Legal implications of Derived Text Formats – a copyright perspective

Ass. iur. Gianna Iacino, LL.M., Dr. iur. Pawel Kamocki, Dr. phil. Keli Du, et al.

Deutsche Nationalbibliothek, Leibniz-Institut für Deutsche Sprache, Trier Center for Digital Humanities
Adickesallee 1, 60386 Frankfurt/Germany, R 5 6-13, 68161 Mannheim, Ludwig-Weinspach Weg, 54296 Trier
g.iacino@dnb.de, kamocki@ids-mannheim.de, duk@uni-trier.de

Abstract

Text and Data Mining (TDM) methods are often used in order to analyse large amounts of text for scientific research. If the analysed text is protected by copyright, the use of such TDM methods has copyright implications. The existing copyright exceptions facilitate TDM within a narrow framework which limits the storage, publication and re-use of datasets. This paper examines the legal framework of converting the source text into a derived text format (DTF) which is no longer protected by copyright in order to allow the use of TDM without legal restrictions. First, the creation itself of a DTF is being examined: it entails copyright relevant acts which are covered by the TDM exception. In a second step the copyright status of the created DTF has to be evaluated based on three criteria: the DTF may not contain elements which are an expression of the intellectual creation of the author of the source material, the source material may not be easily reconstructable based on the DTF and the source material may not be recognizable.
Keywords: TDM, DTF, Copyright

1. Introduction

Digital Humanities research relies significantly on text corpora as foundational data. A key resarch method in this domain is Text and Data Mining (TDM). According to the legal definition in the Digital Single Market Directive (DSM Directive)[1], TDM refers to "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[2]

Conducting TDM typically requires reproducing the source material, and in collaborative settings, sharing it with others or making it publicly available. However, such actions may infringe copyright if the source material is protected by copyright law. Unless a statutory exception applies, permission from rights holders is required.

Such statutory exceptions were introduced into EU law in the art. 3 and 4 of the DSM Directive, yet significant constraints persist—particularly regarding the publication, long-term storage and reuse of datasets derived from copyrighted texts.

Under the TDM exception for scientific research (art. 3 DSM Directive), source material may only be shared within a limited group of researchers for joint research or with third parties for quality assessment. Long-term storage is permitted only if the data were collected by cultural heritage institutions, research organizations, or individual researchers affiliated with such entities. These restrictions contradict the principles of open science, which are central to Digital Humanities. They hinder the reproducibility of results, limit the ability to build on prior work, and create barriers when dealing with currently copyrighted materials.

A potential mechanism to overcome these limitations are Derived Text Formats. By transforming copyrighted source material into formats that no longer contain protected content, DTFs may enable unrestricted storage, sharing and reuse of the material. This analysis focuses on the legal conditions for creating DTFs from copyrighted works under German law, as well as the criteria for determining whether DTFs themselves are protected by copyright.

2. Copyright-relevance of deriving DTFs

Certain acts are considered copyright-relevant acts by law. The 'Right of reproduction' according to art. 16 UrhG is one of those copyright relevant acts. It means the right to produce copies of the work, whether on a temporary or on a permanent basis and regardless by which means of procedure or in which quantity they are made.

[1] 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, DSM Directive, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790.

[2] See art. 2.2 DSM Directive.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 20–24
12 May 2026.

It is necessary to take a closer look at the creation process of a DTF: needless to say, practically the only convenient and feasible way to derive DTFs is by automated means. However, a piece of software used to derive DTFs necessarily copies (large) passages of source texts, even if only copyright-free information is extracted from the text. Even Copies that are only temporarily stored in the RAM are copyright-relevant.

The reproductions necessary to derive a DTF can therefore only be lawfully made with the permission from the right holders, unless they are covered by a statutory exception. Of course, asking for the rightholder's permission to derive a DTF is not a reasonable option, primarily for pragmatic reasons (e.g., multiple and/or unknown copyright holders). And of course: the very point of deriving a DTF in the first place is to be able to share meaningful information about the text without limitations and without having to ask for permission. It is therefore necessary to rely on an existing statutory exception for deriving a DTF.

Copyright law knows several statutory exceptions which might be applicable for the purpose of creating a DTF, due to the fact that they allow reproductions of entire works. In this context the most important exception is § 60d UrhG which allows reproductions necessary for TDM for scientific purposes.

Although first introduced into German law in 2018, § 60d UrhG, containing the exception for Text and Data Mining (TDM) for scientific research purposes, owes its current wording to Article 3 of the 2019 DSM Directive. It can be summarised as follows:

- Research organisations, cultural heritage institutions and citizien scientists[3]
- are allowed to make copies of content
- that they have lawful access to
- in order to carry out TDM
- for scientific research purposes.

The process of creating a DTF entails an automated analysis of the source material in order to generate a representation of a base text. This representation contains information about the source text. Producing a DTF is very similar in nature to applying methods of Text und Data Mining, with the exception that the process is not continued as it normally would, i.e. by summarizing the transformed data, by visualizing a selection of the data, and by drawing conclusions from it. Rather, it stops at an earlier phase. This should not be an obstacle to seeing the creation of DTFs as compatible with the legal definition of TDM and the DTF itself as a result of a TDM process.

Creating DTFs can just be a means, rather than a goal, of a TDM process. In this scenario, the information incorporated in the derived text format is not the goal of the scientific research. Rather, the derived text format serves as a means to an end: it shall now be used as a copyright-free corpus for a subsequent TDM analysis. It is therefore only an interim step in a larger process. This interim step of the TDM process already serves scientific research purposes: Scientific research generally refers to the methodical and systematic pursuit of new knowledge.[4] By already deeming the "pursuit" of new knowledge sufficient, not only the steps directly related to the acquisition of knowledge are included, rather, it is sufficient that the step in question is aimed at a (later) gain in knowledge. The creation of a dataset can be considered scientific research in this aforementioned sense. While the creation of a dataset itself may not yet be associated with knowledge gain, it is a fundamental step aimed at using the dataset for future insights.[5] The TDM regulations do not explicitly provide for a multi-step TDM process, but it is nevertheless covered by the wording as well as the intent and purpose of the regulation.

Therefore, all reproductions necessary to create DTFs are exempted under the TDM exception (§ 60 d UrhG).

3. Copyright-status of DTFs

The question that arises here is whether a DTF still contains copyright-protected content from the source material. In this case, the material continues to be protected in favor of the original author and can only be used within the limitations of copyright law. However, the goal is to create a DTF which no longer entails copyright protected content and can therefore, be used freely without

[3] The **beneficiaries** mentioned in the DSM directive are limited to research organisations (including universities and public research institutes) and cultural heritage institutions (libraries, museums, archives…). Going beyond the wording of the DSM Directive, the German transposition adds to the list "citizen scientists", i.e. researchers without academic affiliation, as long as they are acting for non-commercial purposes. This addition was possible due to a creative fusion of Art. 3 of the DSM Directive with the "general" exception for non-commercial scientific research in Art. 5(3)(a) DSM DIrective.

[4] BeckOK UrhR/Grübler, 42. Ed. 1.5.2024, UrhG § 60c Rn. 5; Dreier/Schulze/Dreier, 7. Aufl. 2022, UrhG § 60c Rn. 1.

[5] LG Hamburg, Urteil vom 27.9.2024, Az.: 310 O 227/23, Rz. 113, https://openjur.de/u/2495651.html.

any constraints of copyright law. It is therefore necessary to look at the following three criteria for the determination of copyright protection.

3.1 Are partial reproductions protected?

Under EU law, a partial reproduction is only protected if it contains elements reflecting the author's intellectual creation—i.e., meets the originality threshold. In Infopaq, the CJEU ruled that even short excerpts may be protected if they are original, but individual words are not.

In Germany, courts generally reject copyright protection for very short texts—such as slogans or brief phrases—due to insufficient originality (*Schöpfungshöhe*). The 2021 BGH ruling in *perlentaucher* confirmed that short snippets (knappe Wortfolgen) are not protected, especially when the summary is sufficiently distinct from the original. The court emphasized the need for "sufficient distance" between source and summary. While 11-word snippets were deemed potentially original in *Infopaq*, this is not a legal threshold. Shorter snippets are extremely unlikely to be protected, and their exclusion from a DTF poses negligible risk. Longer snippets, however, carry higher risk.

Thus, a DTF containing original snippets constitutes a partial reproduction—requiring either permission or a statutory exception. But for most practical purposes, the risk from very short snippets is negligible.

3.2 DTFs and the "reconstructability" criterion

The real challenge lies in ensuring that compilations of snippets do not reconstruct the original work's expressive core.

According to the CJEU snippets are to be regarded as partial reproductions if they are original or if their "cumulative effect (...) may lead to the reconstitution of lengthy fragments which are liable to reflect the originality [of the source text]"[6].

This condition, which can be referred to as "reconstructability", is instructive, but difficult to apply in the realm of language technology.

Reconstruction of source texts from DTFs may in fact be a more complicated task than it appears; in one experiment it was not possible to reconstruct even a very short source text after scrambling the word order.[7] In another, the successful reconstruction of text from a specific kind of language model-based DTF (a BERT-based contextual word embedding model) depended on the availability of the encoder used to build the DTF.[8]

Whether it is possible to do so, has to be evaluated separately for each individual DTF. Due to the constantly evolving technological possibilities, the answer to the question of reconstructibility of source texts is susceptible to changing over time. It appears that with a very large amount of effort (e.g., *ad infinitum* repetition of simple trial and error), DTFs can often be used to reconstruct source material. Therefore, when applying the criterion of reconstructability to DTFs, it appears sensible to restrict it to reconstructions possible with a "reasonable effort." However, currently copyright law does not contain such a standard, and it appears that every reconstructible copy, regardless of the effort invested in the reconstruction, remains an act of reproduction.

3.3 DTFs and the "recognisability" criterion

The third factor influencing the legal status of short textual snippets was recently established by the CJEU in the 2019 *Pelham* ruling. The Court held that the use of a very short sample from a phonogram—specifically, a 2-second rhythm sequence—in another phonogram constitutes an act of reproduction, "unless the sample is altered in a way that renders it unrecognizable to the ear."[9]

Formally, this decision applies only to Article 2(c) of the InfoSoc Directive, which concerns the reproduction rights of phonogram producers. It remains unclear whether this reasoning—particularly the "unrecognizable" threshold—can or should be extended to the broader scope of Article 2, including copyright and other related rights. The CJEU might interpret the concept of "partial reproduction" in copyright law in a manner consistent with the *Pelham-decision*. However, this does not imply that the *Pelham* test—based on recognizability—would supplant the

[6] CJEU, Infopaq, para 50.

[7] Keli Du (2024), "Rekonstruierbarkeit von abgeleiteten Textformaten", https://events.gwdg.de/event/607/contributions/1408/.

[8] Kai Kugler, Simon Münker, Johannes Höhmann, & Achim Rettinger (2023), "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline", Journal of Computational Literary Studies 2(1), 1–18. doi: https://doi.org/10.48694/jcls.3572.

[9] CJEU, judgement of 29 July 2019, Pelham, Case C-476/17, ECLI:EU:C:2018:1002.

established criteria from *Infopaq*, namely originality and reconstructability.[10]

When applied to literary works, the concept of recognizability should not be understood as mere identification of a work or its metadata. Rather, it concerns the recognition of elements that reflect the author's creative expression or distinctive stylistic features. Thus, the use of a well-known literary character's name or the inclusion of publication details does not, by itself, satisfy the recognizability criterion.

4. Conclusions

Technically necessary acts of reproduction are required for the creation of a DTF. If the source material is protected by copyright, the creation of the DTF constitutes a copyright-relevant act. In the absence of the right holder's permission, the creation of the DTF must fall under a copyright exception in order to be lawful. DTFs can be created under the TDM exceptions (§ 60d UrhG). The DTFs can then be used as a basis for a subsequent analysis.

Whether the DTFs can also be *used freely*, made publicly available and stored for an unlimited amount of time, depends on the copyright status of these DTFs. The goal of creating a DTF which no longer contains copyright-protected content is achieved, if

- the DTF does not contain elements which are an expression of the creative individuality of the author of the source material,
- the source material cannot be reconstructed with trivial effort and
- the author's creative individuality is not recognizable.

There are many grey areas in examining the legal status of the many different forms of DTFs. In many cases, legal certainty cannot be achieved. By avoiding reproducibility and recognisability of source texts e.g. through avoidance of longer n-grams.

[10] M. Senftleben, "Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham", IIC, 2020, 51, pp.751 – 769. See also K. Grisse, "Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten", Recht und Zugang, 2020, 2, pp. 143–159.

5. Bibliographical References

Karina Grisse, “Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten”, Recht und Zugang, 2020, 2, pp. 143–159.

Kai Kugler, Simon Münker, Johannes Höhmann, & Achim Rettinger (2023), “InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline”, Journal of Computational Literary Studies 2(1), 1–18. doi: https://doi.org/10.48694/jcls.3572.

BeckOK UrhR/Grübler, 42. Ed. 1.5.2024, UrhG § 60c Rn. 14.

Dreier/Schulze/Dreier, 7. Aufl. 2022, UrhG § 60c Rn. 1.

Keli Du (2024), “Rekonstruierbarkeit von abgeleiteten Textformaten”,

https://events.gwdg.de/event/607/contributions/1408/.

M. Senftleben, “Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham”, IIC, 2020, 51, pp.751 – 769.