LREC 2026

Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026

Workshop Proceedings

Editors
Ingo Siegert, Maria Irena Szawerna, Khalid Choukri, Simon Dobnik, Paweł Kamocki, Therese Lindström Tiedemann, Pierre Lison, Ricardo Muñoz Sánchez, Ildikó Pilán, Lisa Södergård, Kossay Talmoudi, Elena Volodina, Xuan-Son Vu

12 May 2026

Proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026

ISBN 978-2-493814-86-9

Preface

In recent years, Artificial Intelligence, and in particular Large Language Models (LLMs), have moved from promising innovations to transformative forces shaping research, industry, and society at large. This rapid integration has intensified discussions not only around technological capabilities, but also around the governance of data, data protection, authorship, and accountability on a global scale. As these systems become foundational infrastructures, the legal and ethical implications of their development and deployment have become more complex and more urgent across jurisdictions.

Against this backdrop, a joint workshop is organized at LREC 2026, bringing together two previously distinct but highly complementary venues: the Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2026) and the Workshop on Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (CALD-pseudo 2026). By merging these two workshops, this edition explicitly fosters a deeper integration of legal, ethical, and technical perspectives on data governance and privacy in language technologies.

Reconciling innovation with compliance remains a challenge in an increasingly diverse regulatory landscape. While approaches may differ in structure and enforcement, the data protection frameworks of different world regions remain a point of vigilance for the advancement of AI. Japan's Act on the Protection of Personal Information (APPI), South Korea's Personal Information Protection Act (PIPA), Brazil's Lei Geral de Proteção de Dados Pessoais (LGPD), and state-level instruments in the United States such as the California Consumer Privacy Act (CCPA) and the California Privacy Rights Act (CPRA) all reflect this trend. Similarly, countries such as Morocco, through Law 09-08, demonstrate that the establishment of legal frameworks governing personal data is now a widespread and shared priority, even as implementation models and levels of maturity vary.

Within this context, questions surrounding de-identification, pseudonymization, and the assessment of re-identification risk are no longer theoretical concerns but concrete technical and legal challenges, particularly in sensitive domains such as clinical data. Ensuring the accessibility of research data while protecting personal and sensitive information remains a central challenge, requiring approaches that effectively conceal identities while preserving usability and semantic value for downstream research tasks. The detection of indirect personal identifiers and the behavior of embedding representations with respect to personal data further illustrate the complexity of safeguarding privacy in modern AI systems. At the same time, new approaches seek to leverage LLMs themselves to assess privacy sensitivity and to support more robust evaluation of privacy-aware systems, while also accounting for potential biases introduced through pseudonymization processes.

Beyond personal data, the rise of generative and adaptive AI systems also brings to the forefront complex questions of intellectual property rights, liability, and transparency. Issues of authorship, ownership, and derivative works now intersect with algorithmic generation, challenging traditional notions of originality and rights allocation. In response, jurisdictions are progressively enacting AI-specific legal instruments, most notably the European Union's AI Act, which sets harmonized rules for trustworthy and accountable AI.

The increasing use of synthetic data for AI development and fine-tuning introduces additional opportunities, but also raises important legal and ethical questions regarding data provenance, risk mitigation, and compliance. These developments highlight a broader shift from data protection as a constraint toward data governance as an enabling framework for responsible innovation. More broadly, the need to preserve utility and semantic integrity while limiting the

exposure of identifiable entities reflects an ongoing tension between innovation and protection, particularly in light of adversarial re-identification risks.

Held as part of the LREC 2026 Conference, this joint workshop provides a forum for researchers, legal experts, and practitioners to explore the evolving relationship between language technologies and regulatory frameworks worldwide. By explicitly combining the LEGAL and CALD-pseudo communities, the workshop strengthens interdisciplinary exchange and enables a more comprehensive discussion of pseudonymization, data protection, and privacy-preserving technologies.

This year's program reflects a strong focus on privacy-aware methodologies, value-preserving data processing, and compliance-by-design approaches, including the integration of legal requirements into user-oriented infrastructures and consent management tools.

The workshop also addresses emerging concerns around authorship attribution and intellectual responsibility in the age of generative AI, as well as broader governance challenges related to data sharing, contractual frameworks, and the interoperability of infrastructures. The balance between FAIR principles and data protection requirements remains a key issue, particularly in the context of oral and cultural archives.

A notable theme this year is the growing importance of transparency and accountability as system-level properties. The identification of structural compliance gaps and the operationalization of regulatory requirements highlight the need to embed legal considerations directly into the architecture of AI systems.

Ultimately, the workshop underscores the need for a holistic and international approach that integrates technical innovation with legal and ethical responsibility. By fostering dialogue across disciplines and jurisdictions, it aims to contribute to the development of trustworthy, compliant, and socially responsible language technologies in a global context.

This volume presents the proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies (LEGAL2026) and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (CALD-pseudo 2026), co-located with LREC 2026. We would like to express our sincere gratitude to all authors for their valuable contributions, and to the Program Committee for their careful and dedicated reviews. Our further thanks go to the generous support from the Swedish Research Council through its funding to the research environment project *Grandma Karl is 27 years old*.

Workshop Organizers, April 2026

Organizing Committee

- **LEGAL 2026:**
  - Ingo Siegert, Otto-von-Guericke Universität Magdeburg, Germany
  - Paweł Kamocki, Leibniz-Institut für Deutsche Sprache, Germany
  - Kossay Talmoudi, ELDA, France
  - Khalid Choukri, ELDA, France
- **CALD-pseudo 2026:**
  - Maria Irena Szawerna, University of Gothenburg, Sweden
  - Simon Dobnik, University of Gothenburg, Sweden
  - Therese Lindström Tiedemann, University of Helsinki, Finland
  - Pierre Lison, Norwegian Computing Center & University of Oslo, Norway
  - Ildikó Pilán, Norwegian Computing Center, Norway
  - Ricardo Muñoz Sánchez, University of Gothenburg, Sweden
  - Lisa Södergård, University of Helsinki, Finland
  - Elena Volodina, University of Gothenburg, Sweden
  - Xuan-Son Vu, Lund University & DeepTensor AB, Sweden

Table of Contents

Workshop Program

Tuesday, May 12, 2026

9:00–10:10 **Welcome Session**

9:00–9:15 *Welcome and basic information from the organizers*
Workshop Organizers

9:15–10:10 *Introductory Lecture*
Paweł Kamocki

10:10–10:30 **Oral Presentations I: LEGAL**

10:10–10:30 *Transparency as Architecture: Structural Compliance Gaps in EU AI Act Article 50 II*
Vera Schmitt, Niklas Kruse, Premtim Sahitaj and Julius Schöning

10:30–11:00 **Coffee Break**

11:00–11:55 **Keynote**

11:00–11:55 *Keynote Speech*
Maja Bogataj Jančič

11:55–13:00 **Oral Presentations II: LEGAL**

11:55–12:15 *Towards Robust Evaluation for Privacy QA Systems*
Anna Leschanowsky, Zahra Kolagar, Erion Çano, Ivan Habernal, Dara Hallinan, Emanuël Habets and Birgit Popp

12:15–12:35 *LDS Contractual Framework: Principles, Status and Implementation*
Penny Labropoulou, Kossay Talmoudi, Dimitrios Gkoumas, Katerina Gkirtzou, Miltos Deligiannis, Leon Voukoutis, Athanasia Kolovou, Khalid Choukri, Stelios Piperidis and Dimitrios Galanis

12:35–12:55 *Authorship Attribution in the Times of LLMs within the Framework of the CRediT Taxonomy*
Pawel Kamocki and Andreas Witt

Tuesday, May 12, 2026 (continued)

13:00–14:00 **Lunch Break**

14:00–14:55 **Invited Talk**

14:00–14:55 *Privacy and anonymization in NLP: Are we barking up the wrong tree?*
Prof. Dr. Ivan Habernal

14:55–16:00 **Oral Presentations III: CALD-pseudo**

14:55–15:15 *DeID-Clinic: A Risk-Aware Pseudonymization Framework for Clinical Text De-identification and Re-identification Risk Assessment*
Angel Paul, Dhivin Shaji, Lifeng Han, Warren Del-Pinto, Goran Nenadic and Suzan Verberne

15:15–15:35 *Distilling Human-Aligned Privacy Sensitivity Assessment from Large Language Models*
Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer and Marc Tommasi

15:35–15:55 *Birds of a Feather: Do Embedding Representations of Personal Information Flock Together?*
Maria Irena Szawerna and Simon Dobnik

16:00–16:30 **Coffee Break**

16:30–17:45 **Poster Session**

16:30–17:45 *Modelling Legal Compliance in a Consent Wizard Application as Part of a Research-Centered and User-Oriented Data Infrastructure*
Aliena Strathmann, Marc-Levin Joppek, Maryam Mohammadi, Katja Politt, Paul T. Schrader, Annett B. Jorschick and Hendrik Buschmeier

16:30–17:45 *Balancing FAIR and GDPR: A Governance Framework for Oral Archives*
Elvira Mercatanti, Monica Monachini, Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Virginia Niri, Cesarina Vecchia, Giulia Zitelli Conti and Giada Zuccolo

16:30–17:45 *Legal Considerations in the Use of Synthetic Data for AI Development and Finetuning: The Case of LLMs4EU*
Kossay Talmoudi, Khalid Choukri, Amélie Gourgeot and Florine Astruc

Tuesday, May 12, 2026 (continued)

16:30–17:45 *Evaluating Encoder- and LLM-Based Approaches for Robust Indirect Personal Identifier Detection*
Christoph Otto, Ibrahim Baroud, Akiko Aizawa, Sebastian Möller, Roland Roller and Lisa Raithel

16:30–17:45 *VEIL: A Benchmark for Value-Preserving Entity Identification Limitation*
Darina Gold, Shadi Rastegar, Alina Liebel and Alessandra Zarcone

17:45–18:00 Closing Session

17:45–18:00 *Closing Remarks*
Workshop Organizers