# Derived Text Formats as Strategic Transformations of In-Copyright Materials to Support Open Science: A Survey

**Christof Schöch**
Trier Center for Digital Humanities, Trier University
Universitätsring 15, 54296 Trier, Germany
schoech@uni-trier.de

**Abstract**

Derived Text Formats (DTFs) are the result of a strategic transformation of textual materials that are protected by copyright in their original form, such that the resulting data is useful for computational analyses and can be openly shared following best practices of Open Science without infringing copyright law. This paper aims to provide insights into several key aspects of this concept that is closely related to concepts such as corpus masking, non-consumptive research and extracted features. The paper establishes the motivation for using DTFs, discusses several foundational aspects of the concept and practice, describes ongoing research on issues including copyright, reconstructibility, evaluation and standardization of DTFs, and concludes with a roadmap for future work on DTFs. In this way, this paper provides a broad but concise overview of work on DTFs as a contribution to Open Science practices, with a focus on work in the Digital Humanities.

**Keywords:** Derived Text Formats, Non-Consumptive Research, Extracted Features, Copyright, Open Science, Digital Humanities

## 1. Introduction: Why do we need Derived Text Formats?

Three valuable principles contribute to shaping the regulatory framework for research today: academic freedom, Open Science, and copyright and privacy law. The legally and in some cases constitutionally-protected principle of academic freedom means, among other things, that researchers are empowered to investigate (and teach) any domain they choose to and do so in the ways they deem useful or desirable (Menand, 1996; Levy, 2026). Principles of Open Science are designed to support best practices in research, such as collaboration, transparency, accessibility, interoperability, reproducibility, and re-usability (Whyte and Pryor, 2011; Burgelman et al., 2019; Lewis, 2020). They have become increasingly central to research in the Digital Humanities against the backdrop of the 'reproducibility crisis' (Peng, 2015; Bausell, 2021; Gibney, 2022) as well as in the context of the FAIR principles for research data (Wilkinson et al., 2016). The legal framework of copyright is designed to foster innovation and creativity by granting a certain number of (moral and economic) rights to the creators of (textual or other) works (Rose, 1994; Goldstein and Hugenholtz, 2013). Finally, privacy laws are meant to enable individuals to control who is able to obtain, use, share and/or sell information regarding their identity and personal lives, and under what conditions (Bygrave, 2014; Voss, 2016), as implemented for example in the European Union's General Data Protection Regulation (GDPR, European Union, 2016).[1]

When the requirements from these three principles come together, researchers whose work relies on the computational analysis of textual or non-textual sources are forced to follow one of two strategies: Either they avoid copyright and privacy restrictions by investigating only public domain materials published before the contemporary period, that is, accept a limitation in the domains they can investigate in order to be able to practice Open Science; or they chose to investigate contemporary materials that often come with copyright and/or privacy-related restrictions, but then need to scale back their ambitions with respect to best practices of Open Science, essentially by keeping their research data locked away. DTFs are the result of a strategic transformation of in-copyright textual materials, such that these limitations are removed, thereby enabling researchers to investigate contemporary, in-copyright textual materials while both respecting copyright law and following best practices of Open Science.

This paper aims to provide a concise survey of foundational issues surrounding DTFs as well as

[1] While privacy laws are of course hugely important, this paper and most work on DTFs focuses on the copyright aspect of the issue. See, however, Greene et al. (2019) and Peloquin et al. (2020) on the implications of the GDPR for research as well as Altman et al. (2022), Joo and Kwon (2023) and Gadotti et al. (2024) on data anonymization for research in privacy-related contexts.

recent and ongoing investigations of DTFs from the standpoints of copyright law, computational research, and infrastructure development. A particular focus is placed on the fields of Computational Literary Studies, Corpus and Computational Linguistics and Natural Language Processing in the wider context of the Digital Humanities, because researchers in these fields often work with in-copyright materials, such as large newspaper corpora or extensive collections of contemporary literary texts using computational methods of text annotation and analysis also known as text and data mining (TDM), and are directly concerned by the competing requirements of Open Science and copyright law.

In section 2 of this paper, several foundational aspects of DTFs are presented: alternative approaches to providing access to in-copyright materials (subsection 2.1), terms and concepts directly relevant to the approach represented by DTFs (subsection 2.2), prominent examples of DTFs (subsection 2.3), a process-oriented definition of DTFs (subsection 2.4) as well as several typologies of DTFs (subsection 2.5). In section 3, several areas of recent and ongoing scholarly work on DTFs are discussed: the fundamentals of copyright law as they apply to DTFs (subsection 3.1, the evaluation of various kinds of DTFs for specific research questions (subsection 3.3) as well as efforts to establish regulatory, technical and infrastructure-related standards (subsection 3.4). The paper concludes with an assessment of where we stand today and what a roadmap for scholarly work on DTFs holds for the future (section 4).[2]

## 2. Foundational aspects of Derived Text Formats

Having established the need not just for the provision of access to, but also for the ability of sharing materials protected by copyright, we can now turn to several foundational aspects of strategies mobilized to achieve this.

### 2.1. Strategies for providing access to in-copyright materials

Before turning to approaches that – conceptually, if not terminologically – fall into the class of DTFs themselves, it is worth considering several alternative approaches of providing access to materials under restrictions related to copyright and/or licensing.

**Online platforms**. Many databases of textual materials, such as linguistic corpora based on newspaper articles or other contemporary sources, are available online for searching and querying. The query results are usually provided in the shape of keyword-in-context views with limited context. The set of results is sometimes confined to a random sample rather than the full set. Further limitations of this approach include that they usually provide only a fixed set of query routines, do not permit merging of corpora from multiple, independent sources, do not enable custom annotation routines on the data, hinder the deployment of custom algorithms on the data, and make transparent and open sharing of results and reproducibility difficult.

This approach is useful primarily for rather simple and predictable usage scenarios of corpora. The close intertwining of platform, data (both raw text and annotations) and analytical procedures is both a conceptual weakness (because the principle of a separation of concerns is not respected) and a considerable risk in terms of sustainability (because the underlying data cannot be made available in simple, static forms independently of the platforms that are usually rather maintenance-intensive and resource-hungry. Examples of such platforms are COSMAS of the Leibniz-Institut für deutsche Sprache, providing access to the DeReKo (Deutsches Referenz-Korpus, see Kupietz et al., 2018), the DNB's korap system with access to DeLiKo-XL (Deutsches Literatur Korpus, see Jannidis et al. (2026)), or ATILF's (Analyse et Traitement Informatique de la Lange Française) Frantext database (Montémont, 2020), each providing access to large or very large corpora of historical and contemporary language materials.[3]

**On-site access.** In this approach, access to in-copyright materials owned or licensed by a particular institution are made available to users in a closed-room scenario on the physical premises of the institution. Usually, terminals provide access to the dataset but do not permit copying or otherwise transferring the materials. In contrast to the online querying approach, the on-site access approach enables much more sophisticated research scenarios: Access to the full data is possible, deployment of custom annotation schemes and advanced analysis scenarios is possible. Again, however, it is not usually possible to combine the datasets available on-site with other, third-party datasets. The main limitation of this approach, however, is obviously the need for researchers to be physically-present at a particular site to work with the data. The same requirement is also true for all collaborators and peer reviewers, more generally implying a strong limitation on transparency, accessibility, and reproducibility of the work performed. An example of

[2] All references are available online at https://www.zotero.org/groups/6473556/.

[3] See: https://korap.dnb.de/ and https://cosmas2.ids-mannheim.de/cosmas2-web/ as well as https://frantext.fr.

an institution offering this approach is the German National Library (DNB).

**(Data / storage) capsules.** The Hathi Trust Research Center (HTRC) has developed a strategy that aims to combine the advantages of the online platform approach with those of the onsite access approach, but without the need for a researcher to actually be on site, leveraging the idea of 'storage capsules' (also called 'data capsules', initially proposed by Borders et al., 2009, see also Wang et al., 2019): "The HTRC Data Capsules provides the virtual machine with two modes: a maintenance mode during which a user can access the network and install software freely, but cannot access copyrighted data; and secure mode where copyrighted texts become accessible to the user while the network access and file system access is highly constrained" (Zeng et al., 2014, 10). The advantages are obvious, but the infrastructural and technical demands remain high compared to DTFs and similar approaches where the data can be shared freely.

**Sampling.** Another approach is pursued by the XSamples group (Andresen et al., 2023), whose fundamental idea is to leverage the fact that copyright law allows for parts (specifically, 15%) of works to be copied, used and shared in a research and teaching context even outside provisions of the Text and Data Mining exception in European copyright law (see subsection 3.1 below). Based on this idea, the authors design an infrastructure that provides the appropriate amount of samples from a dataset to any one researcher, based on queries they can run on a suitable platform.

The approaches described so far all fundamentally aim to control how users can interact with unmodified in-copyright materials. DTFs and similar approaches, by contrast, aim to provide open accessibility to and allow unconstrained interactions with data that has been modified for this purpose.

## 2.2. The terminological space around Derived Text Formats

There are a number of competing terms circulating, all fundamentally describing the same idea, but with varying focus points or from differing perspectives. The most influential ones are documented and compared in this section.

**Corpus masking**. A pioneering and foundational proposal for using transformed texts rather than document-level metadata in order to avoid copyright restrictions is corpus masking. This proposal came from Corpus and Computational Linguistics, where researchers often work with contemporary texts that, in addition to being in-copyright, are often subject to licensing contracts. Also, such corpora often include multiple layers of linguistic or other annotations provided by different people or institutions and with varying degrees of copyright protection (Lehmberg et al., 2008). After initially experimenting with Treebank corpora (Rehm et al., 2007b), the authors went on to propose a more general framework. The term corpus masking places the focus on a masking operation, where the annotation layer(s) – such as morphological or syntactical annotations – are preserved and published, but the underlying lexical content layer – that is, the corresponding word forms – is replaced by placeholder tokens (Rehm et al., 2007a, 2026).

A related approach recently proposed by Arnold and Jäschke (2026) also separates the content and (stand-off) annotation layers, but additionally relies on a partially-masked version of the original full text that is too sparse to allow reconstruction, but informative enough to allow researchers to merge publicly available annotation layers and independently-obtained in-copyright texts layers.

**(Non-consumptive / non-expressive) (use / reading / research)**. This is a family of terms used to underline the fact that in large-scale corpus analyses, whether platform-based or not, texts are not actually read and intellectually assimilated by any person, but algorithmic analysis processes are run on these texts at scale to identify complex features and patterns (Schreibman, 2014; Bhattacharyya et al., 2015; Kamocki, 2018; Samberg and Hennesy, 2019; Layne-Worthey, 2024; Baudry, 2023; Gruber and van Atteveldt, 2025; Zeng et al., 2014). This term sometimes encompasses or implies strategies such as the one discussed as 'online platforms' or 'data capsules' (subsection 2.1).

**Extracted Features**. This is an influential term mostly used in the context of the Hathi Trust Research Center's Extracted Features dataset, underlining not so much the usage but the creation of such datasets, namely by identifying and counting various kinds of features in texts – such as the number of typographical lines or the counts of all word types, per page – and making that descriptive data available (Bhattacharyya et al., 2015; Jett et al., 2020; Organisciak and Downie, 2021).

Finally, **Derived Text Formats** (in German: abgeleitete Textformate). A term introduced by Schöch et al. (2020b) that places the focus on the derivative nature of the resulting datasets, without specifying the nature of the transformation process or the kind of use being made of the data. As the most general term, this is the one preferred in this paper (see also Schöch et al., 2020a; Raue and Schöch, 2020; Genêt et al., 2025; Trippel et al., 2026; Du and Schöch, 2026a,b; Iacino et al., 2026; Ecker and Schneider, 2026.

DTFs avoid many of the limitations of the alternative approaches, in particular those connected to the need for an interactive platform. While their production and provision does have important infras-

tructural components and requirements, once DTFs have been produced and published, they are fundamentally low-maintenance, static datasets that researchers can download and work with in any way they find appropriate.

This does not mean that DTFs do not have limitations of their own. One key limitation is related to the fact that researchers do not, when working with DTFs, have access to the full, readable, original text that the DTF is based on. Currently, any DTF is designed to strike a strategic balance between the optimal obfuscation of any copyright-related features, on the one hand, and the best possible enabling of interesting and relevant research questions that can be investigated using the DTF, on the other hand. Investigating this trade-off from both a legal and a computational perspective is the object of considerable current work discussed in subsection 3.2 and subsection 3.3.

## 2.3. Existing Derived Text Formats

The basic idea of using not the original, full texts for research, but some proxy that stands in for them, is of course nothing new and may also be practiced for reasons other than copyright restrictions (such as privacy concerns or simply lack of full texts). A rather strong version of this strategy is using (document-level) metadata describing texts without considering their textual content at all. Examples include the use of rich qualitative and quantitative metadata (Paige, 2020), book titles in multiple languages (Patras et al., 2021) or library catalogue data (Fischer and Jäschke, 2022). In fact, one may argue that many DTFs (statistical DTFs, in particular) are simply very precise token-level metadata, in the sense that they consist of detailed information about the frequencies, distribution, co-occurrences or morpho-syntactic similarity of textual features, rather than the texts themselves.

A relatively early and very well-known example of a DTF is the **Google Ngram Viewer Dataset** (Ngram-Dataset, 2020), first published in 2009 (see also Lin et al., 2012).[4] The basic idea of the Ngram Dataset is to aggregate very large amounts of texts by year and then count the occurrences of ngrams of various sizes, allowing for an aggregated view of the rise and fall words and expressions over the years in the Ngram Viewer. Readability, recognizability and reconstruction are clearly excluded, because ngram data is aggregated by years and languages, not by individual works.[5]

While not primarily intended as a solution to copyright issues, the DLINA group's **Zwischenformat** (engl.: 'intermediary format') can also be understood as an early form of a DTF (Kampkaspar et al., 2015). It is suitable for dramatic texts encoded in XML-TEI and replaces the text contained in each scene by simple statistical information regarding the number of speeches and the number of words spoken by each character present in the scene. This information is sufficient for many analyses, in particular for network analysis, which often primarily relies on structural features.

The DTF most used in the context of Digital Humanities is probably the Hathi Trust Research Center's **Extracted Features** dataset (EF2.5, 2025) (see also Jett et al., 2020). It is also one of the more carefully-designed and well-documented such datasets, including through a JSON-LD schema.[6] The data is described as a "derived dataset consisting of metadata and data elements extracted from volumes in the HathiTrust Digital Library" (Jett et al., 2020). The data format is JSON, with metadata at the volume and page levels, and with descriptive statistics (e.g. number of lines and words) and token-level (word form and POS tag) frequency information encoded for each individual page. This dataset has been used to great effect by researchers in Digital Humanities (Underwood, 2014; Piper, 2022a; Sobchuk and Beheim, 2025), with a focus on investigating particular research problems, rather than explicitly or primarily evaluating performance compared to original full texts.

Recent work in the wider context of two projects, the consortium *Text+* in the framework of the German National Research Data Infrastructure (NFDI) and the research project *Forschen mit Derivaten* (engl.: 'Doing Research with Derivatives') have been concerned with three types of DTF used in various studies by the authors involved (Kocula, 2021; Du and Schöch, 2024; Du et al., 2025; Du and Schöch, 2026b):

- **Chunk-based term-document matrix**: Here, the original full text is tokenized, optionally receives some level of linguistic annotation, and is then split into chunks of equal (or nearly-equal) token sizes. The frequencies of all types in each chunk are then established and organized in a tabular format, that is a term-document matrix. The key parameter of this DTF is the chunk size.

[4] This resource became infamous in Digital Humanities circles not just because it is based on the Google Books Corpus of copyright settlement fame (Matulionyte, 2016; Borghi and Karapapa, 2011), but also because it was used in a highly ambitious and controversial paper using the Google Ngram Viewer to perform so-called 'Culturomics' (Michel et al., 2011).

[5] The Google Books Ngram Viewer is available at https://books.google.com/ngrams/.

[6] See Bhattacharyya et al. (2015); Jett et al. (2016); Downie (2015) and https://schemas.hathitrust.org/EF_Schema_v_3.0.

- **Chunk-based randomization of token order**: Again, the original full text is tokenized, optionally receives some level of linguistic annotation and is then split into chunks of equal (or nearly-equal) token sizes. However, instead of establishing frequencies, the token order is then randomized within each chunk, while the order of the chunks in the text remains intact.[7]

- **Selective replacement of word forms by POS**: To produce this DTF, the original full text is tokenized and annotated at least with the POS tag information. Then, a predefined proportion of randomly-selected word forms is replaced by their corresponding POS tag. The key parameter here is the replacement proportion (e.g. causing a slight irritation at less than 5% or massively interfering with readability at 40% or more).

There are many other datasets that can be understood as DTFs, including datasets that combine several types of DTF. One example of this latter type is the dataset about language use in pop culture (Songkorpus, 2022) published by Roman Schneider as accompanying data to his study on this domain (Schneider, 2022). This dataset contains word form and lemma frequencies, n-gram frequencies as well as a GloVe embedding model based on in-copyright materials. Another example is the CONLIT dataset (CONLIT, 2022) that contains a large range of derived data – from POS bigram and bookNLP supersense frequencies to protagonist statistics – describing 2,700 contemporary books, both fiction and non-fiction (see Piper, 2022b).

What emerges primarily from this brief overview is that researchers have shown a large amount of creativity with respect to copyright-safe and useful DTFs: Corpus annotations in XML, ngram statistics or randomized texts in large CSV files, scene-level statistics in XML-TEI, extracted features in JSON, to name a few. However, while each of these formats has its justification, each project or data provider appears to have developed their own formats and strategies, with limited concern for standardization, community consensus or re-usable pipelines (a notable exception being the HTRC's Workset Ontology; see Jett et al., 2016). This is what recent and ongoing efforts in the German DH community aim to remedy.

[7] Strictly speaking, these two DTFs contain the same information, provided that the annotations and the chunk sizes are identical. In practical terms, however, the latter format is a lot more similar to the original, annotated text.

## 2.4. Defining Derived Text Formats

What is common to the terms discussed in subsection 2.2 is that they all describe a strategy that enables work on in-copyright materials while respecting both the principles of Open Science and the rules of copyright law. Indeed, these approaches can be described as strategic transformations of original full texts that pursue a dual goal: on the one hand, removing or obfuscating any features of the texts that make them subject to copyright; and on the other hand, maintaining as much information as possible so that the resulting data remains useful for the investigation of one or several research questions.[8]

The authors of the emerging DIN standard on Derived Text Formats (see subsection 3.4 for context) describe this aspect of DTFs as follows: "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" (Trippel et al., 2026). In practice, this generally means subjecting the original texts to several procedures that can be understood as both an enrichment and a reduction of information (for details, see Schöch et al., 2020b; Trippel et al., 2026).

The **enrichment** of texts means making information explicit that is implicit in the text and understood by human readers, but may not be directly accessible to machines. For example, this could mean adding information to tokens about their lemma, their part of speech, whether or not the token is a named entity, or whether or not a token is part of direct speech (such as a quote in a newspaper article or character speech in a novel). In many cases, such enrichment can only be performed on the original full texts, because the process relies on linguistic structure and contextual information.

The **reduction** of information can take the form of (selective) retention or deletion, replacement, removal or randomization. For example, a certain proportion of tokens, or a certain class of tokens, may simply be deleted. Or they may be replaced, that is either masked (i.e., replaced by a placeholder token) or replaced by information at a different (and often more abstract) level of linguistic analysis (i.e., a word form could be replaced with its corresponding POS tag). Finally, removing information can also involve randomization, e.g. removing the sequence information for tokens by randomizing their

[8] This is in line with the initial definition of DTFs in Schöch et al., 2020b: "We propose derived text formats as a solution: here, copyrighted textual materials are transformed in such a way that copyright-relevant features are removed, but that the use of various relevant methods of TDM remains possible."

order in a document or within each of several segments of a document.

Importantly, such operations of enrichment and reduction can operate at various levels of linguistic description, in particular the character, token, n-gram, sentence, paragraph, chunk of arbitrary length, section, or work level. In addition, a given type of DTF typically has parameters, such as the proportion of tokens to be replaced by their POS tag, or the size of the chunks within which the word order is randomized.

### 2.5. Types of Derived Text Formats

Given the large range of theoretically possible (see subsection 2.4) and already existing (subsection 2.5) DTFs, researchers have proposed typologies of DTFs that help better understand the range of options available.

In a first approximation, Schöch et al. (2020b) have distinguished between token-based and corpus-based DTFs. In their typology, token-based DTFs rely on the manipulation of the original full text primarily at the level of the tokens (that can be enriched, deleted, replaced or see their order randomized), while the unit of production and publication is typically the individual document or work. By contrast, corpus-based DTFs identify, represent and/or count features such as ngrams (sequences of characters or words of a fixed length) or static word embeddings across multiple documents within or across a corpus.

Recently, Iacino et al. (2025) have distinguished three fundamental types of DTFs: statistical, transformative and Language-Model-based. Statistical DTFs involve extracting descriptive textual features (such as tokens, n-grams, typographical lines, or paragraphs) along with statistical metadata (e.g., frequency, length, or sequence). An example is the Google Ngram Viewer Dataset (Ngram-Dataset, 2020). Transformative DTFs introduce controlled noise to original texts by altering word order or replacing words with part-of-speech tags or placeholders, rendering the texts less readable while preserving structural and lexical information. An example is DTF600 (2025), containing texts with segment-wise randomized word order. Finally, language model-based DTFs utilize copyrighted texts to train models (such as topic models, word embeddings, or large language models), which encode textual information into algebraic vector spaces, enabling context-dependent semantic analysis or model fine-tuning for specific research tasks. Any language model can be considered a DTF in this sense.

Finally, Trippel et al. (2026) (in this volume) distinguish four types of DTFs: token-based (such as term-document-matrices or n-gram frequencies), vector-based DTFs (such as static or contextual embeddings), structured DTFs (such as formats based on shuffled segments) and multi-feature formats (combining several kinds of features, such as token statistics and embeddings).

It appears fair to say that the terminology is not yet entirely settled and that, as more such datasets are being published, there will be an opportunity to have another systematic view at the matter.

## 3. Recent and ongoing work on Derived Text Formats

Recent and ongoing research in particular in the Digital Humanities community has investigated legal issues around DTFs, both from a legal and an empirical perspective; has evaluated the usefulness of DTFs for specific research methods (or simply used them for research); and has concerned standardization efforts, both on a conceptual and on a technical level. This section details some of these efforts.

### 3.1. Derived text formats and copyright law

Chief among the legal issues is the question of how to determine whether a given DTF is actually 'safe' from the standpoint of copyright law, that is, whether the features that ground the copyrighted status of a text really have been removed or obscured, so that the data can safely be made openly and publicly available without infringing the copyright that protects the original, full-text version. In addition, the legal basis for producing DTFs in the first place is an important concern as well.

The introduction of the 'Text and Data Mining exception' for research into European copyright law in 2018/2019 (European Union, 2019) was an important adaptation of copyright to the digital age and a significant re-balancing between the interests of academics and those of copyright holders (Raue, 2017; Durantaye and Raue, 2020; Raue, 2022; Margoni and Kretschmer, 2022). In the absence of a doctrine comparable to 'fair use' in the USA and other jurisdictions, the TDM exception enables researchers in Europe to make copies of in-copyright materials without a limitation concerning the amount of the material, albeit under a certain number of conditions: their research is non-commercial in nature (note, however, that there is also a more narrow provision for commercial use cases); they have legal access to the in-copyright materials, for example through purchase or subscriptions; they intend to use these materials for the purposes of Text and Data Mining, as defined by law; and they do not openly share the full texts (data sharing is limited to the very narrow settings of direct project-based collaborations and quality

checks by peers). Because of this last provision, in particular, which balances the freedoms provided to researchers against the legitimate interests of copyright holders, DTFs remain a vital strategy.

However, there is another aspect of the TDM exception that is relevant to DTFs. Because the creation of a DTF requires the creation of (albeit temporary) copies of these original texts, something which is not allowed under copyright law, it is important to ensure that there are provisions that allow the creation of DTFs in the first place. As a preparatory step of TDM, similar to cleaning, structuring and annotating texts, the creation of DTFs is covered by the TDM exception, as established by Iacino et al. (2025) and Iacino et al. (2026). However, the TDM exception requires but does not create legal access to the original full texts, so this needs to be established independently from the TDM exception.

Given that DTFs are still necessary, what legal requirements are there for DTFs in order to make sure they indeed do not infringe on the copyright holders' rights? Fundamentally, we can distinguish three aspects: readability, recognizability, and reconstructibility (Grisse, 2020; Jotzo, 2020; Iacino et al., 2025, 2026): a DTF should not allow people to read (and understand or enjoy) the text in the same way that they can read the original text;[9] a DTF should not allow people to recognize or experience the original ways in which the author of the original text used language to express their ideas and their individuality; and it should not be possible, at least not with trivial effort, to reconstruct the original full text from a DTF.

In essence, DTFs need to obfuscate or remove those aspects of the original texts in which their protection by copyright is grounded (Jotzo, 2020, 129), while making sure that the resulting DTF remains useful for research. The next section introduces various strategies that have so far been used to achieve this balance, or to identify this sweet spot.

### 3.2. Reconstructibility of Derived Text Formats

As shown in subsection 3.1, one of the key criteria for a suitable DTF is reconstructibility, that is, whether or not it is possible to reconstitute the original full text from one or several DTFs, and if it is possible, what effort and expertise are required (Grisse, 2020; Iacino et al., 2025). The degree of reconstructibility varies for each different type of DTF, but also with the specific parameters that were chosen when producing a particular implementation of that type of DTF.

Arguably, the deterministic reconstitution of information that has been removed (such as words forms) from the more abstract information that has been retained (such as POS tags) appears almost impossible. Similarly, randomization of word order is, in principle, an irreversible process. However, Large Language Models (LLMs) could in principle be a game-changer for this kind of task, albeit using fundamentally probabilistic, non-deterministic approaches such as vec2text or embedding inversion (Morris et al., 2023; Zhuang et al., 2024; Seputis et al., 2025). Several studies have, as a consequence, also leveraged LLMs for attempts to reconstruct various kinds of DTFs.

Work by (Kugler et al., 2024) has investigated under which conditions full texts represented as BERT-based contextualized word embeddings, a highly relevant kind of DTF that falls into the group of Language-Model-Based DTFs (in the typology by Iacino et al., 2025), can successfully be reconstructed. They conclude that the answer depends on the attack scenario: if the encoder used to produce the DTF is available, then reconstruction becomes feasible (albeit not without significant technical expertise and time); if, however, the encoder architecture is not known (and cannot itself be reverse-engineered), then any efforts of reconstructing the original texts remain futile.

Using a somewhat different approach, Du et al. (2025) have attempted to reconstruct one kind of transformative DTFs (in the sense of Iacino et al., 2025) using LLMs. Generally speaking, their findings show that while LLMs produce text based on DTFs that is somewhat more similar to the original texts than the DTF itself, in terms of an actual reconstruction of the original texts, the results are rather disappointing.

Finally, recent work by Du and Schöch (2026b) has provided a complement to performance-oriented investigations of the degree to which reconstruction of the original full text from DTFs is feasible. The authors have instead aimed to discover typical patterns of errors that occur when the reconstruction of original texts from DTFs is attempted using generative LLMs. They conclude that such reconstructed texts are typically shorter than the original texts, that LLMs tend to generate more general phrasings with missing modifiers than found in the originals, and that even when they use the same word forms, they often put them together quite differently, resulting in largely different meaning of generated texts.

For the time being, then, it does not appear feasible to reliably and easily reconstruct coherent sections of original texts from some of the more common types of DTFs. However, being unsuccessful

[9] This is related to the commercial value of a copyrighted text: a DTF should not serve as a replacement to the commercial offers that exist, that is, it should not infringe on the author's capacity to derive an economic benefit from their creation.

in such reconstruction attempts using technologies available today is of course not the same as proving that reconstruction will not be feasible in a few years' time or that it is, in principle, impossible.

With respect to the legal status of DTFs, however, it also needs to be reiterated that a copyright infringement based on DTFs only happens if and when someone actually reconstructs significant portions of original texts outside of a research context (where the TDM exception would very likely cover such processes), not when someone publishes a DTF that, potentially and with significant effort, would enable someone else to perform such a reconstruction.

### 3.3. Evaluating Derived Text Formats' usefulness for research

Evaluating the usefulness for research of a given DTF is just as important as probing its reconstructibility. The primary way of investigating this issue is by evaluating the comparative performance of specific research methods or approaches on the original full-text versions and on one or several DTFs.

Kocula (2021) has done so using a collection of 129 British novels. When comparing topic coherence for the original corpus and several different DTFs, the author demonstrated that, as expected, LDA-based Topic Modeling, as a method based on a bag-of-words representation of texts, shows comparable performance for original full texts and DTFs based on term-document matrices or randomized word order. By contrast, performance drops substantially when using a DTF based on selective replacement of word forms by POS tags. Presumably, the lexical material becomes more and more impoverished the larger the proportion of replacements becomes.

A study by Du (2023) has used a similar setup to investigate the influence of the proportion of tokens replaced by their respective POS tag on stylometric authorship attribution. Using several corpora containing German dramatic texts, French novels, and English scientific prose, respectively, the author can show that attribution accuracy drops gradually and continuously as the proportion of POS-tags replacing word forms increases. This is good news, as an appropriate balance between usefulness for research and obfuscation of copyright-related features can in this manner be identified.

A somewhat different approach is used by Du and Schöch (2024). The authors used several DTFs based on two corpora, English-language movie reviews and German-language fiction. Rather than analyzing the performance of a classifier applied to DTFs, they used the DTFs to train a DistilBERT-based sentiment classifier and then evaluated the performance of the resulting classifiers on original full texts. Interestingly, they found a sweet spot of replacing 40% of tokens with their respective POS tags, a level where readability and reconstructibility are significantly hampered, while sentiment classification performance is essentially preserved.

Finally, the paper by Ecker and Schneider (2026) contained in this volume uses similar DTFs to investigate the accuracy of text classification using DistilBERT. Using two datasets and two perturbation strategies (POS-consistent token replacement at several rates, and word-order randomization), the results show that randomization reduced accuracy by 5%, while POS replacement reduced classification accuracy by 4–9%, with performance continuously declining as perturbation intensity increases. The authors make the highly interesting observation (in line with the findings by Du and Schöch (2024) described above) that models trained on perturbed DTF data generalize better to clean text than those trained on clean data, indicating that perturbation-based training fosters more robust representations.[10]

It becomes clear from these studies that the usefulness of a given DTF may vary widely depending on the specific method chosen, and it is important to understand the systematic relationship between the kind of method and the type of DTF (see Du and Schöch, 2026a). For example, any method that relies on a bag-of-word representation of text, such as stylometric authorship attribution, topic modeling, or certain kinds of sentiment analysis, will be entirely undisturbed by token sequence randomization, at least when it is performed chunk-wise. Conversely, any method that relies on word sequence or syntactic structure will be strongly affected by DTFs that do not preserve word order and/or syntactic annotation. Also, the usefulness of a given kind of DTF will vary strongly with certain key parameters of the DTF. The key takeaway here is that many methods are surprisingly robust against the introduction of moderate amounts of noise through replacement or randomization, though at higher rates of such noise, performance will suffer more significantly.

### 3.4. Standardization for Derived Text Formats

An important aspect of ongoing work in the context of DTFs is their standardization. This is an aspect that, as we have seen above in subsection 2.2, has largely been neglected in the early times of the

[10] This issue, however, remains an area of active research in both DH and NLP; see Eder (2013); Hill and Hengchen (2019); Aepli and Sennrich (2022); Lendvai et al. (2025).

discussion on datasets avoiding the limitations of copyright or privacy law. However, standardization of formats and documentation as well as the certification of pipelines that produce DTFs are essential: only if researchers have good reasons to trust that the DTFs they want to use contain exactly what they are supposed to contain will they be able and willing to actually use them for research.

A working group within the German NFDI consortium Text+ has recently developed a DIN (Deutsches Institut für Normung) standard for DTFs with national scope, DIN 19461:2026-04 E. The key contribution of this standard is to define concepts and a vocabulary for describing specific DTFs. It does so by specifying, on the one hand, a certain number of operations that can be performed on the original full texts to strategically transform the information contained in a DTF relative to the original full text. And it does so by defining, on the other hand, a certain number of levels of granularity at which such operations can be applied to the original text (see subsection 2.4). The key takeaway from this work is the requirement for anyone producing a DTF to document in detail all aspects of the production process in order to foster trust and reproducibility. For further details, see Trippel et al. (2026) in this volume.

Equally important are technical pipelines to derive DTFs from full texts as well as (to a lesser degree) platforms able to handle publication of DTFs. An example of a successful integration of DTFs as the output of a text processing pipeline is MONApipe (Barth et al., 2025). This is an NLP library built on spaCy (Honnibal and Montani, 2015–2026) that is designed to process literary texts in such a way that a considerable number of textual features that are of interest to scholars in Computational Literary Studies are identified and represented (e.g. temporal expressions, speakers and speeches, or events). In its latest iteration, this pipeline is also able to produce a certain number of DTFs.

It is true that the infrastructural requirements for making sets of DTFs accessible are usually not very different from publishing regular digital corpora, given that they are usually static files intended for download and offline use, not for in-platform processing. However, some specific provisions in terms of metadata and file storage are required. An example of a repository that has been enhanced to allow publishing of several types of DTFs in XML-TEI is the TextGrid Repository (Calvo Tello et al., 2024).[11]

[11] See: https://textgridrep.org/.

# 4. Conclusion

## 4.1. Where we stand today

In the past five to ten years, Derived Text Formats and related strategies have developed from relatively isolated initiatives linked to individual data collections or institutions to a well-described strategic instrument in the context of Open Science and copyright and privacy law. At the same time, this strategy has become integrated within relevant legal, technical and infrastructural contexts, chiefly in the context of the NFDI consortium Text+.

During this time, Derived Text Formats have arguably also become more necessary than ever. Practices of Open Science, for instance around the open data movement and the FAIR principles, are increasingly gaining traction in the Digital Humanities, particularly but not only in the related fields of Computational Literary Studies, Corpus and Computational Linguistics and Natural Language Processing. At the same time, concerns around copyright and research have become highly visible within the research community. Debates concerning the copyright reforms in the European Union around the years 2018/2019, the high-stakes court cases and discussions around the issue of copyright law and training corpora used for generative AI as well as the question of rights applicable to products of generative AI have received considerable media attention. In this context, solutions like DTFs are becoming more and more timely and their broad adaptation ever more urgent.

The work done so far, including the considerable work represented in this volume, certainly shows that DTFs are a robust and reliable strategy that is an important element in the Open Science toolbox supporting accessibility, transparency, reproducibility, re-usability and sustainability of datasets and results in the Digital Humanities.

DTFs have also found their way into research practices, at least to some extent: beyond the publication by institutions such as libraries of larger datasets as DTFs for research, DTFs have proven useful as a way for researchers who work on in-copyright corpora to make as much data as possible available to others, to support transparency and reproducibility. Examples include publications such as Schneider (2022); Du et al. (2022); Sperling et al. (2024). However, copyright concerns are also frequently used to justify not making any data available and there is certainly still a lot of room for increased adoption of DTFs for this kind of data sharing attached to a particular publication.

## 4.2. A roadmap for DTFs

The current state of affairs, as described in this paper, points to a number of challenges for the

future, both with respect to specific DTFs and with respect to the legal and technical environment in the age of Artificial Intelligence.

The most fundamental challenge for the next years is to foster the publication of very large sets of DTFs, for example by institutions holding large amounts of materials that are protected by copyright and are of interest for researchers in the Digital Humanities, i.e. first and foremost national libraries such as the German National Library (DNB). Making such collections available will be a game changer for research on the domains covered by these materials.[12]

Related to such institutional provision of research data, but at a different level, it is also important to continue to leverage DTFs for publication of ad-hoc datasets used in individual research papers, in order to increase their transparency and reproducibility. This is a roadmap item that primarily concerns the policy and community levels. For example, journals and conferences should be encouraged and supported in their efforts to develop data deposit policies that make use of DTFs, in order to better integrate such best practices into publication habits.

An important element of the roadmap for research into DTFs concerns reconstructibility, for several reasons: new specific DTFs are likely to emerge and consolidate from ongoing experimentation; increasing attention will likely be given to vector-based DTFs on the one hand, hybrid types of DTFs (combining properties of statistical and transformative DTFs), on the other; generative LLMs will increasingly be used for reconstruction tests and are likely to become better at this task over time, including through effects of memorization as training datasets continue to increase in coverage (often with little regard for copyright, or under broad interpretations of 'fair use', see Ahmed et al., 2026); and reconstruction from multiple DTFs of different kinds but created from the same full-text originals may become relevant. All these factors point to an increased importance of research into reconstructibility of DTFs, including developing useful measures to quantify the degree of reconstructibility of different DTFs. This area of future research includes further theoretical investigations into the very concept of DTFs, following the lead of several papers in this volume (Trippel et al., 2026; Du and Schöch, 2026a).

In this context, an important avenue of further research concerns the methods of information reduction or obfuscation. There are clear limitations to approaches relying on randomization (for example by shuffling the order of tokens within chunks of text) and/or on abstractive replacement (for example replacing a certain proportion of word forms with their respective POS tags). The degree of information reduction is often quite substantial, e.g. when 50% of the word forms are missing from a text, or when the information on token sequence becomes highly imprecise. In both these approaches, the twin goals of maximal usefulness for research and minimal reconstructibility are at odds with each other, forcing researchers into a trade-off scenario. In addition (as noted above), the feasibility of reconstruction risks will likely rise in the future, with expected further improvements in generative LLMs.

An alternative approach currently being investigated against this background at Trier University relies on a specific kind of encryption called homomorphic encryption that maintains token order and token identity while making the text unreadable to humans: it is demonstrably impossible to reconstruct the full text from such encrypted data, while the process is reversible if and only if the decryption key is known. This promises to open up entirely new avenues for both computational analyses and the interpretability of the results while effectively moving beyond the trade-off between usefulness and copyright-safety.

[12] This is all the more urgent and timely as the future of one of the largest such datasets, the HTRC's Extracted Features dataset, is currently unclear; see https://web.archive.org/web/20250503202415/https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center/.

## 5. Acknowledgements

This work has been conducted in the context of two projects: One is the National Research Data Infrastructure (NFDI) consortium *Text+* (grant number 460033370). The NFDI is funded jointly by the Federal Republic of Germany and the 16 federal states through the German Research Foundation (DFG). The other is the project *Forschen mit Derivaten* (grant number 564393508) within the DFG funding programme 'Digitalisierung und Bereitstellung (noch) rechtebewehrter Objekte'. Many thanks to all the collaborators in these two projects for the many productive discussions and activities.

## 6. Bibliographical References

Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models.

Micah Altman, Aloni Cohen, Francesca Falzon, Evangelia Anna Markatou, Kobbi Nissim, Michel Jose Reymond, Sidhant Saraogi, and Alexandra Wood. 2022. A principled approach to defining anonymization.

Melanie Andresen, Markus Gärtner, Sibylle Hermann, Janina Jacke, Nora Ketschik, Felicitas Kleinkopf, Jonas Kuhn, and Axel Pichler. 2023. Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-) nutzen. *Zeitschrift für digitale Geisteswissenschaften*, 2022(7).

Frederik Arnold and Robert Jäschke. 2026. Sharing is Caring: A Text Alignment Approach for Sharing Annotations of Copyrighted Texts. In *New Trends in Theory and Practice of Digital Libraries*, pages 135–145, Cham. Springer Natur.

Florian Barth, George Dogaru, Tillmann Dönicke, and Mathias Göbel. 2025. Infrastructures for a Community-Developed Text Processing Library. *Selected Contributions of the 5th Conference for Research Software Engineering in Germany*, 85.

Julien Baudry. 2023. Les non-consumptive research uses des ressources numériques.

R. Barker Bausell. 2021. *The Problem with Science: The Reproducibility Crisis and What to Do About It*. Oxford University Press.

Sayan Bhattacharyya, Peter Organisciak, and J. Stephen Downie. 2015. A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features. *Interdisciplinary Science Reviews*, 40(1):61–77.

Kevin Borders, Eric Vander Weele, Billy Lau, and Atul Prakash. 2009. Protecting Confidential Data on Personal Computers with Storage Capsules. In *18th USENIX Security Symposium, Montreal, Canada, August 10-14, 2009, Proceedings*, pages 367–382. USENIX Association.

Maurizio Borghi and Stavroula Karapapa. 2011. Non-display uses of copyright works: Google Books and beyond. *Queen Mary Journal of Intellectual Property*, 1(1):21–52.

Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data*, 2.

Lee Andrew Bygrave. 2014. *Data Privacy Law: An International Perspective*. Oxford University Press.

José Calvo Tello, Mathias Göbel, Ubbo Veentjer, Stefan E. Funk, Nanette Rißler-Pipka, and Keli Du. 2024. FAIR Derived Data in TEI and its Publication in the TextGrid Repository. *Journal of the Text Encoding Initiative*, 2024(18).

DIN 19461:2026-04 E. 2026. DIN 19461:2026-04 (e). Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF).

J. Stephen Downie. 2015. The HathiTrust Research Center: Providing analytic access to the HathiTrust Digital Library's 4.7 billion pages. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, page 5, New York, NY, USA. Association for Computing Machinery.

Keli Du. 2023. Understanding the impact of three derived text formats on authorship classification with delta. In *Open Humanities, Open Culture: 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd2023)*. Zenodo.

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).

Keli Du, Julia Dudar, and Christof Schöch. 2022. Evaluation of measures of distinctiveness: Classification of literary texts on the basis of distinctive words. *Journal of Computational Literary Studies*.

Keli Du and Christof Schöch. 2024. Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning. In *Digital Humanities Conference 2024: Book of Abstracts*, Lisbon. ADHO.

Keli Du and Christof Schöch. 2026a. A multidimensional constrained framework for derived text formats. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Keli Du and Christof Schöch. 2026b. Why reconstructing scrambled texts fails: A structural analysis of reconstruction outputs. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Katharina De La Durantaye and Benjamin Raue. 2020. Urheberrecht und Zugang in einer digitalen Welt – Urheberrechtliche Fragestellungen des Zugangs für Gedächtnisinstitutionen und die Digital Humanities. *RuZ - Recht und Zugang*, 1(1):83–94.

Jennifer Ecker and Roman Schneider. 2026. Multi-label text classification of derived text formats with distilbert. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Maciej Eder. 2013. Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4):603–614.

European Union. 2016. 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.

European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

Frank Fischer and Robert Jäschke. 2022. Ein quantum literatur. empirische daten zu einer theorie des literarischen textumfangs. In Fotis Jannidis, editor, *Digitale Literaturwissenschaft: DFG-Symposion 2017*, pages 777–812. Metzler, Stuttgart.

Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. 2024. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053.

Philippe Genêt, José Calvo Tello, Florian Barth, Peter Leinen, and Christof Schöch. 2025. Abgeleitete Textformate: Die Chance für Bibliotheken und Wissenschaft zugleich [slides].

Elizabeth Gibney. 2022. Could machine learning fuel a reproducibility crisis in science? *Nature*, 608(7922):250–251.

Paul Goldstein and Peter Bernt Hugenholtz. 2013. *International Copyright: Principles, Law, and Practice*, 3rd ed edition. Oxford University Press, Oxford.

Travis Greene, Galit Shmueli, Soumya Ray, and Jan Fell. 2019. Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers. *Big Data*, 7(3):140–162.

Karina Grisse. 2020. Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten. *RuZ - Recht und Zugang*, 1(2):143–159.

Johannes B. Gruber and Wouter H. van Atteveldt. 2025. Sharing is Caring (about Research): Three Avenues for Sharing (Protected) Text Collections and the Need for Non-Consumptive Research.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Matthew Honnibal and Ines Montani. 2015–2026. spacy: Industrial-strength natural language processing in python.

Gianna Iacino, Pawel Kamocki, and Keli Du. 2026. Legal implications of derived text formats – a copyright perspective. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Gianna Iacino, Paweł Kamocki, Keli Du, Christof Schöch, Andreas Witt, Philippe Genêt, and José Calvo Tello. 2025. Legal status of Derived Text Formats – 2nd deliverable of Text+ AG Legal and Ethical Issues –. *RuZ – Recht und Zugang*, 2025(3):149–172.

Fotis Jannidis, Philippe Genêt, Leonard Konle, Marc Kupietz, Steffen Martus, Carolin Müller-Spitzer, and Samira Ochs. 2026. Empirische Untersuchungen zur Gegenwartsliteratur. Das Literatur-Korpus DeLiKo@DNB und erste Analysen. In *Digital Humanities im deutschsprachigen Raum 2026*, Vienna. DHd-Verband.

Jacob Jett, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie. 2020. The HathiTrust Research Center Extracted Features Dataset (2.0).

Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data*, 2:e1.

Moon-Ho Joo and Hun-Yeong Kwon. 2023. Comparison of personal information de-identification policies and laws within the EU, the US, Japan, and South Korea. *Government Information Quarterly*, 40(2):101805.

Florian Jotzo. 2020. Der Schutz großer Textbestände nach dem UrhG – Die Nutzbarmachung fremder Textbestände für die Forschung. *RuZ - Recht und Zugang*, 1(2):128–142.

Paweł Kamocki. 2018. The argument for 'non-consumptive use' in the EU: How copyright could be redefined to allow text and data mining. In

*Intellectual Property Perspectives on the Regulation of New Technologies*, pages 237–258. Edward Elgar Publishing.

Dario Kampkaspar, Frank Fischer, and Peer Trilcke. 2015. Introducing Our 'Zwischenformat'.

Martin Kocula. 2021. Volltext vs. abgeleitetes textformat: Systematische evaluation der performanz von topic modeling bei unterschiedlichen textformaten mit python.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. InvBERT: Reconstructing text from contextualized word embeddings by inverting the BERT pipeline. *Journal of Computational Literary Studies*, 3(1).

Marc Kupietz, Harald Lüngen, Paweł Kamocki, and Andreas Witt. 2018. The German Reference Corpus DeReKo: New Developments–New Opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC.

Glen Layne-Worthey. 2024. Copyright Is the Lock; Non-Expressive Fair Use Is the Key: Research with In-Copyright Texts. In *The Routledge Companion to Libraries, Archives, and the Digital Humanities*. Routledge.

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.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Instruction Finetuning to Attribute Language Stage, Dialect, and Provenance Region to Historical Church Slavic Texts. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 654–662, Varna, Bulgaria. INCOMA Ltd.

Jacob T. Levy. 2026. Conceptualizing Academic Freedom. *Annual Review of Political Science*.

Neil A. Jr. Lewis. 2020. Open Communication Science: A Primer on Why and Some Recommendations for How. *Communication Methods and Measures*, 14(2):71–82.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic Annotations for the Google Books NGram Corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Thomas Margoni and Martin Kretschmer. 2022. A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8):685–701.

Rita Matulionyte. 2016. 10 years for Google Books and Europeana: Copyright law lessons that the EU could learn from the USA. *International Journal of Law and Information Technology*, 24(1):44–71.

Louis Menand, editor. 1996. *The Future of Academic Freedom*. University of Chicago Press, Chicago.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Véronique Montémont. 2020. De Frantext 1 à Frantext 2: La cure de jouvence d'une vieille dame. *La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen*.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Peter Organisciak and J. Stephen Downie. 2021. Research access to in-copyright texts in the humanities. In Koraljka Golub and Ying-Hsang Liu, editors, *Information and Knowledge Organisation in Digital Humanities*, pages 157–177. Routledge.

Nicholas D. Paige. 2020. *Technologies of the Novel. Quantitative Data and the Evolution of Literary Systems*. Cambridge University Press.

Roxana Patras, Carolin Odebrecht, Ioana Galleron, Rosario Arias, J. Berenike Herrmann, Cvetana Krstev, Katja Poniž Mihurko, and Dmytro Yesypenko. 2021. Thresholds to the 'great unread': Titling practices in eleven eltec collections. *Interférences littéraires/Literaire interferenties*, 25:163–187.

David Peloquin, Michael DiMaio, Barbara Bierer, and Mark Barnes. 2020. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*, 28(6):697–705.

Roger Peng. 2015. The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance*, 12(3):30–32.

Andrew Piper. 2022a. Biodiversity is not declining in fiction. *Journal of Cultural Analytics*, 7(3):768.

Andrew Piper. 2022b. The CONLIT Dataset of Contemporary Literature. *Journal of Open Humanities Data*, 8(0).

Benjamin Raue. 2017. Text und Data Mining. *Computer und Recht*, 33(10):656–662.

Benjamin Raue. 2022. Text und Data Mining in Einrichtungen des Kulturerbes – Die neuen Möglichkeiten des § 60d UrhG n.F. aus Sicht von Gedächtniseinrichtungen. *RuZ - Recht und Zugang*, 3(1):4–18.

Benjamin Raue and Christof Schöch. 2020. Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung. *Recht und Zugang*, 1(2):118–127.

Georg Rehm, Thorsten Trippel, and Andreas Witt. 2026. Revisiting masking after fifteen years: Early approaches to non-reconstructable linguistic data in the current context. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

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*, volume 1 of *NEALT Proceedings Series*, pages 127–138.

Mark Rose. 1994. *Authors and Owners: The Invention of Copyright*, 2. print edition. Harvard Univ. Press, Cambridge, Mass.

Rachael G Samberg and Cody Hennesy. 2019. Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis. In *Copyright Conversations: Rights Literacy in a Digital World*. UC Berkeley.

Roman Schneider. 2022. Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. *Sprachreport*, 38(1):38–50.

Susan Schreibman. 2014. Non-Consumptive Reading. In Naomi Segal and Daniela Koleva, editors, *From Literature to Cultural Literacy*, pages 148–165. Palgrave Macmillan UK, London.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020a. Abgeleitete Textformate: Prinzip und Beispiele. *Recht und Zugang*, 1(2):160–175.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020b. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*, 5.

Dominykas Seputis, Yongkang Li, Karsten Langerak, and Serghei Mihailov. 2025. Rethinking the Privacy of Text Embeddings: A Reproducibility Study of "Text Embeddings Reveal (Almost) As Much As Text". In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pages 822–831, New York, NY, USA. Association for Computing Machinery.

Oleg Sobchuk and Bret Beheim. 2025. Does literature evolve one funeral at a time? *Proceedings of the Royal Society B: Biological Sciences*, 292(2040):20242033.

Dorothy Henriette Modrall Sperling, Mike Kestemont, and Vincent Neyt. 2024. The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis. *Journal of Computational Literary Studies*, 2(1).

Thorsten Trippel, Florian Barth, Jose Calvo Tello, Phillipe Genêt, Piroska Lendvai, and Christof Schöch. 2026. Din 19461: A national standard for derived text formats. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Ted Underwood. 2014. Understanding Genre in a Collection of a Million Volumes. White Paper Report, University of Illinois, Urbana-Champaign.

W. Gregory Voss. 2016. European Union Data Privacy Law Reform: General Data Protection Regulation, Privacy Shield, and the Right to Delisting. *The Business Lawyer*, 72(1):221–234.

Lun Wang, Joseph P. Near, Neel Somani, Peng Gao, Andrew Low, David Dao, and Dawn Song. 2019. Data Capsule: A New Paradigm for Automatic Compliance with Data Privacy Regulations. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 3–23, Cham. Springer International Publishing.

Angus Whyte and Graham Pryor. 2011. Open Science in Practice: Researcher Perspectives and Participation. *International Journal of Digital Curation*, 6(1):199–213.

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.

Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM Workshop on Scientific Cloud Computing*, pages 9–16, New York, NY, USA. Association for Computing Machinery.

Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. 2024. Understanding and Mitigating the Threat of Vec2Text to Dense Retrieval Systems. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pages 259–268, New York, NY, USA. Association for Computing Machinery.

## 7. Language Resource References

CONLIT. 2022. *CONLIT*. Edited by Andrew Piper. Figshare. PID https://doi.org/10.6084/m9.figshare.21166171.v1.

DTF600. 2025. *600 French Novels in Derived Text Format*. Edited by Christof Schöch, Keli Du, and Julia Röttgermann. Beyond Words, 1.0. PID https://github.com/Zeta-and-Company/dtf600.

EF2.5. 2025. *The HathiTrust Research Center Extracted Features Dataset (2.5)*. Edited by John A Walsh et al. HathiTrust Research Center, 2.5. PID https://doi.org/10.13012/PXP0-F135.

Ngram-Dataset. 2020. *Google Books Ngram Viewer Dataset*. Google, v3. PID https://storage.googleapis.com/books/ngrams/books/datasetsv3.html.

Songkorpus. 2022. *Abgeleitete Textformate zu popkultureller Sprache*. Edited by Roman Schneider. Leibniz-Institut für deutsche Sprache (IDS). PID https://grammis.ids-mannheim.de/download.