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**[‡]

[|||]University of Tübingen
Keplerstraße 2, D-72074 Tübingen
thorsten.trippel@uni-tuebingen.de

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

[*]University of Göttingen
Papendiek 14, D-37073 Göttingen
florian.barth@uni-goettingen.de, { kurzawe,calvotello}@sub.uni-goettingen.de

[§]University of Trier
Universitätsring 15, 54296 Trier
{duk,schoech}@uni-trier.de

[‡]German National Library
Adickesallee 1, D-60322 Frankfurt am Main
{P.Genet, Ar.Zimmermann}@dnb.de

[+]Technical University of Darmstadt
Karolinenplatz 5, 64289 Darmstadt
peter.leinen@tu-darmstadt.de

[¶]Bavarian Academy of Sciences and Humanities
Alfons-Goppel-Str. 11, 80539 München
piroska.lendvai@badw.de

Abstract

We present DIN 19461:2026-06 (E), a German draft national standard that defines categories, terminology, and process requirements for Derived Text Formats (DTFs) created from text documents in natural language. The standard specifies enrichment and information reduction operations, requirements for combining multiple DTFs, and documentation obligations for publication, archiving, and reuse. Its aim is to enable legally compliant sharing and analysis of texts–especially where copyright or data protection prevents distributing originals–while maintaining scientific utility and reproducibility through explicit process and parameter recording. We outline the scope, the key concepts, the four core reduction operations (retain, delete, replace, randomise), together with examples across token-, structure-, and vector-based DTFs, and implications for infrastructures (e.g., ISO 24622-based metadata). Finally, we discuss limitations, open questions (e.g., reconstruction risks with modern ML models), and next steps for adoption and maintenance.

1. Introduction

Access to large text collections and their analysis via Text and Data Mining (TDM) is foundational for many research domains, yet legal constraints often restrict sharing copyrighted texts in their original form. The standard DIN 19461:2026-06 (E), which is a standard developed within the national organization of standardisation in Germany (DIN), addresses this by standardising how Derived Text Formats (DTFs) are defined, produced, and documented with the goal that they can be shared without enabling trivial reconstruction of the source text.

This paper contributs: (i) a concise overview of DIN 19461's scope and rationale; (ii) a presentation of its conceptual framework, requirements, and operations; (iii) worked examples spanning token-, structure-, and vector-based DTFs; and (iv) implications for infrastructures and reproducibility.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 67–75
12 May 2026.

2. Background and Motivation

In the context of the German National Research Data Infrastructure (NFDI), it is essential to provide researchers with reliable and legally compliant access to a wide range of research data. NFDI (Kraft et al., 2021) is a public initiative in Germany to build a research infrastructure for all research disciplines build on the FAIR principles (Wilkinson et al., 2016) with a clear focus on research data management established by the federal government and the German states (GWK – Gemeinsame Wissenschaftskonferenz, 2018). The consortium Text+ (see e.g. Hinrichs and Trippel, 2024) addresses language and text based research data. For language- and text-based data—central in linguistics, digital humanities, and natural language processing— the requirement of providing reliable and legal access to data is particularly challenging. Such data are frequently subject to legal and ethical constraints that restrict their distribution, while at the same time posing technical challenges related to processing, documentation, and reproducibility. This section outlines these two major dimensions: on the one hand, the legal and ethical considerations that shape the conditions under which text data may be shared, and on the other hand, the technical prerequisites needed to create formats that remain useful for research without compromising applicable restrictions, such as limited reusability due to licenses not granting this, or possible privacy related data.

2.1. Legal and ethical constraints

Language and text data are frequently subject to restrictions arising from both copyright and privacy considerations. In many jurisdictions, copyright law limits what can be done with contemporary texts and governs how authors' economic rights are protected. These rights ensure that creators can control the use and dissemination of their works and can benefit financially from them. As a result, research projects cannot always redistribute original texts, even when these texts are needed for scientific analysis. In addition to such copyright-related constraints, ethical obligations also arise from the nature of the data. Text collections may contain sensitive or personal information—for example, learner corpora that document individual writing performance. Even if such corpora do not always reach the threshold of originality required for copyright protection, they may still raise privacy concerns (see European Parliament and Council of the European Union (2026)) if the individuals who produced the text samples could be identified or exposed, especially in contexts where the content may be perceived as personal or potentially embarrassing. In Germany, the rights of authors to creative works are established by the ¨Urheberrechtsgesetz¨(UrhG).

DIN 19461 addresses these issues by positioning DTFs as a mechanism to reduce and abstract content in ways that preserve the ability to answer research questions while preventing reconstruction of the original text or making such reconstruction require disproportionate effort. The determination of whether a specific workflow meets legal obligations remains the responsibility of implementers.

2.2. Technical Challenges and Gaps

Technical challenges in working with language and text data arise at several levels. Prior approaches to "masking"(see for example Rehm et al., 2007 and Lehmberg et al., 2008) or abstracting text differ substantially across projects and infrastructures, and until recently no common terminology or requirements catalogue has existed. As a consequence, practices for reducing or altering text to meet legal and ethical constraints have been inconsistent and often difficult to compare. DIN 19461 responds to this inconsistent practices by introducing a unified vocabulary–for example the notions of DTFs (see also Schöch et al., 2020), information reduction, transformation, and generalisation–and by specifying that all operations and parameters must be documented at clearly defined levels of granularity, such as token, sentence, paragraph, document, or collection.

Beyond these conceptual gaps, practical technical issues must also be considered. On the one hand, many research workflows rely on large quantities of text data that are already available in digital form. These may include formats such as EPUB publications or HTML-based sources. Once such digital formats are available, the central question becomes how they may be used and what technical means exist to transform them automatically into representations that continue to support research needs while reducing legal and ethical risks.

Automated procedures play a key role in this transformation. When large-scale collections of digital text are involved, it is often impractical to rely on manual curation. Instead, workflows must support efficient, reproducible, and scalable operations that alter or abstract content without requiring excessive computational effort or complex manual intervention. Developing such workflows–capable of handling diverse formats, documenting each processing step, and ensuring that legal and ethical constraints remain respected–constitutes a significant technical challenge.

3. Standardisation Process

The development of DIN 19461 followed established procedures for national standardisation.This section provides an overview of the working group, its methodology, and the alignment of the resulting standard with related terminology and models from existing standards. For more information on the standardisation process see also (Preissner and Heid, 2025).

Standards developed within formal processes such as those of national standards, in Germany by the standardisation organisation DIN, or ISO, the International Organization of Standardisation, follow highly structured and rigorously regulated procedures. The documents themselves typically conform to established templates and organisational principles, which we do not discuss in detail here, as they are shared across many standards. Nevertheless, it is useful to provide a brief overview of the types of provisions contained in DIN 19461 in order to clarify how the standard supports the production, documentation, and publication of DTFs.

3.1. Working Group and Methodology

As in other formal standardisation processes, the creation of DIN 19461 began with the submission of a work-item proposal intended to assess whether a standard on DTFs was necessary and feasible. The responsible standards committees reviewed the proposal and concluded that, although only a limited number of data-holding institutions were at that time providing DTFs, it was nonetheless important to establish a structured and consistent basis for evaluating the legal situation surrounding derived formats. A standard would allow such evaluations to follow reproducible criteria rather than ad-hoc institutional decisions.

Based on this assessment, a working group was established within the German Standards Organization (DIN) working group NA 105-00-06 AA “Sprachressourcen und Sprachtechnologie”, drawing on expertise from the national research data infrastructure. The group developed a draft standard that defines terminology, describes the relevant operations and procedures, and specifies requirements for the creation and documentation of DTFs.

The question of why an international standard was not pursued from the outset was considered in the early stages. Although other jurisdictions, particularly within the broader European legal context, may eventually recognise the usefulness of such a standard, the immediate use case was rooted in national requirements and the needs within national research data providers. DIN 19461 therefore focuses first on addressing needs emerging from research data infrastructures in Germany, while leaving open the possibility that its concepts and procedures may inform international efforts in the future.

3.2. Alignment and References

DIN 19461 does not exist in isolation. It draws on several concepts and models already established in existing standards for linguistic resources. Within ISO/TC 37/SC 4, numerous standards are relevant to the creation, documentation, and referencing of language resources, and these form part of the conceptual background against which DIN 19461 was developed.

For example, persistent identifiers (PIDs) as defined in (ISO 24619:2011) are essential for reproducibility. When information-reduction operations are applied and a DTF no longer contains the original text in recognisable form, PIDs ensure that the underlying source can still be referenced reliably. Similarly, ISO 24622 (CMDI, see ISO 24622-1:2015 and ISO 24622-2, 2019) provides a component-based framework for metadata and serves as a foundation for modelling the metadata required for DTFs. Using CMDI components allows the documentation of enrichment steps, reduction operations, provenance, and processing parameters in a structured and interoperable manner.

Beyond these standards, several other models are relevant as conceptual or technical precursors. These include frameworks such as the Linguistic Annotation Framework (LAF, ISO 24612:2012) and feature-structure-based models (ISO 24610:2008) used for representing linguistic information, which can underpin enrichment steps and serve as starting points for generating DTFs based on stand-off annotations. Such standards and frameworks do not prescribe specific procedures for creating DTFs, but they provide established terminology, structural patterns, and annotation practices that can be employed within the workflows described in DIN 19461.

4. Overview of DIN 19461

For researchers, it is not only important to understand that DTFs provide a way to make text data usable under legal and ethical constraints, but also that a standardised approach exists for creating and documenting such formats. Data-holding institutions carry responsibility for complying with legal requirements while still enabling collaboration, transparency, and reliable long-term access. A common standard helps ensure that decisions about which data can be shared, and under what conditions, are based on consistent criteria rather than local interpretations.

Against this background, the following section

introduces the scope, conceptual framework, and requirements defined in the standard, and explains how these elements assist institutions in creating DTFs that are both legally robust and technically transparent.

4.1. Scope and Purpose

DIN 19461 applies to the classification and uniform description of methods and procedures used for creating DTFs from natural-language text documents. It covers both semi-structured data, such as XML or TEI representations, and unstructured sources, such as plain text, provided that these materials encode language at the character level. The focus of the standard lies on identifying how enrichment and information-reduction operations produce derived formats that remain analytically useful while preventing reconstruction of the original text in ways that could infringe legal or ethical constraints.

The scope of DIN 19461 explicitly excludes representations of text as images; only once such materials have been transformed into machine-readable digital text–for example through OCR–do they fall within the domain of the standard. Moreover, the standard does not make legal determinations about the status of any particular DTF. Instead, it provides the conceptual and procedural framework that allows practitioners to produce, document, and evaluate DTFs in a consistent, transparent, and assessable manner.

Within this scope, DIN 19461 defines terminology, units of granularity, categories of operations, and requirements that govern the creation of DTFs. It also establishes the documentation obligations necessary to ensure that such formats can be interpreted, compared, and reused across institutions and projects. The standard thereby provides a stable foundation for producing derived formats that simultaneously support research goals and respect the legal and ethical boundaries of the underlying source material.

4.2. Conceptual Framework

The conceptual framework of DIN 19461 introduces the fundamental notions required to describe, generate, and evaluate DTFs.

The standard distinguishes several key concepts. *Information reduction* refers to operations that remove, alter, or generalise textual content in a controlled manner. *Transformation* captures the conversion of text segments into new representational forms, such as linguistic categories or numerical vectors. *Generalisation* describes abstraction steps that replace specific linguistic content with higher-level descriptors. To ground these operations, DIN 19461 formalises the units on which they may apply, including text, tokens, and sequence information such as sentences, paragraphs, or larger structures.

The conceptual framework also incorporates terminology from linguistic and computational methods, including part-of-speech categories, lemmas, named entities, syntactic relations, and embeddings. Such concepts enable the enrichment of source material with linguistic annotations that may subsequently form the basis for reduction or transformation into a DTF.

Finally, the standard enumerates units of granularity and their associated types of annotation. These granularities–ranging from individual tokens to entire document collections–provide the structural reference points for both enrichment and reduction. By defining concepts and units systematically, DIN 19461 establishes a coherent vocabulary for describing how DTFs are produced and how their properties can be evaluated across projects and institutions.

4.3. Requirements Structure

DIN 19461 provides the requirement structure setting out the central requirements that govern the creation, documentation, and publication of DTFs. These requirements address the entire workflow from initial enrichment to the evaluation and dissemination of the resulting formats. First, the standard defines how enrichment procedures must be described, including the linguistic or structural annotations applied to the source material and the tools, models, and parameters used. Second, it specifies the requirements for information reduction, which is implemented through four well-defined operations: selective retention, deletion, replacement, and randomisation. Each operation must be applied at an explicitly stated level of granularity, and its effects must be documented in a way that enables transparent assessment of the resulting DTF.

Beyond individual operations, the standard also provides requirements for evaluating combinations of DTFs that originate from the same source material. Such combinations may increase the risk of reconstructing the original text, and DIN 19461 therefore mandates that their joint effects be considered when assessing whether a set of DTFs remains compliant with legal and ethical constraints. Finally, the standard outlines the prerequisites for publication, including mandatory metadata describing methods, tools, granularity levels, parameters, and any additional contextual information required to ensure reproducibility and to support the evaluation of reconstruction risks (see for example (Du et al., 2025)). Collectively, these requirements create a framework that enables consistent production and responsible sharing of DTFs across institutions and projects.

5. Core Operations for DTFs

In the following subsections, we describe the four core operations defined in the standard: enrichment, selective retention, deletion, and randomisation.

5.1. Enrichment

Enrichment is the first step in producing a DTF and serves as the foundation for all subsequent information-reduction operations. In DIN 19461, enrichment refers to the addition of linguistic, structural, or statistical information to the source text before any transformation, deletion, or abstraction takes place. At the same time, the standard recognises that enrichment may also consist of *not* adding any additional annotation. In such cases, the text is used exactly in the form in which it is available, without further linguistic or structural augmentation. This minimal form of enrichment remains a valid option whenever no additional annotation is required for the intended reduction steps.

If enrichment *is* performed, it must precede all information-reduction operations. This ensures that any subsequent transformations rely on consistent, traceable, and high-quality input data. Enrichment can include a wide range of annotations, depending on the research context and the level of granularity relevant to the intended DTF. Examples include the assignment of part-of-speech categories, lemmas, named entities, syntactic or dependency relations, and other forms of linguistic analysis. It may also involve information obtained through computational methods, such as vector-based representations or statistical measures extracted from the source material.

5.2. Selective Retention

Selective retention refers to the controlled preservation of certain pieces of information from the source text, provided that these elements are relevant for subsequent analytical tasks and can be retained at a given level of granularity without, on their own, revealing the original textual content. In DIN 19461, selective retention does *not* aim to produce a legally unproblematic DTF by itself. Instead, it constitutes an initial step that ensures that the information required for later processing remains available while preparing the ground for further information-reduction operations.

The key requirement is that selective retention should preserve only those elements that remain compatible with the intended reduction workflow. At the chosen granularity level—whether token, sentence, paragraph, or document—retained information must support the analytical purpose without obstructing the subsequent deletion, replacement, or randomisation that may be necessary to ensure that the final DTF cannot be trivially reconstructed. Selective retention therefore contributes to shaping the representational basis on which later reduction steps operate, but it does not by itself guarantee non-reconstructibility or legal compliance.

In practice, selective retention may involve keeping structural delimiters, metadata, positional information, segment boundaries, or statistical properties that support later analytical methods. Where linguistic annotations such as lemmas, part-of-speech categories, or named-entity labels are retained, this must be done with the understanding that further reduction operations may still be needed to prevent reverse mapping to the original lexical items. DIN 19461 therefore requires transparent documentation explaining why the retained information is relevant for the intended analysis, how it relates to the chosen granularity level, and how it fits into the broader sequence of reduction steps. Through this staged approach, selective retention helps ensure that the resulting DTF can be transformed into a legally robust format by applying the subsequent operations defined in the standard.

5.3. Deletion

Deletion constitutes one of the fundamental operations in the derivation of a DTF, focusing on the removal of elements from the source material that are not required for the planned analytical tasks. As defined in DIN 19461, deletion reduces the textual detail contained in the intermediate representation and thereby contributes to limiting the potential for reconstructing the original text. However, deletion alone does not produce a format that is legally or ethically unproblematic; rather, it functions as one coordinated step within a broader sequence of reduction operations.

In this operation, specific units of the text–such as individual tokens, multi-word expressions, sentences, or larger structural segments–are removed according to defined criteria. These criteria may be rule-based, algorithmic, or derived from annotations produced during the enrichment phase. The granularity selected for deletion must be compatible with later reduction operations, ensuring that the workflow as a whole leads toward a representation that can fulfil the legal and methodological aims of the DTF.

DIN 19461 requires that every deletion step be documented precisely. This includes a description of which elements are removed, the procedures or heuristics used to identify them, and the granularity level at which the deletion occurs. Such documentation clarifies how deletion supports the transformation of the data and how it interacts with the other operations–replacement and randomisation–that may still be necessary to ensure that the final DTF

cannot be interpreted or reverse-engineered as the original text.

5.4. Replacement

Replacement is an information-reduction operation in which selected elements of the source text are substituted with abstract or categorical representations. In DIN 19461, this operation serves to transform concrete linguistic material into forms that retain analytical value while reducing the possibility of reconstructing the original wording. As with other reduction operations, replacement does not by itself ensure that the resulting Derived Text Format (DTF) complies with legal or ethical requirements; rather, it forms part of a coordinated sequence of steps that collectively lead toward a non-reconstructible representation.

In this operation, textual units–such as tokens, multi-word expressions, or larger linguistic structures–are replaced according to explicitly defined rules. These replacements may take the form of linguistic categories (e.g., part-of-speech labels, lemma identifiers, named-entity types), structural abstractions, or numerical or vector-based representations. Replacement may operate at various levels of granularity, and the chosen level must be compatible with the intended analytical purpose as well as with subsequent reduction steps such as deletion or randomisation.

DIN 19461 requires that each replacement step be documented thoroughly. This documentation includes the description of the transformation rules applied, the models or algorithms used (for instance, tagging models or embedding frameworks), version information, and any parameter settings relevant to the operation. Such transparency ensures that the replacement is interpretable, reproducible, and assessable within the broader reduction workflow. Through this mechanism, replacement helps preserve analytical features of the text while progressively distancing the derived representation from the original content.

5.5. Randomisation

Randomisation is an information-reduction operation in which the order or internal structure of textual units is deliberately altered according to defined randomness parameters. In DIN 19461, randomisation contributes to distancing the resulting DTF from the original text by disrupting sequential patterns that could otherwise support reconstruction. As with the other operations, randomisation is not sufficient on its own to guarantee a legally or ethically unproblematic DTF; instead, it operates as one coordinated element within a broader reduction workflow.

Randomisation can be applied at various levels of granularity. At the token level, shuffling or reordering disrupts the syntax and surface structure of the text, effectively eliminating the sequence information needed to reconstruct meaningful sentences. At the sentence or paragraph level, randomising segment order breaks larger structural relationships, while still maintaining the internal consistency of each segment if required for analysis. At the document level, randomising whole-document order affects only the arrangement of documents within a collection and does not, by itself, prevent reconstruction of the content of individual documents. The effects of randomisation therefore depend strongly on the chosen granularity, and this choice must be aligned with the intended analytical purpose and with the overall reduction strategy.

DIN 19461 requires that all randomisation procedures be documented precisely. This includes specifying the units subject to randomisation, the algorithms or methods used, parameters such as random seeds or shuffle constraints, and any rules governing how randomisation interacts with preserved or enriched information. Such documentation is essential for reproducibility and for assessing how randomisation contributes to reducing reconstructive potential when combined with other reduction operations such as deletion or replacement. By systematically altering structural order at defined levels of granularity, randomisation supports the creation of DTFs that maintain analytical utility while further limiting the possibility of recovering the original text.

6. DTF Generation Workflow

The workflow is designed so that each step builds on the previous one, with enrichment establishing the informational basis and reduction operations progressively removing or abstracting content.

6.1. Enrichment Phase

Before any information-reduction steps are applied, enrichment may be used to add linguistic, structural, or statistical information to the source text. This can include annotations such as lemmas, part-of-speech categories, named-entity labels with authority links, syntactic or dependency structures, coreference chains, or disambiguation results. In addition to linguistic annotations, enrichment may also involve the extraction of statistical measures or vector-based features. These annotations may be represented in formats such as XML, TEI, or JSON, depending on the needs of the workflow and the characteristics of the data.

DIN 19461 does not prescribe specific tools, models, or annotation schemas, but it requires that all

enrichment steps be documented clearly and in detail. Such documentation must include the methods and tools used, version and parameter information, the level of application (e.g., token or sentence), and the file formats in which annotations are stored. By providing this information up front, the enrichment phase establishes a consistent and interpretable starting point for the subsequent reduction steps.

6.2. Reduction Phase and Granularity

Once enrichment is complete (or if no enrichment is required), one or more of the four core reduction operations—selective retention, deletion, replacement, and randomisation—are applied. These operations may be used individually or in combination, depending on the analytical purpose of the DTF and the legal or ethical constraints associated with the underlying material.

One aspect defined within DIN 19461 is that each reduction operation must be applied at a clearly specified level of granularity. Possible levels include individual tokens, sentences, paragraphs, documents, or entire collections. The standard provides illustrative examples showing how the choice of granularity affects both the usefulness of the resulting DTF and its resistance to reconstruction. For instance, segment-wise randomisation disrupts intra-segment structure while preserving overall grouping; selective replacement of tokens with part-of-speech categories abstracts away from lexical form while retaining syntactic patterns. These examples highlight that granularity is not an arbitrary choice but a decisive factor in the structure, utility, and safety of the resulting DTF.

6.3. Reproducibility and Metadata

To support scientific reuse, evaluation of non-reconstructibility, and long-term preservation, every DTF must be accompanied by comprehensive metadata describing the full generation workflow. This includes all processing steps, the tools and models used, segmentation decisions, random seeds (where applicable), parameters for enrichment and reduction operations, as well as software versions. Such documentation makes it possible to reproduce the DTF creation process and to assess whether the combination of reduction operations sufficiently prevents reconstruction of the source text.

The standard recommends using structured, interoperable metadata representations, such as CMDI components defined in ISO 24622, to ensure that enrichment and reduction steps are linked to the final DTF in a machine-readable manner. This linkage enables consistent archiving, facilitates dissemination across infrastructures, and allows data-holding institutions to provide transparent accounts of how a given DTF was produced.

7. Challenges in the Development of DIN 19461

The development of DIN 19461 involved navigating several challenges arising from the intersection of legal, linguistic, and technical requirements. One central difficulty was balancing these perspectives in a way that would allow the standard to be broadly applicable across institutions while remaining sufficiently precise to support reliable evaluation of derived text formats (DTFs). Legal considerations emphasised the need to avoid creating formats from which the original text could be reconstructed, whereas linguistic and technical considerations focused on preserving analytical utility and ensuring that workflows could be implemented with existing tools and methods.

Another challenge concerned the level of detail required to document enrichment and reduction operations. The standard needed to be specific enough to support reproducibility and assessment—particularly regarding parameters, granularity choices, and processing steps—while also remaining tool-agnostic so that institutions could use different software environments without deviating from the standard's requirements. Achieving this balance required careful formulation of definitions, documentation rules, and workflow descriptions.

A further challenge was clarifying how combinations of DTFs derived from the same source material should be evaluated. While individual DTFs may meet the requirements for non-reconstructibility, their combination can increase the potential for recovering information from the original text. This issue becomes particularly relevant in light of modern machine-learning capabilities, which may detect patterns or correlations across multiple representations that are not apparent in any single DTF. The standard therefore emphasises the need to assess reconstruction risks not only at the level of individual DTFs but also for sets of derived formats taken together.

8. Discussion

Benefits. DIN 19461 provides a shared terminology and a unified set of requirements for the creation and description of DTFs. By defining enrichment and information-reduction operations, specifying levels of granularity, and requiring explicit documentation of tools, parameters, and workflow decisions, the standard enables reproducible and transparent production of derived formats. This contributes not only to methodological clarity but

also to the lawful and privacy-respecting dissemination of text-based data, as institutions are supported in assessing what information may be preserved, transformed, or removed within a controlled process. The standard therefore serves as a foundational reference for data-holding institutions and research infrastructures seeking to balance analytical utility with legal and ethical responsibilities.

Limitations. While DIN 19461 establishes a structured framework for deriving text formats, it explicitly refrains from making case-by-case legal determinations. Implementers must independently assess the applicable copyright and data-protection requirements for their specific use cases. Moreover, the concept of non-reconstructibility is not an absolute property but depends on context, available auxiliary information, and the evolving capabilities of analytical tools.

As computational methods continue to advance, especially in machine learning, the risk that certain derived formats could be partially reconstructed may increase. For this reason, the standard recommends conservative combinations of reduction operations and thorough documentation to support informed assessment of residual reconstruction risks.

Open Questions. Several aspects lie beyond the current scope of the standard and remain open for future work. These include formal models for assessing reconstruction risks–particularly in the presence of modern language models—benchmark tasks or evaluation suites for validating non-reconstructibility, and the development of CMDI profiles specifically tailored to DTFs. In addition, mappings to repository schemas and guidance for integrating DTF workflows into existing archival infrastructures require further elaboration.

9. Conclusion

DIN 19461 systematises how DTFs are defined, produced, combined, and documented. By establishing a unified terminology, specifying enrichment and information-reduction operations, and outlining clear requirements for granularity, metadata, and workflow transparency, the standard provides a structured foundation for creating derived formats that can be shared lawfully and used reliably in research contexts where original texts cannot be distributed. It thereby supports data-holding institutions and research infrastructures in enabling analytical work while respecting legal and ethical constraints.

The standard also aims to foster community engagement. We encourage infrastructures, projects, and research communities to pilot DTF workflows, provide practical feedback on their applicability, and contribute to the continued development and refinement of the standard. Such collaboration will help ensure that future revisions reflect emerging needs, evolving technologies, and potentially open pathways toward broader—possibly international—alignment.

10. 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). The authors are affiliated with 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. We acknowledge the work of the DIN committee NA 105-00-06 AA "Sprachressourcen und Sprachtechnologie" and contributing institutions, who were providing initial feedback and discussion, resulting finally in the draft standard, that is expected to be finalized as a national standard in the course of 2026.

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, including notes that were created during the development of the national standard. The authors retain full responsibility for the content.

11. Bibliographical References

DIN 19461:2026-06 (E). 2026. Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF). National Standard, Deutsches Institut für Normung (DIN), Berlin.

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).

GWK – Gemeinsame Wissenschaftskonferenz. 2018. Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018. Accessed: 2026-03-24.

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

ISO 24610:2008. 2008. Iso 24610-1:2008 – language resource management – feature structures – part 1: Feature structure representation. Technical report, International Organisation for Standardization (ISO), Geneva, Switzerland.

ISO 24612:2012. 2012. Language resource management — linguistic annotation framework (LAF). 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. Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language. International Standard, International Organization for Standardization (ISO), Geneva.

Sophie Kraft, Angela Schmalen, Hendrik Seitz-Moskaliuk, York Sure-Vetter, Jennifer Knebes, Eva Lübke, and Elena Wössner. 2021. Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele. *Bausteine Forschungsdatenmanagement*, (2):1–9.

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.

Annette Preissner and Ulrich Heid. 2025. *The life of an ISO standard*, pages 427–446. De Gruyter.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007. 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.

UrhG. 2021. Gesetz über Urheberrecht und verwandte Schutzrechte. https://www.gesetze-im-internet.de/urhg/. Accessed 24 March 2026.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.