LREC 2026

Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026

Workshop Proceedings

Editors
Florian Barth, Keli Du, José Calvo Tello, Philippe Genêt, Piroska Lendvai, Christof Schöch, Thorsten Trippel

12 May 2026

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026

ISBN 978-2-493814-77-7

Preface

We are pleased to present the *Book of Abstracts* for the workshop **"Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science"**, held in conjunction with **LREC 2026** in Palma de Mallorca.

As language resources grow in scale and significance across linguistics, digital humanities, and language technology, the research community continues to grapple with the challenge of restricted-access textual data — particularly collections encumbered by copyright, licensing terms, or privacy constraints. **Derived Text Formats (DTF)**, also referred to as *extracted features*, have emerged as a promising pathway for enabling scientific inquiry and best practices of Open Science while at the same time respecting restrictions imposed by copyright law. By transforming texts into structured, interpretable, but non-reconstructible representations, DTF open important opportunities for reproducibility, comparability, and transparency in research while upholding legal and ethical obligations.

This workshop brings together researchers, legal scholars, infrastructure developers, and standardization experts to explore the multifaceted landscape of DTF. The contributions reflect active work and community experience on topics such as:

- methodologies for creating and processing Derived Text Formats
- legal and ethical considerations surrounding derived data publication
- practical use cases across digital humanities, linguistic research, corpus linguistics, and NLP
- tools, workflows, and infrastructure supporting DTF-based research
- theoretical investigations and standardization efforts

The breadth of submissions demonstrates the increasing relevance of derived data across fields that rely on sensitive or proprietary textual sources. The discussions and presentations showcased in this volume not only illuminate current practices but also point toward the development of robust community standards and sustainable infrastructures that support open science in legally complex environments.

We thank all authors for their thoughtful contributions and the members of the programme committee for their careful and constructive reviews. We are equally grateful to the participants — onsite and online — whose engagement makes this hybrid event a dynamic forum for exchange. Finally, we acknowledge the LREC 2026 Organising Committee for their support in hosting this workshop.

We hope that the papers compiled in this volume will inspire continued collaboration and innovation in leveraging Derived Text Formats to responsibly and effectively broaden access to textual resources.

The Workshop Organizers
LREC 2026
Palma de Mallorca

Organizing Committee

- Florian Barth, University of Göttingen
- Keli Du, University of Trier
- José Calvo Tello, University of Göttingen
- Philippe Genêt, German National Library
- Piroska Lendvai, Bavarian Academy of Sciences and Humanities
- Christof Schöch, University of Trier
- Thorsten Trippel, University of Tübingen and Leibniz-Institut for the German Language (IDS)

Table of Contents

Workshop Program

Tuesday May 12, 2026

14:00–15:30 **Session 1: Overview**
Room: Room 9
Chair: Philippe Genêt

14:00–14:10 ***Welcome and Introduction***

14:10–14:30 *Derived Text Formats as Strategic Transformations of In-Copyright Materials to Support Open Science: A Survey*
Christof Schöch

14:30–14:50 *A Multi-dimensional Constrained Framework for Derived Text Formats*
Keli Du and Christof Schöch

14:50–15:10 *Legal implications of Derived Text Formats - a copyright perspective*
Gianna Iacino, Pawel Kamocki and Keli Du

15:10–15:30 *Revisiting Masking After Fifteen Years: Early Approaches to Non-Reconstructable Linguistic Data in the current context*
Georg Rehm, Thorsten Trippel and Andreas Witt

15:30–16:00 ***Break***

16:00–18:00 **Session 2: Applications**
Room: Room 9
Chair: Piroska Lendvai

16:00–16:20 *Multi-Label Text Classification of Derived Text Formats with DistilBERT*
Jennifer Ecker and Roman Schneider

16:20–16:40 *Training data generation for context-dependent rubric-based short answer grading*
Pavel Šindelář, Filip Prášil, Dávid Slivka, Christopher Bouma and Ondrej Bojar

16:40–17:00 *DUO_DE A1: An Annotated Corpus of Online Learning Material for Beginning Learners of German as a Foreign Language*
Jammila Laâguidi, Vitaliia Ruban, Ronja Laarmann-Quante and Anastasia Drackert

17:00–17:20 *Why Reconstructing Scrambled Texts Fails*
Keli Du and Christof Schöch

Tuesday May 12, 2026 (continued)

17:20–17:40 *DIN 19461: A National Standard for Derived Text Formats*

Thorsten Trippel, Florian Barth, Jose Calvo Tello, Keli Du, Philippe Genêt, Daniel Kurzawe, Peter Leinen, Piroska Lendvai, Christof Schöch, Andreas Witt and Arden Zimmermann

17:40–18:00 ***Final discussion and closing***