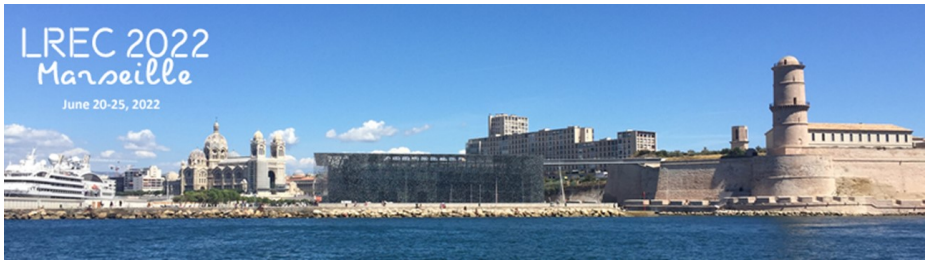

LREC 2022 Joint Workshop
Language Resources and Evaluation Conference
20-25 June 2022

Legal and Ethical Issues In Human Language Technologies and Multilingual De-Identification of Sensitive Language Resources
(LEGAL - MDLR)

PROCEEDINGS

Editors: Mickaël Rigault, Victoria Arranz, Ingo Siegert

Proceedings of the LREC 2022 Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Language Resources (LEGAL - MDLR 2022)

Edited by:
Mickaël Rigault, Victoria Arranz, Ingo Siegert

ISBN: 979-10-95546-96-2
EAN: 9791095546962

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface

The legal framework affecting access to and re-use of language data in the European Union has evolved very significantly since the last LREC conference (7-12 May 2018). The main objective of the workshop is to discuss the major issues around legal and related technological directions of Human Language Technologies.

The workshop is meant to study different interactions between legal and technical aspects of data collection, processing, and distribution. Such interactions may concern text crawling, speech and voice recordings and the impact of the text and speech data mining exception introduced by the European legislation in 2018. These interactions may also concern the compatibility of the legal requirement for (text, audio, video) data collection and their processing as imposed by the GDPR, together with the technical feasibility of the different anonymisation and pseudonymisation techniques. This workshop looks into the various approaches to effective and reliable text de-identification, focusing on some particularly sensitive domains such as the medical and legal domains, but not only.

This workshop also aims to discuss larger issues such as ethics and morality, as well as trust and their interactions as a whole on data collection and distribution and how they may be inserted into binding legal instruments (code of ethics, best practices). The purpose of this workshop will attempt to build bridges between technology and legal experts and discuss current legal and ethical issues in the Human Language Technology sector. This will be addressed by bringing together researchers and scholars working on Intellectual Property, Public Sector Information, Personal Data and possibly ethics, both from the legal and technical perspectives.

This volume documents the Proceedings of the LREC Joint Workshop on Legal and Ethical Issues In Human Language Technologies and Multilingual De-Identification of Sensitive Language Resources, held on Friday, June 24, 2022, as part of the LREC 2022 Conference (International Conference on Language Resources and Evaluation).

We would like to thank our keynote speakers for their enlightening speeches, as well as the authors who contributed to this workshop with their papers and discussions. We are also very grateful to the members of the Program Committee for the time and effort devoted to the reviewing of the papers.

LEGAL Organizing Committee

Ingo Siegert – Otto von Guericke University Magdeburg (GERMANY)
Mickaël Rigault – ELDA/ELRA (FRANCE)
Khalid Choukri – ELDA/ELRA (FRANCE)
Pawel Kamocki – IDS Mannheim (GERMANY)
Andreas Witt – IDS Mannheim (GERMANY)
Krister Linden – University of Helsinki (FINLAND)
Claudia Cevenini – University of Bologna (ITALY)

MDLR Organizing Committee

Victoria Arranz – ELDA/ELRA (FRANCE)
Montse Cuadros – Vicomtech Foundation (SPAIN)
Aitor García Pablos – Vicomtech Foundation (SPAIN)
Cyril Grouin – Université Paris-Saclay, CNRS, LISN (FRANCE)
Manuel Herranz – Pangeanic (SPAIN)

Program Committee:

Khalid Choukri, ELDA/ELRA (FRANCE)
Hercules Dalianis, Stockholm University (SWEDEN)
Amando Estela, Pangeanic (SPAIN)
Thierry Etchegoyhen, Vicomtech Foundation (SPAIN)
Albert Gatt, Malta University (MALTA)
Lucie Gianola, Université Paris-Saclay, CNRS, LISN (FRANCE)
Ona de Gibert, BSC (SPAIN)
Marwa Hadj Salah, ELDA/ELRA (FRANCE)
Udo Hahn, University of Jena (GERMANY)
Thomas Kleinbauer, COMPRISE Project (GERMANY)
Maite Melero, BSC (SPAIN)
Patrick Paroubek, Université Paris-Saclay, CNRS, LISN (FRANCE)
Naiara Perez, Vicomtech Foundation (SPAIN)
Stelios Piperidis, Athena Research and Innovation Center (GREECE)
Prokopis Prokopidis, Athena Research and Innovation Center (GREECE)
Mike Rosner, Malta University (MALTA)
Roberts Rozis, TILDE (LATVIA)
Özlem Uzuner, George Mason University (U.S.A.)
Emmanuel Vincent, Inria Nancy-Grand Est (FRANCE)
Rinalds Vīksna , TILDE (LATVIA)
Pierre Zweigenbaum, Université Paris-Saclay, CNRS, LISN (FRANCE)

Table of Contents

Conference Program

Friday, June 24, 2022

09:00–09:15 *Welcome and Introduction*
Ingo Siegert, Victoria Arranz

09:15–10:10 Keynote Speech - Major Developments in the Legal Framework Concerning Language Resources
Pawel Kamocki, IDS Mannheim

10:10–10:30 Session A: COVID Issues and Policy Amendments

10:10–10:30 *Sentiment Analysis and Topic Modeling for Public Perceptions of Air Travel: COVID Issues and Policy Amendments*
Avery Field, Aparna Varde and Pankaj Lal

11:00–12:00 Session B: GDPR and Legal Aspects

11:00–11:20 *Data Protection, Privacy and US Regulation*
Denise DiPersio

11:20–11:40 *Pseudonymisation of Speech Data as an Alternative Approach to GDPR Compliance*
Pawel Kamocki and Ingo Siegert

11:40–12:00 *Categorizing Legal Features in a Metadata-Oriented Task: Defining the Conditions of Use*
Mickaël Rigault, Victoria Arranz, Valérie Mapelli, Penny Labropoulou and Stelios Piperidis

Friday, June 24, 2022 (continued)

12:00–13:00 Session C1: Data Protection: Anonymisation, De-Identification and Legal Aspects

12:00–12:20 *About Migration Flows and Sentiment Analysis on Twitter data: Building the Bridge between Technical and Legal Approaches to Data Protection*
Thilo Gottschalk and Francesca Pichierri

12:20–12:40 *Transparency and Explainability of a Machine Learning Model in the Context of Human Resource Management*
Sebastien Delecraz, Loukman Eltarr and Olivier Oullier

12:40–13:00 *Public Interactions with Voice Assistant – Discussion of Different One-Shot Solutions to Preserve Speaker Privacy*
Ingo Siegert, Yamini Sinha, Gino Winkelmann, Oliver Jokisch and Andreas Wendemuth

14:00–15:00 Keynote Speech - Anonymisation and the GDPR
Brij Mohan Lal Srivastava, Co-Founder of Nijta Startup Studio, Lille

15:00–16:00 Session C2: Data Protection: Anonymisation in Practice

15:00–15:20 *Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats*
Olle Bridal, Thomas Vakili and Marina Santini

15:20–15:40 *Generating Realistic Synthetic Curricula Vitae for Machine Learning Applications under Differential Privacy*
Andrea Bruera, Francesco Aldà and Francesco Di Cerbo

15:40–16:00 *MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents*
Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek and Pierre Zweigenbaum

Friday, June 24, 2022 (continued)

16:30–17:30 Session D: Privacy and Ethical Challenges in Data

16:30–16:50 *PriPA: A Tool for Privacy-Preserving Analytics of Linguistic Data*
Jeremie Clos, Emma McClaughlin, Pepita Barnard, Elena Nichele, Dawn Knight, Derek McAuley and Svenja Adolphs

16:50–17:10 *Legal and Ethical Challenges in Recording Air Traffic Control Speech*
Mickaël Rigault, Claudia Cevenini, Khalid Choukri, Martin Kocour, Karel Veselý, Igor Szoke, Petr Motlicek, Juan Pablo Zuluaga-Gomez, Alexander Blatt, Dietrich Klakow, Allan Tart, Pavel Kolčárek and Jan Černocký

17:10–17:30 *It is not Dance, is Data: Gearing Ethical Circulation of Intangible Cultural Heritage Practices in the Digital Space*
Jorge Yánez and Amel Fraisse

17:30–18:00 ***Closing Ceremony***