Revisiting Masking After Fifteen Years: Early Approaches to Non-Reconstructable Linguistic Data in the Current Context

Georg Rehm*, **Thorsten Trippel**†**, **Andreas Witt****

*Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Salzufer 15/16 D-10587 Berlin, Germany
Humboldt-Universität zu Berlin, Dorotheenstraße, 24, 10117 Berlin, Germany
georg.rehm@dfki.de

†University of Tübingen, Keplerstraße 2, D-72074 Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

**Leibniz Institute for the German Language, R 5, 6-13, D-68161 Mannheim, Germany
{trippel, witt}@ids-mannheim.de

Abstract

This paper revisits corpus masking approaches introduced in 2007 for enabling the distribution of linguistically annotated corpora without exposing copyrighted or sensitive source texts and situates them within the contemporary framework of Derived Text Formats (DTFs). While the original work demonstrated how syntactic and morphological information could be preserved through parameterised masking, today's landscape, which is shaped by large language models, FAIR requirements, and emerging standardisation efforts, demands more formalised, robust and reproducible methods. We outline how DTFs extend early masking concepts by introducing explicit abstraction levels, reversibility classes, and machine-actionable provenance, supported by standards such as TEI, ISO linguistic annotation models, CMDI metadata, and the draft DIN DTF specification. Building on these foundations, we present a modern workflow for DTF generation, including enrichment pipelines, structural abstractions, statistical and embedding-based representations, and non-reversible transformation layers, illustrated through the MONA-pipe framework. A range of linguistic, digital-humanities and NLP use cases demonstrates the analytical utility of DTFs while maintaining legal compliance. We conclude that DTFs constitute a sustainable and infrastructure-ready solution for open, reproducible and legally secure text-based research in the decades to come.

Keywords: Derived Text Format, Masking, Linguistic Resources

1. Introduction

Research in linguistics and the digital humanities relies on empirical data. Text corpora are the backbone of many studies, enabling quantitative analyses, comparative investigations, and structural observations across a wide range of linguistic phenomena. Despite their central role in scholarly work, these corpora are often subject to legal restrictions that severely limit their distribution and reuse or even contractual restrictions that remain in force regardless of added annotation layers.

In 2007, these challenges led to early explorations of masking techniques, which aimed to separate linguistic annotations from protected textual content to enable data sharing in a legally compliant way. These initial approaches laid the conceptual groundwork later taken up and systematised within today's Derived Text Format (DTF) frameworks.

Now, almost 20 years later, the core problems remain, but the context has changed dramatically. Machine learning and large language models (LLMs) have heightened concerns about partial reconstruction of masked data, increasing the need for robust, non-reversible transformation methods. At the same time, the scientific community places greater emphasis on reproducibility, transparency, and open science, creating new incentives to standardise such transformations. Initiatives like the European Open Science Cloud (EOSC), the German National Research Data Infrastructure (NFDI) and the emergence of a DIN DTF standard (DIN 19461, currently in draft status) demonstrate that the ideas first explored in 2007 have now become part of a broader, coordinated effort to enable the lawful and sustainable use of text data in research.

This paper revisits and extends the early approaches in light of these developments. We examine how masking-based techniques can be integrated into modern DTF frameworks, how they respond to today's legal and technical challenges, and how they can support reproducible research.

2. From Masking to Derived Text Formats: A Historical Perspective

The idea of making annotated linguistic resources without distributing the underlying textual content has its origins in early discussions within compu-

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 25–33
12 May 2026.

tational linguistics and the digital humanities. In 2007, two complementary approaches were proposed that would later become foundational for what is now conceptualised as DTFs.

The first, introduced at the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007), (Rehm et al., 2007b)) focused on *masking treebanks*. In the same year, a more general framework was presented at Digital Humanities 2007. This *corpus masking* (Rehm et al., 2007a) approach explored the possibility of legally bypassing licensing restrictions by systematically obfuscating the source text while preserving the annotation layers.

Both papers were motivated by the same underlying tension: while annotations constitute independently copyrightable intellectual contributions of researchers, their dissemination was limited by the rights attached to the underlying source texts. The masking approaches offered an early strategy to keep annotations usable and shareable even when the legal conditions made the distribution of the original source texts impossible.

Looking back, these works anticipated several themes that have since become central to research data infrastructures. They recognised the *importance of sustainability*, the *separation of data and annotation layers*, and the *need for flexible licensing models*. They also outlined distinctions between legal regimes and research uses that are now standard in infrastructures such as Text+ (Hinrichs and Trippel, 2024) and the National Research Data Infrastructure (NFDI) in Germany, and European initiatives such as CLARIN (Fišer and Witt, 2022), META-SHARE (Piperidis et al., 2014), European Language Grid (Rehm, 2023) or European Language Data Space (Rehm et al., 2024).

At the time, however, these approaches remained largely conceptual. The field lacked formal standards, consistent metadata vocabularies, and infrastructural support for transformations of this kind. As a result, the masking work of 2007 was forward-looking but mostly confined to experiments with specific corpora and local workflows.

From today's perspective, the significance of these early approaches has become more clear. The rise of large-scale machine learning, increasing legal scrutiny, and the emergence of formalised DTF standards have created a landscape in which the ideas of 2007 are not only relevant but foundational. The conceptual separation between text and annotation, and the possibility of distributing transformed, non-reversible representations, laid the groundwork for many of the practices and standards currently being developed.

2.1. Masked Treebanks (TLT 2007)

The work presented at TLT 2007 (Rehm et al., 2007b) introduced one of the earliest systematic attempts to address the legal barriers surrounding the distribution of annotated corpora. The masked treebank approach demonstrated that many linguistic research questions rely primarily on syntactic and morphological structure rather than on lexical content. By replacing word forms with neutral placeholders while keeping constituent hierarchies, dependency relations and grammatical features intact, the method ensured analytical usefulness without exposing protected content. Different parameter settings allowed researchers to control how much morphological or formal information was preserved. This combination of structural fidelity and legal safety made masked treebanks one of the earliest viable strategies for sharing annotated resources under restrictive conditions.

2.2. Corpus Masking (DH 2007)

The second major contribution from 2007, presented at the Digital Humanities conference (DH 2007) (Rehm et al., 2007a), generalised the idea from treebanks to any XML-annotated corpora. The proposed *corpus masking* framework applied a parameterised randomisation algorithm that replaced surface forms with automatically generated strings while preserving the annotation layers. Depending on the configuration, selected formal properties – such as token length, case patterns or affix information – could be retained to balance legal constraints with analytical usefulness. This made it possible to redistribute annotated corpora under more permissive licensing conditions and supported a range of use cases, including unlexicalised parsing, NLP evaluation, teaching scenarios and sustainability efforts where the underlying texts could not be shared.

Many of the principles introduced in 2007 – such as parameterisation, structural preservation, non-reversibility and the separation of text and annotation – anticipate the requirements of today's DTFs. Contemporary work shows that these transformations enable legally compliant text and data mining while supporting reproducible research, provided that protected content is sufficiently abstracted (Schöch et al., 2020). Empirical studies further demonstrate that even strong obfuscation methods can retain substantial analytical value while preventing reconstruction (Du et al., 2025). As a result, DTFs have developed from early experimental masking approaches into infrastructure-ready components for sustainable data stewardship and long-term reusability.

3. New Challenges in 2026

Although the early masking approaches anticipated many of the ideas that now underpin DTFs, the landscape in which such transformations operate has changed profoundly over the past fifteen years. Advances in machine learning – most notably in LLMs – have created new risks related to the reconstruction and inference that were not conceivable at the time. At the same time, the growing emphasis on transparent and reproducible research has increased the demand for datasets that can be shared, cited and reused without infringing on copyright or privacy laws. These developments intersect with a complex legal environment shaped by copyright, contractual restrictions and data protection regulations, all of which impose increasingly stringent requirements on the handling of textual data. Together, these factors create a new constellation of challenges that modern DTFs must address, extending far beyond the assumptions and technical constraints of 2007.

3.1. LLMs and Reconstruction Risks

Modern machine learning models are capable of inferring missing lexical or semantic material from sparse cues. Even when surface forms have been fully removed, LLMs can exploit preserved structural patterns, token-level regularities or distributional signals to reconstruct plausible fragments of the original text or to approximate its content. Du et al. (2025) point out that currently the reconstruction is far from perfect, but caution needs to reign to avoid legal implications, especially if the quality of transformation further increases.

A masked corpus that appears non-reversible to a human analyst may not be non-reversible to a state-of-the-art model trained on vast amounts of data. As a consequence, masking is no longer sufficient on its own; it must be accompanied by a *formalised assessment of non-reversibility and information leakage*. Such assessment requires evaluating the transformation against realistic attack models that reflect modern inference capabilities, rather than assuming that the removal of surface forms is intrinsically safe.

These developments highlight the need to reconsider masking not merely as a technical transformation, but as a risk-managed process grounded in formal guarantees. Contemporary DTFs must therefore include explicit definitions of abstraction levels, measurable criteria for leakage, and transformation metadata documenting which information has been removed, retained or generalised.

3.2. Reproducibility Requirements

Reproducibility has become a central requirement in contemporary machine learning and computational linguistics. This kind of research requires accessible datasets, legal frameworks often prohibit sharing the underlying content.

DTFs offer a way to reconcile these conflicting demands. The exact transformed datasets used in a workflow can be published, allowing others to replicate results with full transparency regarding preprocessing steps and data provenance.

Furthermore, DTFs allow for the creation of benchmarks that remain usable over time even when the original corpora cannot be redistributed or when licensing conditions change. By separating protected content from its derived representational layers, DTFs make it possible to document and preserve the exact data conditions under which a model was trained or evaluated.

In this sense, reproducibility is not merely a desirable feature but an operational requirement – one that DTFs are uniquely positioned to support.

3.3. Legal and Ethical Context

Legal and ethical constraints surrounding text content remain major challenges for the distribution and reuse of linguistic corpora. Copyright and contractual restrictions often prevent redistribution of source content, even when annotation layers are fully owned by researchers (Lehmberg et al., 2007). Corpora frequently comprise multiple content layers (source content, annotations, metadata, derived structures), each subject to different legal regimes, which complicates reuse across institutions.

Privacy and data protection laws add further constraints. GDPR (European Parliament and Council of the European Union, 2026) prohibits sharing corpora containing personal or identifiable information, regardless of copyright status. While masking can reduce identifiability, it must be designed carefully: insufficient abstraction or the inferential power of modern LLMs can allow partial reconstruction, similar to failures in anonymisation.

Recent legal developments relevant to machine learning include the German text-and-data-mining (TDM) exception (UrhG) (also see Text+ on TDM), which permits certain forms of model training under specific conditions, provided that resulting models do not reproduce protected content and deletion obligations are met. However, the exception does not allow the redistribution of the underlying corpora, leaving reproducibility dependent on legally shareable derived formats.

Complex licensing structures further complicate matters. Large corpora may involve multiple rights holders, annotation layers contributed by different groups, and tools or contracts imposing additional restrictions. Clear provenance documentation is essential to specify which rights apply to each layer and how they are transformed in a DTF workflow.

By separating protected source texts from non-reversible derived representations, DTFs reduce legal risks and support compliance with copyright, contractual obligations and data protection requirements. Earlier analyses (Lehmberg et al., 2008) already highlighted the need for such differentiated data handling; in the 2026 landscape, it has become an operational requirement. DTFs thus function not only as technical artefacts but as legally and ethically robust instruments for sustainable and compliant research data practices.

4. Standardisation Landscape

The increasing legal, technical and methodological complexities surrounding textual data have intensified the need for clear, formalised frameworks that govern how DTFs are created, documented and shared. While the early masking approaches were primarily technical experiments developed in the absence of established standards, today's research data ecosystem demands structured, interoperable and machine-actionable specifications. Here, standardisation plays a crucial role: it transforms ad-hoc masking techniques into reproducible, transparent and legally robust workflows. Emerging efforts such as the DIN standard for Derived Text Formats (DIN 19461:2026-04 , E, currently a draft national standard), along with existing ISO standards for linguistic representation and metadata (such as ISO 24622-1:2015, ISO 24622-2:2019, ISO 24619:2011) provide a coherent foundation upon which modern DTF practices can be built. Below, we outline this evolving standardisation landscape and its implications for sustainable research data management.

4.1. DTF Standard DIN 19461

The emerging DTF DIN standard (DIN 19461:2026-04 , E) provides the first formalised framework for describing, generating and documenting transformed text representations in a standardised and methodologically transparent way. Unlike the early approaches, the new standard offers a systematic categorisation of transformation types, information-reduction operations and metadata obligations that apply to all formats derived from natural-language text.

At its core, the standard defines DTFs as structured text formats produced through *targeted enrichment and information reduction* of an original source text. It distinguishes clearly between enrichment steps that add analytical layers such as POS tags, lemmas, NER information or syntactic structures, and information-reduction operations designed to eliminate or generalise protected content. The explicit modelling of transformation procedures formalises what earlier work on corpus masking and masked treebanks treated implicitly.

A central requirement of the DIN standard is the assurance of *non-reconstructability*. The standard recognises that reconstruction risks may arise not only from a single transformation but also from the *combination of multiple DTFs* generated from the same source material. Section 4.4 therefore mandates a systematic evaluation of combined leakage risks – an issue anticipated in 2007, but now formulated explicitly as a normative obligation.

The standard also introduces comprehensive *documentation and metadata requirements*. All enrichment and reduction steps must be fully documented, including tools, algorithms, parameters and the granularity at which each operation was applied. Documentation should be linked to machine-readable CMDI metadata following (ISO 24622-1:2015) and (ISO 24622-2:2019), ensuring traceability and long-term sustainability. This establishes a formal reproducibility framework that earlier masking work lacked.

In addition, the standard defines canonical *categories of derived formats*, including token-based DTFs (e. g., bag-of-words and n-gram representations), vector-based DTFs (e. g., TF/IDF, Word2Vec or contextual embeddings), structured DTFs (e. g., shuffled segments) and multi-feature formats (e. g., HathiTrust Extracted Features). By codifying these patterns, the DIN standard extends concepts implicit in masking – replacing words with placeholders, removing sequence information, randomising tokens – into a general typology that reflects contemporary NLP and text-analytic needs.

While the canonical DTF categories define token-based units as the smallest standardised representational layer, subtoken-level information units – such as character n-grams, morphologically defined graphemic segments, or tokenizer-produced subword units – can also be modelled within the DTF framework. Their inclusion changes both analytical utility and leakage profiles. DIN 19461 therefore treats subtoken-level DTFs as special cases of information-reduction or enrichment steps whose granularity must be documented. Modern masking workflows often implicitly operate on such units (e. g., affix-aware masking), and evaluation must consider reconstruction pathways that exploit LLM tokenizers or subword distribution patterns.

4.2. International Standards

A crucial element of the emerging DTF ecosystem is its alignment with established standards for linguistic representation and metadata. While the DIN DTF standard defines the conceptual and operational framework for DTFs, its practical implementation relies on interoperable modelling languages, annotation formats and metadata infrastructures (DIN 19461:2026-04 , E).

The *TEI Guidelines*, in particular the module for *Feature Structures*, provide a flexible mechanism for representing linguistic information independently of the underlying text. Feature structures support hierarchical, attribute–value-based annotations and are already used in real-world DTF implementations where annotation layers remain intact while textual content is masked or transformed (DIN 19461:2026-04 , E).

Beyond TEI, a suite of *ISO standards for linguistic annotation* ensures cross-infrastructure interoperability. These include models for syntactic structure (ISO 24615-1:2014) (SynAF), morphosyntactic units (ISO 24611-1:2025) (MAF), and lexical representations (ISO 24613-1:2024) (LMF), as well as frameworks for structural, discourse and semantic annotation (ISO 24612:2012) (LAF) and the ISO 24617-x series (SemAF, see for example ISO 24617-1:2012). Together, they provide the conceptual backbone referenced throughout the DIN DTF draft (DIN 19461:2026-04 , E).

The *Component Metadata Infrastructure (CMDI)* (ISO 24622-1:2015; ISO 24622-2:2019) offers a mechanism for documenting linguistic resources. CMDI profiles can capture the enrichment and reduction steps required for DTF, ensuring transparent and machine-actionable provenance (DIN 19461:2026-04 , E).

Finally, *persistent identifiers (PIDs)* support long-term referencing and provenance tracking. Systems such as DOI, Handle or the identifier mechanisms defined in (ISO 24619:2011) allow infrastructures to systematically track all DTF variants derived from a source corpus and thus complement FAIR-aligned data stewardship practices (DIN 19461:2026-04 , E).

Together, TEI Feature Structures, ISO annotation standards, CMDI metadata infrastructure and persistent identifier systems provide the technical foundation that enables DTF to function as transparent, sustainable and interoperable components within modern research data ecosystems.

4.3. Infrastructural Implications

The introduction of a formal DIN standard for DTFs has significant implications for national and international research infrastructures. Infrastructures for language resources including sustainability and repository services depend on interoperable formats, persistent identifiers and machine-readable metadata to ensure long-term access, legal compliance and cross-project reuse. The DIN DTF standard provides a unified framework that enables derived textual data to be integrated into these ecosystems in a transparent and robust manner (DIN 19461:2026-04 , E).

By requiring the documentation of enrichment steps, information-reduction operations, granularity choices and tool parameters – captured through CMDI-compatible metadata profiles (ISO 24622-1:2015; ISO 24622-2:2019) – the standard ensures that DTFs can be archived, validated and reused across institutions. Maintaining internal and external records of all DTF variants derived from a source corpus supports long-term preservation and risk assessment, while the linkage of transformation metadata to persistent identifiers (ISO 24619:2011) allows repositories to manage DTFs as first-class digital objects (ISO 24619:2011).

Beyond technical specifications, the DIN DTF standard acts as an infrastructural enabler. By aligning DTF workflows with established standards for metadata, annotation, provenance and sustainability, it facilitates seamless integration of derived representations into the major research data infrastructures essential for open science and enduring accessibility of textual resources.

5. Revisiting and Modernising Masking Algorithms

The early masking techniques formed an important starting point for generating legally shareable representations of copyrighted or sensitive corpora. They demonstrated that substantial syntactic and morphological information can be preserved even when lexical material is removed or obfuscated. At the same time, the assumptions underlying the original methods – especially concerning reconstructability and acceptable abstraction levels – reflected the technological landscape of their time. With the advent of LLMs, machine-learning architectures and formalised DTF standards, these early approaches require systematic revision.

This section summarises the conceptual foundations of the original masking approach, outlines the requirements modern DTF must meet and presents a contemporary, multi-stage workflow for producing legally compliant and reproducible DTFs.

5.1. The Original Masking Approach

The 2007 masking strategies were built around the idea of *parameterised masking*: configurable transformation settings allowed researchers to adjust how much information was retained, modified or removed. Central to this approach was a *dictionary-based randomisation* procedure that replaced each token with an artificial string – preserving length or formal characteristics where needed – while ensuring internal coherence across the corpus. Additional *affix-aware strategies* enabled the retention of morphological cues such as case, number or tense, even when stems were fully masked. Low-risk closed-class items could be selectively retained to support syntactic modelling without introducing significant leakage risks.

These ideas were operationalised in the *CorpusMasker* tool (Dellert, 2007; Rehm et al., 2007a), which provided a GUI for experimenting with masking parameters and applying transformations to XML-annotated corpora. The resulting principles – parameterisation, affix sensitivity, selective retention and controlled randomisation – remain foundational for contemporary DTF approaches.

5.2. Updated Requirements for DTF

Transforming masking into fully fledged DTFs introduces several new requirements.

First, *robustness against reconstruction* has become essential. Modern LLMs can infer missing lexical or semantic information from subtle structural or distributional cues. Removal or randomisation of surface forms is no longer sufficient on its own; DTFs must be evaluated against realistic attack models and provide quantifiable, machine-verifiable guarantees of non-reversibility.

Second, contemporary DTFs require *explicit abstraction levels and reversibility classes*. Instead of a single continuum of obfuscation, DTF frameworks must define what types of linguistic or statistical information remain in a derived representation and what forms of reconstruction are theoretically possible. This includes classes such as token-level replacement, morphological generalisation, sequence removal or embedding-level abstraction, each with different leakage profiles.

Third, modern infrastructures require *comprehensive provenance and machine-actionable metadata*. All enrichment and reduction operations must be documented – tools, parameters, granularity choices and processing steps – so that repositories and workflows can validate, reproduce and assess the transformation. Such provenance is essential for legal compliance, long-term preservation and transparent scientific practice.

Together, these requirements transform masking from an informal technique into a principled, standardised methodology embedded within the broader DTF framework.

5.3. A Workflow for DTF Generation

Modern workflows for generating DTFs extend far beyond the ad-hoc masking procedures of 2007. They integrate enrichment, structural abstraction, statistical modelling and legally robust information reduction into reproducible multi-stage pipelines.

The process begins with the *extraction of structural representations* from the source text. Prior to any reduction, the text is enriched with annotation layers such as tokenisation, part-of-speech tags, lemmas, morphological features, named entities and syntactic dependencies. These layers form the basis for transformations that preserve analytic value while removing protected content.

Next, the workflow applies *structural abstraction*, producing feature bundles, e. g., POS or lemma classes, dependency graphs or other units defined in the DIN DTF framework. This separates the protected source text from the shareable representational layers.

In parallel, workflows may generate *statistical and embedding-based formats*, including TF/IDF features, topic-modelling inputs, token or sentence embeddings and contextualised representations. These abstractions retain analytical utility without exposing lexical material.

A dedicated *non-reversible transformation layer* ensures that the resulting DTF cannot be used to reconstruct the source text. Depending on the abstraction level and reversibility class, this layer applies operations such as deletion, replacement or randomisation. Modern pipelines must ensure that combinations of retained information do not permit reconstruction, even by advanced systems.

DTFs do not assume that the source material is written text. Any linguistic representation – handwritten, OCR-derived, ASR output, phonetic or graphemic transcription – can serve as source material. For born-derived modalities, the same principles apply: enrichment layers (POS, phonetic features, timing annotations) precede information reduction, and reversibility is assessed with respect to the original modality.

A concrete implementation of this approach is provided by *MONAPipe* (Dönicke et al., 2022; MONAPipe 2023), whose `derived_text_formatter` applies DIN-aligned operations such as `replace`, `randomize` and `keep` at multiple text levels. The use depends on the language models selected, MONApipe's default being German at present. While MONAPipe already supports the main transformation operations, it does not yet generate a complete provenance record for example in a CMDI serialisation, which is an important requirement for future development.

Taken together, these developments transform masking from a single operation into an integrated, multi-stage workflow. Contemporary DTF pipelines combine enrichment, abstraction, statistical representation and non-reversible transformation to produce legally compliant, reproducible and infrastructure-ready representations of textual data.

DTF enable a broad range of research applications, from stylometry and diachronic studies to NLP benchmarking and pedagogical use. These applications benefit from the structural preservation and non-reversibility provided by DTF workflows.

6. Discussion

The emergence of DTFs has opened new possibilities for research on copyrighted or sensitive textual data, but it also highlights several important trade-offs. Foremost among these is the tension between *utility* and *non-reversibility*: retaining too much linguistic or distributional information can increase the risk of reconstruction, while overly aggressive abstraction may reduce the analytical value of the data. Determining the appropriate level of transformation requires explicit modelling of abstraction levels, reversibility classes and information-reduction operations, as well as evaluation against potential reconstruction pathways – including those made possible by modern LLMs.

Several *limitations of current approaches* persist. Even well-designed masking or randomisation strategies cannot guarantee absolute non-reversibility, particularly when external knowledge or powerful generative models can infer missing content from subtle linguistic cues. The ability of LLMs to recover stylistic, syntactic or semantic patterns raises questions about the long-term safety of DTFs that retain traces of original structure. Furthermore, the development of DTF workflows varies across research communities, with some digital humanities domains lacking standardised pipelines. Tools such as MONA-pipe implement DIN-aligned transformation operations, but still do not automatically generate the complete provenance metadata required for reproducible and legally robust publication – an important aspect for future work.

These challenges underscore the *importance of standardisation*. The techniques introduced in 2007 established key principles – parameterisation, structured abstraction and controlled leakage – that underpin modern DTF approaches. The DIN DTF standard extends these ideas by providing formal categories, transformation operations, documentation obligations and provenance requirements.

Several *open questions* remain. Defining thresholds for irreversibility, quantifying leakage risks, automating provenance-aware metadata generation and balancing legal compliance with scientific utility will require continued interdisciplinary collaboration. Ultimately, the promise of DTFs lies not only in enabling access to restricted data but also in establishing a transparent, accountable and reproducible ecosystem for text-based research – one that remains resilient as technologies, legal frameworks and research practices continue to evolve.

Future work includes the development of metrics that quantify both divergence from the original and retained utility. Current practice relies on (i) statistical divergence measures (e.g., KL or JS divergence of POS or dependency distributions), (ii) structural similarity metrics (tree edit distances, graph similarity), (iii) task-based utility evaluations (NER, parsing, text classification performance on DTF vs. original), and (iv) leakage assessments using reconstruction experiments Du et al. (2025). While DIN 19461 introduces reversibility classes and documentation obligations, it does not prescribe specific metrics; formal benchmarks for leakage and utility are therefore an open research need.

7. Conclusion

Nearly two decades separate the first explorations of corpus masking from DTFs in 2026. During this period, the research community has progressed from prototypes toward a mature and standardised ecosystem for legally compliant text transformations. The original work anticipated many of the challenges that would later become central – most notably the need to preserve linguistic structure while preventing reconstruction – and laid the foundations on which modern approaches now build.

The updated framework presented in this paper translates the early insights of parameterised masking, affix-aware strategies and controlled leakage into a formalised model incorporating explicit abstraction levels, reversibility classes, provenance requirements and non-reversibility guarantees suited to an era shaped by LLMs. Tools such as MONAPipe, alongside emerging standards like the DIN DTF draft standard, provide practical implementations of these concepts and illustrate how masking can evolve into transparent, reproducible and extensible workflows.

DTFs also contribute directly to broader goals in research data management. By enabling the publication of legally usable, machine-actionable representations of copyrighted or sensitive corpora, they support the reproducibility of computational experiments, the creation of open benchmarks and the long-term stewardship of linguistic resources. They enhance compliance with legal and ethical requirements without diminishing the analytical potential of derived datasets.

Looking ahead, DTFs could become a sustainable and foundational component of national and international research infrastructures. Their integration into metadata standards, repository systems and NLP pipelines will help ensure that text-based research remains transparent, responsible and legally viable as technologies and legal contexts continue to evolve. Future work on automated provenance generation, leakage evaluation frameworks and interoperable DTF toolchains will further strengthen this role. DTFs represent not only a technical solution but also a structural contribution to open science: a means of enabling rigorous scholarship while respecting the rights and protections inherent in the underlying textual data.

8. Acknowledgements

Though the authors are indebted to various co-authors working on this topic for years, work on this paper was carried out within the National Research Data Infrastructure (NFDI) association. The NFDI is funded jointly by the Federal Republic of Germany and the 16 federal states, and the Text+ consortium is supported by the German Research Foundation (DFG). Georg Rehm is part of NFDI4DS – the National Research Data Infrastructure for Data Science and Artificial Intelligence, grant number 460234259; Andreas Witt and Thorsten Trippel are part of the Text+ consortium, grant number 460033370. The authors gratefully acknowledge this support, as well as the engagement of all institutions and individuals contributing to the NFDI and its goals.

The authors acknowledge the use of Large Language Models (LLMs) as writing aids in phrasing this paper, based on the authors' notes, ideas and concepts. The authors retain full responsibility for the content.

9. Bibliographical References

Johannes Dellert. 2007. Corpusmasker: A tool for parameterised masking of linguistic resources. https://www.lingexp.uni-tuebingen.de/sfb441/c2/corpus-masker-0.1.tar.gz. Version 0.1, SFB 441 "Linguistic Data Structures", University of Tübingen.

DIN 19461:2026-04 (E). 2026. Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF). Technical report, Deutsches Institut für Normung, Berlin.

Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. Monapipe: Modes of narration and attribution pipeline for german computational literary studies and language analysis in spacy. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, Potsdam, Germany.

Keli Du, Sarah Ackerschewski, Uygar Navruz, Nazan Sınır, Julian Valline, and Christof Schöch. 2025. Reconstructing shuffled text. bad results for nlp, but good news for using in-copyright texts. *Journal of Computational Literary Studies*, 4(1).

European Parliament and Council of the European Union. 2026. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

Darja Fišer and Andreas Witt, editors. 2022. *CLARIN: The Infrastructure for Language Resources*. De Gruyter, Berlin, Boston.

Erhard Hinrichs and Thorsten Trippel. 2024. Text+ – concept and benefits for empirical researchers. *Cybernetics and Information Technologies*, 24(4):143–163.

ISO 24611-1:2025. 2025. Language resource management — morphosyntactic annotation framework (MAF) — part 1: Core model. International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24612:2012. 2012. Language resource management — linguistic annotation framework (LAF). International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24613-1:2024. 2024. Language resource management — lexical markup framework (LMF) — part 1: Core model. International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24615-1:2014. 2014. Language resource management — syntactic annotation framework (SynAF) — part 1: Syntactic model. International Standard, International Organisation for Standardizaton (ISO), Geneva.

ISO 24617-1:2012. 2012. Language resource management — semantic annotation framework (SemAF) — part 1: Time and events (semaf-time, iso-timeml). International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24619:2011. 2011. Language resource management – Persistent identification and sustainable access (PISA). International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-1:2015. 2015. Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-2:2019. 2019. Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language. International Standard, International Organization for Standardization (ISO), Geneva.

Timm Lehmberg, Christian Chiarcos, Georg Rehm, and Andreas Witt. 2007. Rechtsfragen bei der nutzung und weitergabe linguistischer daten. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen*, pages 93–102. Gunter Narr Verlag, Tübingen.

Timm Lehmberg, Georg Rehm, Andreas Witt, and Felix Zimmermann. 2008. Digital text collections, linguistic research data, and mashups: Notes on the legal situation. *Library Trends*, 57(1):52 – 71.

MONAPipe 2023. 2023. Monapipe: Modes of narration and attribution pipeline. `GitLab repository`. Accessed: 2026-03-20.

Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi. 2014. META-SHARE: One year after. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1532–1538, Reykjavik, Iceland. European Language Resources Association (ELRA).

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiļjevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitrios Galanis, Maria Gavriilidou, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdiņš, Elena Leitner, Valérie Mapelli, Hélène Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault, and Leon Voukoutis. 2024. Common European Language Data Space. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3579–3586, Turino, Italy. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL). May 20-25, 2024.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007a. Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In *Digital Humanities 2007. Conference Abstracts*, pages 166–170, Urbana-Champaign. University of Illinois.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007b. Masking treebanks for the free distribution of linguistic resources and other applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, pages 127–138, Bergen, Norway.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. Abgeleitete textformate: Text und data mining mit urheberrechtlich geschützten textbeständen. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*, 5.

Text+ on TDM. 2025. Why the training of large language models falls within the scope of the text and data mining exceptions. Last modified: 23 January 2025.

UrhG. 2021. § 44b urhg – text und data mining. `https://www.gesetze-im-internet.de/urhg/__44b.html`. Zugriff am 16. Februar 2026.