pre-marked with the value automatically generated (if present) provided by University partners.

With these new functions, when a user loads a transcription, some field's value are automatically set (i.e. question n. 2 sets as PASS) and if a Synopsis is available, it appears on top of the form as showed in Figure 2.

The data are stored not only in MySQL Database but also in Elastic Search.



SENSEI Monitoring Report Translate Logout Vincenzo Blé

Compile t Prefilled ur evaluation.

Service Transcription file

DECODA 20091112_RATP_SCD_0004.trs Load it

Synopsis predicted

oui bonjour excusez -moi de vous déranger je voudrais savoir l' état du trafic du RER+B parce-que je dois me rendre à et j' ai entendu dire qu' il s' arrêta à la Gare+du+Nord alors.

Description	PASS	FAIL	N.A.	Note
1) Call Opening				
Agent respects opening procedure	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Effective Speech Turn	<input type="checkbox"/> General			
2) Identification of the issue				
Agent listens actively and asks relevant questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
Effective Speech Turn	<input type="checkbox"/> General			

Audio Filename: 20091112_RATP_SCD_0004

[4.364] Bonjour [5.538]

[5.538] Oui, bonjour, excusez-moi de vous déranger, je voudrais savoir l'état du trafic du RER B parce que je dois me rendre à et j'ai entendu dire qu'il s'arrêterait à la Gare du Nord alors ... [17.455]

[17.455] Aujourd'hui l'intercollection ... [18.772]

[19.426] Est reprise? [20.067]

[20.067] Est reprise, oui. [20.876]

[20.876] il y a trois trains sur quatre. D'accord ... [23.016]

[23.016] il y a trois trains sur quatre [24.349]

[24.349] donc... oui [25.036]

[25.036] c'est moins catastrophique [26.073]

Figure 2: Monitoring Prefilled view

New View: Report Elastic Search

The new report is based on Elastic Search and let user run complex query as required for the prototype Evaluation phase described in D1.3.

The form accepts many filtering conditions and allows user to choose which fields to show in the results table. The filtering conditions are all in AND logic, that means the system will return the records that meet all the filtering conditions provided in the search.

SENSEI

Monitoring

Report

Translate

Logout

Vincenzo Blè

You show the monitoring report made.

Date or period:

Inserted/Modified Date

User

Service

Filename

Note

Considerations

Synopsis

Synopsis Predicted

☐ Considerations
 ☒ Synopsis
 ☒ Synopsis Predicted
 ☐ Monit. Start Date
 ☐ Monit. End Date

Subitem	Score	Show column
1 - Agent respects opening procedure	PASS	<input checked="" type="checkbox"/>
2 - Agent listens actively and asks relevant questions	FAIL	<input checked="" type="checkbox"/>
3 - Agent shows the information in a clear, comprehensive and essential way		<input checked="" type="checkbox"/>
4 - Agent manages the objections reassuring the customer and always focusing on client satisfaction		<input checked="" type="checkbox"/>
5 - Agent manages the call with safety		<input type="checkbox"/>
6 - Agent uses positive words		<input type="checkbox"/>
7 - Agent follows the closing script		<input type="checkbox"/>
8 - Agent is polite and proactive with the customer		<input type="checkbox"/>
9 - Agent is able to adapt to the style of client's communication always maintaining professionalism		<input type="checkbox"/>
10 - Agent Management: he negotiates the wait always giving reasons		<input type="checkbox"/>
11 - Ability to listen		<input type="checkbox"/>

Search

☐ Elasticsearch query

Reset

Filtering conditions in AND

Columns to show in the result

Figure 3: Elastic Search Report - Query view

SENSEI		Monitoring		Report		Translate		Logout		Vincenzo Biè				
Results: 26 - 2428		ID	Transcription file	Synopsis	SynopsisPredicted	Service	Score %	Insert Date	Modify Date	Inserted By	Subitem 1	Subitem 2	Subitem 3	Subitem 4
Excel														
		674	20091112_RATP_SCD_0017.trs	le client demande s'il y a de grÃ©ves sur les lignes de bus ce matin et s'il y a des informations prÃ©cises aux dates de passage approximatif. Le conseiller rÃ©spond qu'elle n'a pas des informations precises et les horaires, le client demande aussi du bus 162 et 182 le conseiller confirme que le 162 marchent normalement mais le 182 est en grÃ©ve	d'accord donc euh un sur quatre mais il n' y a pas de moyen de savoir.	DECODA	-7.5	2014-12-16 12:59:03	2015-02-24 14:28:15	a	PASS	FAIL	PASS	FAIL
		699	20091112_RATP_SCD_0122.trs	La Cliente habite dans les Hauts-de-Seine et veut aller Ã la clinique Ambroise ParÃ© Ã Neuilly-sur-Seine. Le conseiller donne des indications pour aller Ã l'hÃ´pital Ambroise ParÃ© Ã Boulogne, qui n'est pas la destination de la Cliente. Puis, il donne les indications correctes : bus 175 direction Gabriel PÃ©ri, descendre Ã pont Bineau puis 164 direction porte de Champerret et descendre Ã l'arrÃªt Victor Hugo.	alors par+contre pour rÃ©cupÃ©rer le bus suivant c'est pas au mÃªme arrÃªt parce+que le cent+soixante-quinze vous laisse sur le quai hein. oui par+rapport+au au cent cinquante sept qui passe dans le secteur sur le quai en+fait hein au+niveau+du quai donc pour aller au l' HÃ´pital+Ambroise+ParÃ© on va dire euh vous avez l' arrÃªt Bellini qui est pas trÃ¨s loin aussi.	DECODA	45	2014-09-26 18:28:37	2015-03-05 12:27:00	co	PASS	FAIL	PASS	NA
		1497	20091112_RATP_SCD_0122.trs	Le client appelle pour savoir comment aller	alors par+contre pour rÃ©cupÃ©rer le bus	DECODA	5	2015-02-24 13:01:56	2015-03-09 15:36:11	m	PASS	FAIL	FAIL	PASS

Figure 4: Elastic Search Report - Result view

Data Export to Excel

Export function allows all authenticated user to extract off Elastic Search all annotated data. The Excel format and the data to include in the export, are the same of the result table in the Elastic Search Report (see previous Figure 4).

Appendix C: State of Data Collections

Size of Speech Data per Language

Speech Data Sets Available	TP Annotation Activity Y1 (M1-12)	TP Annotation Activity Y2 (M13-24)	Language
572 LUNA dialogs	200 different dialogs have been annotated with AOF included segment turn and COF.	359 different dialogs have been annotated with ACOF included segment turn.	Italian
1500 DECODA Conversations	118 different conversations have been annotated with AOF included segment turn and COF.	308 different dialogs have been annotated with ACOF included segment turn.	French

Size of Social Data Sets Per Language

Type	Language	N. of Dialogues or posts	N. of tokens
Social Media, News and Blogs	English	8M	>800M
Social Media, News and Blogs	French	3.2M	>320M
Social Media, News and Blogs	Italian	3.8M	>380M
Social Media, News and Blogs	Spanish	.4M	>40M
Social Media, News and Blogs	Other Languages	1.1M	>110M