D4.1 OVERVIEW AND ANALYSIS OF LAWFULLY INTERCEPTED AND PUBLICLY AVAILABLE DATA

Grant Agreement:	833635
Project Acronym:	ROXANNE
Project Title:	Real time network, text, and speaker analytics for combating organised crime
Call ID: **Call name:**	H2020-SU-SEC-2018-2019-2020, Technologies to enhance the fight against crime and terrorism
Revision:	V1.4
Date:	30 December 2019
Due date:	31 December 2019
m.2 2242 512	SAIL
Work package:	WP4
Type of action:	RIA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Disclaimer

The information, documentation and figures available in this deliverable are written by the "ROXANNE - Real time network, text, and speaker analytics for combating organised crime" project's consortium under EC grant agreement 8833635 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

Copyright notice

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Nature of deliverable:	R	
Dissemination Level		
PU	Public	☒
CO	Confidential, only for members of the consortium (including the Commission Services)	☐
EU-RES	Classified Information: RESTREINT UE (Commission Decision 2015/444/EC)	☐

* R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Revision history

Revision	Edition date	Author	Modified Sections / Pages	Comments
V1.0	17.12.2019	Erinc Dikici (SAIL)	All	Original draft
V1.1	17.12.2019	Petr Motlicek (Idiap)	All	Modifications
V1.2	22.12.2019	UCSC, LUH, AEGIS, NFI	All	Feedback
V1.3	24.12.2019	Erinc Dikici, Gerhard Backfried (SAIL)	All	Modifications
V1.3	28.12.2019	Petr Motlicek (Idiap)	All	Modifications/comments
V1.3	30.12.2019	Theoni Spathi (KEMEA)	All	Modifications/comments
V1.4	30.12.2019	Erinc Dikici (SAIL)	All	Integrated modifications/comments
V1.4	31.12.2019	ITML, KEMEA	All	Comments
V1.4	31.12.2019	Petr Motlicek (Idiap)	All	Modifications/comments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Executive summary

This deliverable D4.1: Overview and analysis of lawfully intercepted and publicly available data is in reference to tasks T4.1: Inventory and analysis of lawfully intercepted data and T4.2: Overview of publicly available data resources for research, training and development of the ROXANNE project. Its purpose is to provide information on all datasets that have been suggested or made into use by the partners of the ROXANNE consortium. It also attempts to summarize the use of data within ROXANNE with respect to the data sources, data types and technologies. As such, it contains a reviewed and modified version of some ideas presented in the Grant Agreement as well as in the deliverable D10.17: Requirement No.20.

This deliverable is a living document; it will be updated as the project progresses and as new datasets, data types or technologies are made available to the consortium. Updates and extensions to this document will also be covered in D4.2: Simulated data for development and demonstration (due M16), D4.3: Final report on ROXANNE data (due M30), and D1.3: ROXANNE's data management plan (due M6, to be updated in M18 and M36).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

1. Introduction

This deliverable "D4.1: Overview and analysis of lawfully intercepted and publicly available data", is the first report to be submitted as part of the "Data management" work package (WP4) of the ROXANNE project. In this introductory section, we present the definition of WP4, the purpose of this document and its scope within WP4, and the outline of this report.

1.1. Background

The objectives of WP4 are given in the ROXANNE Grant Agreement as follows [1]:

> **WP4: Data management**
>
> - *Inventory of relevant data resources (i.e. real data to be used for investigative work) within the consortium, and secure the data applicability for developing, evaluating and demonstrating the technologies involved;*
> - *Suggest and implement solutions to ingest investigative data from criminal proceedings (wiretap records, associated metadata, other lawfully intercepted data from systems installed on LEAs);*
> - *Use data from social media, either publicly available (public profiles related to the use-case), or lawfully intercepted from private profiles;*
> - *Prepare, in cooperation with LEAs, a limited amount of simulated target data to simulate the real criminal investigation scenarios, due to legal and ethical concerns around real investigative data.*

Data play a very critical role in ROXANNE, not only because they are needed to carry out the technical work (developing the tools and components, building and adapting domain-dependent models, etc.), but also because the whole ROXANNE platform is envisioned to be a data- and evidence-driven analysis tool for criminal activities. For this reason, WP4 stands in the center of the information flow in ROXANNE and serves as a common basis for other technical WPs. The relation and interaction of WP4 with the other work packages can be seen in Figure 1.

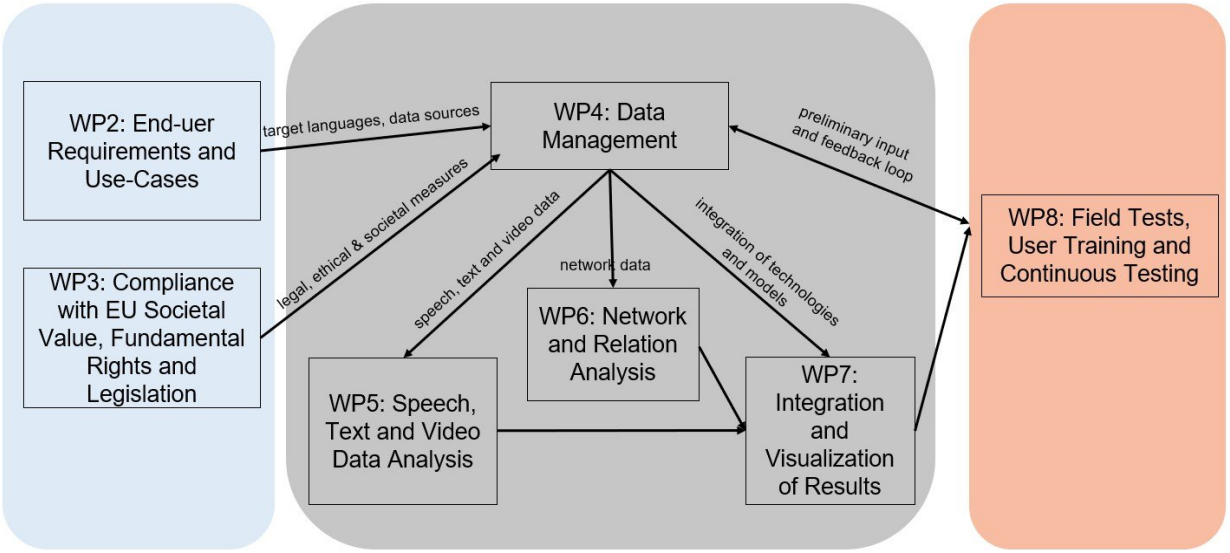

Figure 1. Relation of WP4 with the other work packages.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

1.2. Purpose and scope

The main purpose of this document is described in the ROXANNE Grant Agreement [1] as *"to provide an initial overview of investigation and public data available for ROXANNE, and the associated legal framework"*. Along these lines, this document is produced with the efforts in the tasks "T4.1: Inventory and analysis of lawfully intercepted data" and "T4.2: Overview of publicly available data resources for research, training and development", whose details will be given in Section 2.

The members of the consortium participating in T4.1 and T4.2 were asked to provide the list of datasets that they own or have access to, which they could use in ROXANNE. This deliverable contains the answers to this questionnaire, and discusses the types of datasets, the types of data involved in these datasets, and the technologies in which they will be used.

Besides T4.1 and T4.2, the other tasks in ROXANNE which this deliverable has relations to are:

- T1.5: Data management plan,
- T2.1: Collection of end-user requirements, and
- T4.6: Target data simulation.

Although the submission of this deliverable D4.1 is in M4 of the project, the work in tasks T4.1 and T4.2 will continue as planned until the end of the first project year (M12). Therefore, we expect the content of this deliverable to be improved and extended in the future. Updates and extensions to this document will be covered in the upcoming deliverables of WP4, namely the "D4.2: Simulated data for development and demonstration" (due M16) and "D4.3: Final report on ROXANNE data" (due M30), as well as in "D1.3: ROXANNE's data management plan" (due M6, to be updated in M18 and M36). This deliverable does not contain information on the legal framework for data collection and processing, as these topics have already been extensively covered in several deliverables of WP10, namely the D10.5 (Requirement No.8, M1), D10.7 (Requirement No.10, M1) and D10.9 (Requirement No.12, M1).

1.3. Document structure

This document is organized as follows: Section 2 explains the different data sources and how they will be used. Sections 3 and 4 respectively focus on the two main pillars of data in the project, namely the lawfully intercepted and publicly available data. Finally, we present the open topics and future work regarding data in Section 5. The document ends with a list of references and an appendix with a complete table of datasets.

2. Data sources, types and usage in ROXANNE

Issues related to the sources, types and usage of data in ROXANNE were first covered in Section 1.3.5 of the Grant Agreement [1]. After initial discussions in the project kick-off meeting, the sources of data have been extended and presented in the ethics deliverable D10.17: Requirement No.20 [2]. In this section we combine the ideas presented in these two documents and investigate the data-related aspects with regards to the data sources, data types, and data-driven technologies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

2.1. Data sources in ROXANNE

There are three principle sources[1] of data for technical training, development and evaluation activities in ROXANNE: the LEA partners' operations, publicly accessible media, and research activities, including ROXANNE itself. Let us now have a look at these three categories in detail:

The first data source consists of **lawfully intercepted communications (wire-tap recordings) and videos of criminal network members**. All ROXANNE LEA end-users have access to this type of data as part of their operations. The data that falls under this category can be categorized as follows:

- Resources through training exercises
  These consist of data produced by LEA personnel during training exercises, using actual equipment and mimicking as closely as possible (due to the LEA expertise) communication within organised crime networks. They closely resemble the target domain of ROXANNE.
- Resources from closed proceedings
  These consist of LEA-owned data obtained from cases which have been already closed and is not part of any investigation or court activity anymore. These data are not only realistic, but they are the closest type to the "real thing" for non-LEA members of the consortium. These data, if available, will be anonymised and in a state that allows LEAs to pass them onto (technical) project partners, and accessed through either a secure collaborative platform, or through access to the ROXANNE remote platform on KEMEA premises.
- Resources from hot, ongoing cases
  These consist of LEA-owned data in actual, current cases, with ongoing investigation or court activity. To ensure respect of relevant legal, ethical and confidentiality requirements, this category of data will only be processed by LEAs and only on their own premises, and in principle will never be shared with other consortium partners. This type will be used by LEAs in the ROXANNE evaluations. Results of the evaluations will be communicated to all project partners but in a limited and anonymised manner according to the legal obligations and possibilities of LEAs[2].

The second source of data can be generalized in the term **publicly available data**, and contains the following categories:

- Public datasets
  These could be commercial datasets collected, annotated and sold by data collection and processing companies (Datatang[3], Speechocean[4], Appen[5], etc.), datasets processed and distributed (either commercially or freely) by data distribution agencies (LDC[6], ELRA[7], etc.), or datasets that are publicly available/downloadable from research organization's and individual researchers' web pages. Many technical partners have these and use them for training and evaluation of various models.

[1] We use the term "resources" below as a placeholder for data resources, corpora, recordings, datasets, etc.

[2] More information about this can be found in Section 5 of this document.

[3] https://www.datatang.com/datatang

[4] http://en.speechocean.com/

[5] https://appen.com/

[6] Linguistic Data Consortium, https://www.ldc.upenn.edu/

[7] European Language Resources Association, http://www.elra.info/en/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- **Data collected from public sources**
  These are mostly data available online, such as websites, social media, microblogging and image/video sharing platforms such as Facebook, Twitter, YouTube, Instagram, etc. The public APIs of these sources (e.g., Twitter's streaming API, Facebook Graph-API) will be used to collect such data[8]. Existing supporting infrastructure, such as text crawlers and entity lists from the Internet are available in the consortium. Please note that this category only covers open sources, i.e., only public profiles will be used to collect data for such purposes and no private data will be "hacked". The crawled/collected dataset will not be re-distributed and the regulations of the source sites will not be violated.

The third and final source of data is the **data already available to partners** as part of their past or current commercial and research activities. These can be categorized as follows:

- Data from other (research) projects
  Resources that have been created by other (research) projects and that are available to ROXANNE partners, such as Babel[9] and LORELEI[10]. Some of these datasets may also be publicly available or distributed by the agencies mentioned above.
- Resources created and held by individual technical partners
  Data with technical partners that they have created themselves, e.g., when working on a language where no data are available from or too expensive to collect. These can be shared between project partners or be "private" to individual partners and not shared.
- Simulated data
  Data simulated by (technical) partners using input from and/or the technical infrastructure of LEAs and experts in order to be able to develop, evaluate, improve algorithms and methods. Expertise for simulation would come from LEAs and experts, and enable technical partners to simulate things in the most realistic manner possible.

As the title suggests, this document contains information on the first two types of sources, namely the investigation data and the publicly accessible data. The efforts to prepare a simulated dataset has already been planned as part of T4.7 of ROXANNE, and the outputs of this task will be published in D4.2, due M18.

2.2. Data types in ROXANNE

The overall concept of ROXANNE is based on the efficient identification and tracking of criminals from speech, text and video data combined with criminal network analysis. Based on these aims, the different types[11] of data are as follows:

- Audio/Speech
  This category refers to auditory content and is mostly used to describe speech recordings of individuals (i.e., criminal network actors in ROXANNE). The recordings may also be accompanied by annotations, such as manual transcriptions, time tags of utterances, speaker information, etc. Other auditory data may be environmental (background) sounds.

[8] Terms and conditions typically change over time. This will be monitored and considered for processing.

[9] IARPA Babel Program, https://www.iarpa.gov/index.php/research-programs/babel

[10] DARPA Low Resource Languages for Emergent Incidents (LORELEI) Program, https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

[11] We avoid using the term "modality" here as a single data type listed here can be composed of different modalities, or multiple data types can be grouped into a single modality.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

- Text
  The textual data in ROXANNE will mainly consist of web pages, posts and comments from social media platforms. Automatic speech recognition outputs, time-tagged transcriptions of spoken content, also fall into this category.
- Image/Video
  The data category regarding visuals will focus on faces of individuals, logos, objects, background scenes, either from still or moving images.
- Social network data
  This category contains data representing structural and dynamic connections between actors on a social network and their interaction with each other and with objects (e.g., two people being friends on some online social network, two phone users/numbers having conversations or exchanging messages, two users using the same phone number, etc.). Time/date information may also accompany these type of data.
- Other metadata
  This category contains other data types such as geolocation or information about points in time. These may either be accompanied with another data type (GPS information stored in the EXIF of an image) or exist as they are (knowledge on Person A's check-in to a location at a certain point in time with no concrete evidence, route-tracking of a vehicle).

The data used in ROXANNE may be recorded under different conditions (environment, language, accent, level of noise, etc.), over different communication channels (landline, mobile, VoIP), from different sources (surveillance cameras, mobile cameras) and can vary in quality (resolution, frame rate, dynamic range, used codec and compression). ROXANNE will take the necessary measures to address these issues.

2.3. Data-driven technologies in ROXANNE

ROXANNE aims to develop and combine technologies related to audio, video and natural language processing with those related to network analysis. All of these technologies are based on models which rely on training data. Hence, data form one of the central elements in ROXANNE which are required to create, train and evaluate such models of the various components. The aim is to eventually extract hidden knowledge from large data by new algorithms, components and models developed by technology partners, and to test these on data provided by end-user partners.

ROXANNE will make use of different data sources and data types in developing the following technologies/components:

- Automatic Speech Recognition
- Speaker Identification
- Language Identification
- Accent Identification
- Keyword Search
- Named Entity Detection
- Topic Detection
- Sentiment Analysis
- Face Recognition
- Social Influence Analysis
- Cohesive Group (Community) Analysis
- Hidden Link Prediction
- Network Embedding
- Network Sparsification

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

3. Lawfully intercepted data

The activities towards collecting lawfully intercepted data in ROXANNE falls within the scope of T4.1 with the following description [1]:

> **T4.1 Inventory and analysis of lawfully intercepted data [M2-M12]**
> **Leader:** SAIL, **Participants**: UCSC, AEGIS, ADITESS, PHO, INTERPOL, all LEAs
>
> *This task will involve all LEA partners, and will consider all legal and ethical aspects which will arise w.r.t. using real (sensitive) data in the project. It will also prepare a list of data and their potential use in the project according to the classified sensitivity levels: (1) data for project R&D activities available only on secured LEA premises; (2) data for project R&D activities which can be accessed through either a secured collaborative platform, or through access to the ROXANNE remote platform on KEMEA premises; (3) data for project R&D activities with significantly less constraints on their use within the project (specifically targeted in T4.3). The task activities will be in detail supervised by the ethics board and security advisory board.*

As already discussed in Section 2.1, the data obtained in this task would be the "real" datasets, which would be used to adapt the technologies for the criminal domain and to evaluate the final ROXANNE platform. Nevertheless, most of the LEA partners have been unable to provide any inputs up to the submission of this deliverable. Due to the sensitivity of such data and the legal restrictions, these organizations are not entitled to share whole or part of their data with other consortium partners. Some partners are still considering which data they own could safely be shared with the consortium whereas some others are waiting for consent from their management. Some partners have denoted that they would require a court order to be able to make data available for research, which seems to be unlikely at this point. Several solutions are being pursued to alleviate this problem.

ROXANNE consortium will attend the workshop organised by EC, focusing on the issue of data (January 2020). Several possible solutions will be presented by ROXANNE in this workshop and discussed with others to find a consensus on legal, ethical, but also functional questions on data collection. Below, we give a short introduction to the proposed points:

(a) recording synthetic data (i.e. data provided by volunteers to simulate real criminal cases) through a real environment (e.g. a HW used by LEA to record all telephone calls). The reader should be notified that efforts towards collecting more of such sources will continue throughout the timeline of this task (until M12). As a fallback scenario to the case that no additional LEA data could be obtained, the ROXANNE consortium have already started the activities in T4.6: Target data simulation for development and demonstration activities (which was supposed to start at M6). The idea is to prepare a new dataset that simulates a criminal network activity, by making it as "realistic" dataset as possible. A data collection methodology document has already been prepared and proposed to the consortium by BUT, which contains information on how to collect such dataset. The plan is have the ROXANNE researchers role-play criminals and innocent people, by calling and interacting with each other, similar to a criminal network activity. All details regarding preparation and implementation of this dataset will be covered in D4.2: Simulated data for development and demonstration (due M16).

(b) setting up a self-contained environment (without a network connection) on LEA premises where the data can be processed. The output would be represented by aggregated data (and to be cecked for a presence of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

personnal data). The solution would alleviate risks on LEA sides and may clearly answer all the legal requirements.

(c) Researchers and other partners from technical organisations would work directly on the LEA premises. The data would be fully under the control of LEA and the security clearance on an individual level could be organized.

(d) Installing the whole ROXANNE solution on LEA server(s) without the need for technical partners to be physically present for using the SW.

(e) Relatively small datasets, fully anonimised, which can already be shared by some of LEA partners: In the following paragraphs we provide information about the two datasets which have already been shared:

FRIDA (provided by NFI):

The FRIDA dataset contains telephone conversations of over 200 speakers in Dutch language. There are a total of 16 unique recordings per speaker, simultaneously recorded by multiple recording devices, making up a total of 72 files per speaker. The average duration of recordings is 6 minutes and the sides of the telephone conversation are available as two separate files. 8 recordings of each speaker are transcribed in the utterance level, with start and end times marked. This dataset is designed as a substitute for lawfully intercepted telephone conversations, for use in R&D projects. The speakers are not criminals, but people "from the street", who are recruited for the construction of this dataset.

BALSAS_200LT (provided by LTEC):

The LTEC's voice database BALSAS_200LT contains intercepted GSM telephone conversations from real cases in Lithuanian language. It consists of three parts: (i) 200 mono recordings (conversations) between two speakers, (ii) 203 manually-segmented recordings having only a single speaker (side) in each, (iii) 10 recordings of 5 known speakers. The average duration of unsegmented recordings are about 50 seconds, whereas the average duration of segmented recordings are about 20 seconds. All the data are anonymized.

4. Publicly available data

The activities towards collecting publicly accessible data in ROXANNE falls within the scope of T4.2 with the following description [1]:

> **T4.2 Overview of publicly available data resources for research, training and development [M2-M12]**
> **Leader:** SAIL, **Participants:** ADITESS, ITML, BUT, PHO, AEGIS, USAAR
>
> *This task will provide a survey of existing publicly available resources, mitigating the unavailability of LEA data for technology development. Focus will be on: (i) Commercial databases as the ones available from Linguistic Data Consortium (LDC). (ii) Multimedia data from which relations can be inferred and where a significant amount of meta information is available. (iii) Publicly available data - YouTube, Vimeo, etc., allowing tests of the developed approaches on data that was not used in training. Publicly available data might as well complement the investigation data (T4.1).*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Per definition, T4.2 covers not only the second data source (publicly available data) but also partly the third data source (data already available to partners) that was introduced in Section 2.2. Indeed, the difference between the second and third data sources is a vague one, and was only categorized so as to distinguish between the actual ownership of data (outside sources or anonymous data which need to be acquired vs. data which the consortium partners already own or will create). At the time of the preparation of this document, 6 technical partners (including those who are not participating in this task) have provided information on the datasets they own and have access to. In the following paragraphs we explain some of these in detail:

AMI Meeting Corpus

The AMI Meeting Corpus [3] is a multi-modal data set consisting of 100 hours of meeting recordings. Is was collected as part of the European-funded AMI project (FP6-506811) by a 15-member multi-disciplinary consortium. Around two-thirds of the data has been elicited using a scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The rest consists of naturally occurring meetings in a range of domains. The recordings use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants also have unsynchronized pens available to them that record what is written. The meetings were recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers. ROXANNE may benefit from the AMI Meeting Corpus in training and evaluating speech recognition, speaker recognition and face recognition systems.

IARPA Babel Corpora

The Babel Program [4] was a research action funded by IARPA in order to develop agile and robust speech recognition technology that can be rapidly applied to any human language in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech. It required innovations in how to rapidly model a novel language with significantly less training data that are also much noisier and more heterogeneous than what has been used in the current state-of-the-art. Telephone conversations of varying data quantity over 26 under resourced languages were collected as part of the project. The datasets are distributed by LDC. ROXANNE may benefit from the IARPA Babel Corpora in training and evaluating speech recognition and keyword search systems.

CoNLL-2002 and CoNLL-2003 Language-Independent Named Entity Recognition

CoNLL, the Conference on Computational Natural Language Learning, is a top-tier conference, yearly organized by SIGNLL (ACL's Special Interest Group on Natural Language Learning). Every year the conference organizes a shared task for researchers to collaborate on a common subject. The shared tasks of CoNLL-2002 [5] and CoNLL-2003 [6] concerned language-independent named entity recognition and the aim was to detect four types of named entities: persons, locations, organizations and names of miscellaneous entities. The CoNLL-2002 data consist of files covering Spanish and Dutch, and the CoNLL-2003 data consists of files covering English and German. The Spanish data is a collection of newswire articles from May 2000, made available by the Spanish EFE News Agency. The Dutch data consist of four editions of the Belgian newspaper "De Morgen" of 2000. The English data was taken from the Reuters Corpus with news stories between August 1996 and August 1997. The text for the German data was taken from the ECI Multilingual Text Corpus with news stories from the German newspaper Frankfurter Rundshau. ROXANNE may benefit from these datasets in training and evaluating named entity recognition systems.

KONECT Corpora

KONECT (the Koblenz Network Collection) [7] is a project to collect large network datasets of all types in order to perform research in network science and related fields, collected by the Institute of Web Science and Technologies at the University of Koblenz–Landau. KONECT contains several hundred network datasets of various types, including directed, undirected, bipartite, weighted, unweighted, signed and rating networks. The networks of KONECT cover many diverse areas such as social networks, hyperlink networks, authorship networks, physical networks, interaction

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

networks, and communication networks. ROXANNE may benefit from these corpora in developing network analysis techniques and systems. Some particular datasets in KONECT which may be of particular use in ROXANNE are the Actor Collaborations Dataset [8] (actors connected by an edge if they both appeared in the same movie), the Twitter dataset [9] (the follower network from Twitter, containing 1.4 billion directed follow edges between 41 million Twitter users), the Enron email network dataset [10] (1,148,072 emails sent between employees of Enron between 1999 and 2003), and the St. Louis Crime dataset [11] (870 indviduals involved in 557 crime events in St. Louis in 1990s, as victim or/and suspect)".

Speakers In the Wild Corpus

The Speakers in the Wild (SITW) [12] was a speaker recognition challenge held by SRI for Interspeech 2016. The SITW database contains hand annotated speech samples from open source media for the purpose of benchmarking speaker recognition technology on single and multi-speaker audio acquired across unconstrained or 'wild' conditions. The conditions represented in the SITW database provide samples of nearly 300 individuals across clean interview, red carpet interviews, stadium conditions, outdoor conditions, and multi-speaker scenarios. Each individual also has speech acquired using camcorders or cellphones and void of professional editing. All noise, reverb, compression and other artifacts in the corpus are natural characteristics of the original audio. ROXANNE may benefit from this corpus in developing robust speech and speaker recognition systems against channel effects and noise.

More information on all of the collected corpora can be found in the Appendix of this document.

5. Other topics and future work related to data

This deliverable intends to inform the reader about the datasets currently available to the ROXANNE consortium for developing and evaluating the ROXANNE platform. It does not touch upon how (parts of) these datasets will be selected. The procedure of establishing an appropriate subset, considering all data types and technologies is still under development and will be discussed in the upcoming data-related deliverables.The main topics related to data selection and usage in ROXANNE can be summarized as follows:

Languages: The consortium has not yet decided on the final set of languages. The first field test is expected to cover at least the English and German languages. As it has been already indicated in the Grant Agreement, Part B Section 1.3.5 Languages, the following languages which we expect to support are Dutch, Spanish, Czech, Greek, Hebrew, and Arabic. A questionnaire is currently being prepared to gather end-users' requirements. We aim to be flexible in selecting the languages to be processed, as many of the project partners have experience in adapting speech and language technologies for a large number of languages.

Social media platforms and content: The questionnaire for end-users also contain questions regarding the social media platforms and the types of content within those platforms. For instance, a Facebook page can contain posts, embedded audio/video content, links, replies, likes, etc. Similarly, A YouTube page would contain video, description of the video, comments of the video, etc. Knowing the exact content our end-users are interested in will help us shape our components and build better models that make use of such data.

Availability of ground-truth: The datasets with which the technological components are trained already have ground-truth information. For the datasets with which the ROXANNE platform will be evaluated, we also aim to obtain such ground-truth information. Availability of ground-truth, especially from the closed criminal cases, will be a plus in the project, as it will allow the comparison of results/features/traits obtained from the processing performed by ROXANNE with the results obtained by investigators working on the case (e.g. an identity or connection obtained by ROXANNE as compared to the same fact as obtained by human investigative efforts).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Personal data and anonymisation: The consortium prefers that the data will not be anonymised wherever possible, since not only the acoustic information, but also the content information will be systematically used to extract different traits (e.g. names, abbreviations, relations among different words, etc.) of an individual. Other privacy preserving techniques will be deployed in order to ensure full respect of relevant legal and ethical requirements, which will be monitored as part of WP3 and WP10.

Processing of real data within the project: The project relies on real operational data to be available for evaluating the technology. The lack of data can generate problems of uncertain information (e.g. from the statistical point of view). On the other hand, the consortium is aware that data gathered from lawful interceptions and other investigative sources may be very sensitive in nature. This data shall remain confidential and cannot be shared with other consortium partners. The consortium will therefore make sure that such data stays all the time within the LEAs premises. Access to data within the LEAs premises may be granted to representatives from other consortium partners in accordance with LEA's rules, and only as an exception. In general, such real data can be used in two stages:

- *To enhance/adapt the developed models:* The simulated data provided by the LEAs or collected by the consortium will resemble the data available in investigative work. This can be achieved by collecting a set of telephone or other communications, recorded for instance by LEAs under similar conditions as in real cases. With their consent, this data then can be used for developing or adapting the technology.
- *To evaluate the ROXANNE platform:* This will be done directly by the LEAs. The software will be installed locally, on their equipment and will be evaluated on real data by a closed group comprising of LEAs representatives and only exceptionally limited number of technical partners (i.e. continuous testing activities). In order to preserve privacy and personal data of individuals, once the evaluation is finished, the LEAs involved will produce a report and share their findings with technical partners. Their feedback will not include any personal data and will focus on the technical aspects of the ROXANNE technology.

Open research data pilot: ROXANNE opts out from the Open Research Data Pilot for the following reasons:

- The most important data used for validation of the technology comes from real operational scenarios of LEAs and is very sensitive. Even within the consortium, accessing such data for research partners will mostly be not possible, and the field tests will be conducted at the premises of LEAs and on their hardware.
- The simulated data are less sensitive, as they will be recorded by the consortium members, with appropriate consent forms signed. These data, however, will contain simulated criminal cases (i.e. scenarios will be drafted by LEAs), and as such, will partially disclose the proceedings of criminal investigations. As such, they are potentially dangerous if misused by the criminals, and should not be distributed completely publicly. The consortium plans to share this data with proven LEA partners both in Europe and elsewhere.

Bibliography

1. "Grant Agreement number: 833635 — ROXANNE — H2020-SU-SEC-2018-2019-2020/H2020-SU-SEC-2018", 2019.
2. "ROXANNE D10.17: Requirement No.20", edited by TRI, v1.0, 2019.
3. J. Carletta, "Announcing the AMI Meeting Corpus", *The ELRA Newsletter 11(1)*, January-March, p. 3-5, 2006.
4. M.J.F. Gales, K.M. Knill, A. Ragni, "Low-Resource Speech Recognition and Keyword-Spotting", SPECOM 2017 Lecture Notes in Computer Science, vol 10458. Springer, Cham, 2017.
5. Erik F. Tjong Kim Sang, "Introduction to the CoNLL-2002 shared task: language-independent named entity recognition", *Proceedings of CONLL '02*, Taipei, Taiwan, 155–158, 2002.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

6. Erik F. Tjong Kim Sang and Fien De Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition", *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*, Association for Computational Linguistics, USA, 142–147, 2003.
7. Jérôme Kunegis. "KONECT - The Koblenz Network Collection", *Proc. Int. Conf. on World Wide Web Companion*, pp. 1343-1350, 2013.
8. Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks", *Science*, 286(5439):509-512, 1999.
9. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is Twitter, a social network or a news media?", *Proc. Int. World Wide Web Conf.*, pp. 591-600, 2010.
10. Bryan Klimt and Yiming Yang, "The Enron corpus: A new dataset for email classification research", *Proc. European Conf. on Machine Learning*, pp. 217-226, 2004.
11. Crime network dataset -- KONECT, April 2017.
12. M. McLaren, L. Ferrer, D. Castan and A. Lawson, "The 2016 Speakers in the Wild Speaker Recognition Evaluation," *Proc. INTERSPEECH 2016*, pp. 823-827, September 2016.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

Appendix

Partner	Dataset Name	Data Type	Year Created	Language	Amount of Data	Quality of Data	Notes/Comments	License Type	Application Domain
IDIAP, PHO	AMI Meeting Corpus	audio/speech video xml tags	< 2010	English, many accents	about 100 hours	close-talk vs. far-field microphones	Around two-thirds of the data has been elicited using a scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The rest consists of naturally occurring meetings in a range of domains. Detailed information can be found in the documentation section.	Creative Commons Attribution 4.0 International Licence (CC BY 4.0).	speech recognition, speaker recognition, face recognition
BUT	IARPA Babel	audio/speech transcriptions	2011 >	26 languages	varies with respect to language	telephone		LDC License	speech recognition keyword search
USAAR, NFI	CoNLL-2002 CoNLL-2003 Language-Independent Named Entity Recognition	text	2002 2003	Spanish, Dutch, English, German	273,037 Spanish tokens 218,737 Dutch tokens 22,137 English sentences 18,933 German sentences	high	This is a commonly used dataset for named entity recognition tasks	LDC License	named entity recognition
LUH	Cornell movie dialog corpus	text	2011	English	220,579 conversational exchanges between 10,292 pairs of movie characters; involves 9,035 characters from 617 movies		Transcription of dialogs in movies	Publicly and freely avaiable for research purposes	linguistics, social network analysis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

PHO	Fearless steps NASA	audio/speech	2019	English	91 hours	low quality recordings in wav	wide range of recording devices	CC0	voice activity detection, diarization, speech recognition, speaker recognition, sentiment analysis
SAIL	Forensic database of voice recordings of 500+ Australian English speakers	audio/speech	2015	Australian English	This database currently contains recordings of 552 Australian English speakers. female male speakers 322 231 recorded on 1 session 90 62 recorded on 2 sessions 72 43 recorded on 3 sessions 155 107 recorded on 4+ sessions 5 19	landline + close-talking	On each occasion, each speaker was recorded in three speaking styles (tasks): casual telephone conversation (cnv), information exchange task over the telephone (fax), pseudo-police-style interview (int).	Freely available for non-commercial research and forensic casework	speaker recognition accent identification
SAIL	Forensic Voice Comparison Databases forensic_eval_01	audio/speech	< 2016				A set of training and test data representative of the relevant population and reflecting the conditions of an actual forensic voice comparison case, and operational forensic laboratories and research laboratories are invited to use these data to train and test their systems.	Freely available for research	speaker recognition
LUH	KONECT	social networks	2004-constantly updated	Several languages	Datasets of various sizes, from tens to hundreds millions nodes, tens to billions edges.			Publicly and freely available for research purposes	Social network analysis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

BUT	LDC data	audio/speech text	2004-2019	Several languages	several TB of data, see details in LDC catalogue for detailed sizes	various, from telephone to broadcast	BUT is a member of LDC since 2004, which allows us for free download of a dozen of LDC-created corpora every year. Separate licenses were purchased for important corpora published prior to 2004, such as Switchboard, Fisher, etc.	Varying depending on corpus, BUT can use the data for academic R&D	speech recognition, speaker recognition, language modeling
SAIL, PHO	LibriSpeech	audio/speech	2015	English	Approximately 1000 hours	16kHz read speech, close talking	LibriSpeech corpus is prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.	Creative Commons Attribution 4.0 International Licence (CC BY 4.0).	speech recognition
PHO	Mozilla Common Voice Dataset	audio	2019	Several languages	each language has a different size	PC recordings in mp3		CC0	speech recognition, voice activity detection
USAAR	Named Entity Recognition in Estonian	text	2013	Estonian data	572 news stories 184,638 tokens			Publicly available	Name-entity extraction
SAIL, PHO	Speakers in the Wild	audio/speech	2016	English	The database consists of recordings of 299 speakers, with an average of eight different sessions per person.		The SITW speech data was collected from open-source media channels in which 299 well-known public figures, or persons of interest (POI), were speaking. Specifically, the data have considerable mismatch in audio conditions as they were acquired both from high quality studio-based interviews and from raw audio captured on, for example, a camcorder.Duration of speech for each speaker is unconstrained, as are the audio conditions. All noise, reverb, vocal effort, and other acoustic artifacts in the corpus are natural characteristics of the original audio. Speaking conditions include monologues, interviews, and more conversational dialogues with dominant backchannel and speaker overlap.	SRI License	speaker recognition

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.

LUH	SNAP	social networks	2004 >	Several languages	Datasets of various sizes, from tens to hundreds millions nodes, tens to billions edges.			Publicly and freely avaiable for research purposes	social network analysis
NFI	SoNaR	text	2015	Dutch	A 500M word corpus, of which 1M is hand annotated, including named entities. It includes a number of genres.		Data collected as part of the project between 2008-2011.	Publicly and freely available.	named entity recognition
SAIL	Spanish Speaker Verification Corpus	audio/speech	2016	Spanish	54 native Spanish speakers. Every speaker uttered six different sentences and there at least 10 repetitions for all but 2 speakers.			Freely available for research	speaker recognition
SAIL	The Spoken Wikipedia Corpora	audio/speech	2017	English, German, Dutch	German English Dutch #articles 1010 1314 3073 #speakers 339 395 145 total audio 386h 395h 224h aligned words 249h 182h 79h phonetic aligned 129h 77h —			CC BY-SA 4.0	speech recognition
SAIL, BUT, PHO	VoxCeleb	audio/speech video metadata	2017	English, many accents from 145 nationalities, but not annotated.	# of POIs 1,251 (690 m / 561 f) # of videos 22,496 # of utterances 153,516 # of hours 352 (8.2 sec per utterance) Avg # of videos/utterances per POI 18/116	Youtube data, mostly professional recordings	The audio/speech part also contains speaker labels, and the video part also contains cropped face videos. Approximately 7000 speakers, 1000000 segments, 2000h. Contains some overlap between the speakers in this dataset and SITW.	Creative Commons Attribution 4.0 International License	speaker recognition accent identification speech separation face recognition
SAIL, PHO	VoxCeleb2	audio/speech video	2018	English, many accents	# of POIs 6,112 (3,761 m / 2,351 f) # of videos 150,480 # of utterances 1,128,246 # of hours 2,442 (7.8 sec per utterance) Avg # of videos/utterances per POI 25/185			Creative Commons Attribution 4.0 International License	speaker recognition, face recognition

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833635.