Urheberrecht

Ass. iur. Gianna Iacino, LL.M., Dr. iur. Paweł Kamocki, Dr. phil. Keli Du, Prof. Dr. Christof Schöch, Prof. Dr. Andreas Witt, Philippe Genêt and Dr. José Calvo Tello*

Legal status of Derived Text Formats

– 2nd deliverable of Text+ AG Legal and Ethical Issues –

This document was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

I. Introduction

A key aspect of many Digital Humanities projects is the use of texts as research data. Text and Data Mining (TDM) is an umbrella term for a range of methods

* Ass. jur. Gianna Iacino, LL.M., specialised in media law and works at the law department of the German National Library.
Dr. iur. Paweł Kamocki is a Legal Expert at the Leibniz-Institut für Deutsche Sprache, Mannheim, co-chair of the Text+ Working Group on Legal and Ethical Issues, and chair of the CLARIN Legal and Ethical Issues Committee
Dr. phil. Keli Du is a PostDoc researcher in Computational Literary Studies at the Trier Center for Digital Humanities, Trier University, Germany.
Prof. Dr. Christof Schöch is Professor of Digital Humanities and Co-Director of the Trier Center for Digital Humanities at Trier University, Germany.
Prof. Dr. Andreas Witt is Professor of Computational Humanities and Text Technology at the University of Mannheim, Head of the Department of Digital Linguistics at the Leibniz Institute for the German Language in Mannheim, and Spokesperson of the Text+ consortium within the German National Research Data Infrastructure (NFDI).
Philippe Genêt works at the German National Library and coordinates the Task Area Collections in the consortium Text+ oft he German National Research Data Infrastructure (NFDI).
Dr. José Calvo Tello works as a researcher and subject librarian at the Göttingen State and University Library.

 DOI: 10.5771/2699-1284-2024-3-149

used to analyse texts for scientific research. According to the legal definition in the Digital Single Market Directive (hereinafter: the DSM Directive),[1] TDM is "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] To conduct TDM, it is necessary to reproduce the source material, and in collaborative research projects also to communicate it to the public. Such acts are copyright-relevant if the source material is protected by copyright. In such cases, performing TDM requires the authorisation of the rights holders unless a statutory exception applies. With the DSM Directive, new copyright exceptions regarding TDM have been introduced into the EU legal framework (see art. 3 and 4 DSM Directive). Still, TDM encounters limitations concerning the storage, publication, and re-use of datasets derived from copyrighted texts: According to the TDM exception for scientific research, the source material may only be shared with a limited circle of persons for joint scientific research or with third persons for quality evaluation purposes. It can only be stored long-term if it was collected for research purposes by cultural heritage institutions, research organisations or individual researchers belonging to a research organisation.[3] Such limitations, however, run counter to the principles of open science in research which, like in many other fields, play an important role in Digital Humanities and make it difficult to replicate or verify the results of existing studies, or to build on earlier work when current, in-copyright materials are concerned.

This paper will focus on Derived Text Formats (DTFs) as a possible way to avoid such limitations by using statutory exceptions to transform the source material into formats which no longer contain copyrighted content. It will discuss the legal requirements to create DTFs from copyright-protected material, as well as the legal criteria to determine the applicability (or not) of copyright to DTFs according to German law. Copyright law is heavily influenced by EU directives, so these will play a significant role throughout this analysis.

II. B. What are derived text formats (DTFs)

DTFs have been described as extracted features for non-consumptive research.[4] They are systematically generated representations of a base text, which allow the application

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.

3 For a comprehensive analysis of the TDM exceptions in the DSM Directive, their transposition into German law and the limitations concerning storage, publication and re-use of datasets, see: G. Iacino, P. Kamocki, P. Leinen, Assessment of the Impact of the DSM-Directive on Text+, https://zenodo.org/doi/10.5281/zenodo.12759959.

4 See e.g., Y. Lin, J.-B. Michel, E. Aiden Lieberman, J. Orwant, W. Brockman, S. Petrov, Syntactic Annotations for the Google Books NGram Corpus, in: Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, Jeju Island, Korea,

of specific TDM methods. They can be produced in such a way that, on the one hand, the resulting representation still allows for the application of at least one research method, while, on the other hand, the representation is no longer protected by copyright.

The basic idea behind DTFs is to selectively remove specific pieces of information, particularly copyright-relevant features, from in-copyright texts to transform copyright-protected material into DTFs which no longer contain copyright-relevant features and therefore, are no longer affected by copyright restrictions. Additionally, if the material cannot be 'humanly' read and understood in order to intellectually assimilate its content (Baudry, 2023),[5] making it available to the public is less likely to affect the interests of copyright holders. At the same time, such materials are still suitable for a variety of TDM tasks in the Digital Humanities, such as simple quantification of words and other linguistics features, stylometry and authorship attribution, topic modelling, or the training of machine learning models.

There are many ways to create DTFs from source texts, but they can be roughly divided into three groups[6] listed below, with examples for reference.

1) **Statistical DTFs:** The first method is to extract textual features from texts (e.g. tokens, lemmata, n-grams, sentences, lines, paragraphs or pages) and their corresponding statistical information, such as length, absolute/relative frequency or sequence. Such extracted, descriptive information can then be published for text analysis tasks. Examples of such DTFs are the "Hathi Trust Extracted Features" (see e.g., Jett et al. 2020, Organisciak et al. 2017, Parulian et al. 2022) and the "Google Books Ngram Datasets" (see e.g., Michel et al. 2011, El-Ebshihy et al. 2018, Richey & Taylor 2020).
2) **Transformative DTFs:** The second method and the idea is to artificially add some "noise" to the original text, which reduces its readability (Schöch et al. 2020). More precisely, different kinds of transformation are being applied to the source

2012, pp. 169–174, https://aclanthology.org/P12-3029; S. Bhattacharyya, P. Organisciak, J. S. Downie, A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features, Interdisciplinary Science Reviews 40 (2015) 61–77, http://www.tandfonline.com/doi/; J. Jett, B. Capitanu, D. Kudeki, T. Cole, Y. Hu, P. Organisciak, T. Underwood, E. Dickson Koehl, R. Dubnicek, J. S. Downie, The HathiTrust Research Center Extracted Features Dataset (2.0), 2020, https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId= 79069329, doi:10.13012/R2TE-C227; C. Schöch, F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. Hinzmann, J. Röpke, Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen (2020), http://zfdg.de/2020_006. doi:10.17175/2020_006; P. Organisciak, J. S. Downie, Research access to in-copyright texts in the humanities, in: Information and Knowledge Organisation in Digital Humanities, Routledge, 2021, pp. 157–177.

5 J. Baudry (2023), Non-consumptive research use, an analysis of the legal situation, on Couperin.org, https://www.couperin.org/le-consortium/actus/non-consumptive-research-use/.

6 Some authors distinguish between "token-based" and "vector-based" DTF, see Schöch et al. 2020, F. Barth, J. Calvo Tello, K. Du, P. Genêt, L. Keller, J. Knappen, Liste der Abgeleiteten Textformate (working title), forthcoming, 2025.

texts (e.g. removing the sequence information by randomizing the order of the words and/or randomly replacing a certain proportion of words in texts with their corresponding part-of-speech tags or with a placeholder token) and the transformed texts can then be published in different formats (plain text, JSON, XML, tabular formats) as research data.[7] One recent example of this can be found in the TextGrid-Repository, with TEI files containing metadata, structure and lexical information but with the tokens in randomized order (Calvo Tello et al. 2025).[8]

3) **Language model-based DTFs:** The third method is to train a language model using the copyrighted texts and publish the model (e.g. topic model, static / contextualized word embedding model, or large language model). In this way, the information contained in texts (including for example the frequency and the context of words) is mapped to and represented in an algebraic vector space instead of the usual symbolic character system. Information in this format could be used for e.g context-dependent semantic analysis of individual words or fine-tuning of the large language models for specific analysis procedures[9].

It should be noted that DTFs are usually published along with information about the texts, such as data about the structure of the text (if it contains paratexts, verses, images, etc.) and metadata of various kinds, such as details of the composition of the files, title and other descriptive information about the work(s) it contains, the author and other agents (such as publishers, translators) involved in the process, etc. This is the case for Statistical and Transformative DTFs and most of the examples mentioned above and in Schöch et al. (2020). Even in cases with poorer metadata, such as Google N-grams, some metadata is needed to be able to use them for analysis. As Burnard points out, "without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity" (2004). This is an important feature of textual data, and could have important implications for the criteria for copyright status of DTFs.

7 Note that for a number of kinds of statistical DTFs, a conversion into transformative DTFs with identical information content is possible, and vice versa, so that these two types are not necessarily fundamentally different.

8 https://textgridrep.org/search?query=&order=relevance&limit=20&mode=list&filter=format:application%2Fxml%3Bderived%3Dtrue&filter=project.id%3ATGPR-8b44ca41-6fa1-9b49-67b7-6374d97e29eb.

9 See e.g. C. Schöch, F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. Hinzmann, J. Röpke, Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen (2020), http://zfdg.de/2020_006,. doi:10.17175/2020_006; Hessel, J., & Schofield, A. (2021, August). How effective is BERT without word ordering? implications for language understanding and data privacy. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 204–211); Keli Du & Christof Schöch: "Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning" (long presentation). In: Karajgikar, J., Janco, A., & Otis, J. (2024). DH2024 Book of Abstracts. Zenodo, https://doi.org/10.5281/zenodo.13761079.

The way in which DTF's metadata is published is closely related to the format chosen. For plain text files, metadata is often stored separately in a tabular format or in the name file in a precarious way. Other formats, such as JSON or XML-TEI, allow metadata to be encoded in a structured way, facilitating better compliance with the FAIR principles (Wilkinson et al. 2016).

III. Copyright implications of creating DTFs

This section will discuss in which cases creating DTFs from copyright-protected source material qualifies as a copyright-relevant act and therefore can only be carried out with authorisation of the rights holders or under a statutory exception.

1. Copyright protection of the source material

This deliverable concerns only the creation of DTFs from copyright-protected source material. The reader should be aware, however, that while a great majority of texts (novels, poems, song lyrics, news, blog posts, letters, diary entries…) meet the originality threshold required for copyright protection, some texts are copyright-free (i.e., in the public domain). This is the case if:

- Copyright has reached its term, which generally happens 70 years after the death of the author[10] OR
- The text is expressly excluded from copyright by a statutory provision; under German law, this is the case of "official works";[11]
- The text fails to meet the originality threshold, e.g. because it's too short for the originality (German: persönliche geistige Schöpfung) to manifest itself[12] (some tweets, slogans), or because it's very commonplace (e.g., consists of expressions commonly used in the given context[13], such as "I wish you a merry Christmas and a happy new year"), or because the author had to follow some strict formal constraints (this could be the case of some product descriptions or user manuals).

Public domain texts can be freely copied and shared, so there is limited practical interest in deriving DTFs from them (at least in the context of copyright concerns); but since they can also be freely modified, deriving and sharing DTFs from such texts is not in any way restricted by copyright law.

10 See § 64 Urheberrechtsgesetz (UrhG) for the German legal framework, https://www.gesetze-im-internet.de/englisch_urhg/.
11 See § 5 UrhG.
12 The Court of Justice of the European Union ruled that texts as short as 11 consecutive words can be protected by copyright; however, very short texts (around 4 words and shorter) are generally regarded as too short to be protected by copyright, see CJEU, judgement of 16 July 2009, Infopaq, Case C-5/08, ECLI:EU:C:2009:465, https://curia.europa.eu/juris/document/document.jsf?docid=72482&doclang=EN.
13 See e.g. LG München I, Urteil vom 12.12.2017, 33 O 15792/16, https://openjur.de/u/970910.html.

2. Copyright-relevance of creating DTFs

Two scenarios need to be distinguished. On the one hand, if the outcome of the creation process, the DTF itself, still contains copyright protected-source material, the creation of the DTF necessarily entailed reproductions of the source material, independently from how it was created. On the other hand, if the DTF no longer contains copyright-protected source material, the creation process might either not have entailed any acts of reproduction or might have required (technically necessary) acts of reproduction. Therefore, It is necessary to take a closer look at the creation process itself before analysing whether the result of the process, the DTF itself, is still protected by copyright. As it will be demonstrated, a distinction can be made between manually and automatically creating DTFs.

a) Manually deriving DTFs

Theoretically, one could derive DTFs, especially Statistical DTFs, entirely manually, e.g. by manually counting words, characters, etc., and then presenting the results in a chart or a diagram. Although certainly time-consuming and inefficient in the digital age, manual derivation of Statistical DTFs possibly can be a copyright-irrelevant act, if the source text is not reproduced (copied) in the process, and only pure copyright-free information is extracted from the text (e.g., there are 3 occurrences of the word "kerfuffle" in the text, the text consists of 772 sentences, the most common sentence length is 5 words, etc.). However, this is of course a time-consuming and error-prone process that does not scale well to larger amounts of text; it is therefore of very limited practical relevance.

b) Automatically deriving DTFs

Needless to say, the most 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. It is interesting to realise that an action that would "normally" be free from constraints may fall within the ambit of an exclusive right just because it is performed with the aid of information technology.

Copies that are only temporarily stored in the RAM are also copyright-relevant.

The wording of Article 2 of the InfoSoc Directive[14] and of § 16(1) UrhG, defining the exclusive right of reproduction, leaves no doubts as to the fact that all acts of reproduction, even technically necessary acts of temporary reproductions, are copyright-relevant:

14 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, InfoSoc Directive, https://eur-lex.europa.eu/eli/dir/2001/29/oj.

Article 2 InfoSoc Directive:

Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part: (a) for authors, of their works (...).

§ 16(1) UrhG:

'Right of reproduction' means the right to produce copies of the work, whether on a temporary or on a permanent basis and regardless of by which means of procedure or in which quantity they are made.

3. Creating DTFs under statutory exceptions

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). 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.

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 the context of this deliverable, two of these exceptions deserve a closer look: § 60d UrhG which allows reproductions necessary for TDM for scientific purposes and § 44a UrhG which allows temporary acts of reproduction. Other statutory exceptions, such as § 51 UrhG (citation) or § 53 UrhG (private copy) are too narrow in scope. Even § 60c UrhG which is the specific statutory exception for scientific research, generally only allows reproductions of 15 % of a work, with entire reproductions being allowed only in very limited cases (for so-called "small-scale works" or out-of-commerce works).

When interpreting statutory exceptions, the so-called three-step test must be considered, before turning to a more detailed examination of the potentially relevant statutory exceptions for creating DTFs.

a) Three-step-test

According to Article 9(2) of the Berne Convention, signatory countries can adopt exceptions to the reproduction right in their national laws, provided that these exceptions are limited to "certain special cases", and that the allowed uses "[do] not conflict with a normal exploitation of the work and [do] not unreasonably prejudice the legitimate interests of the author[s]". This provision, called the "three-step test", was first introduced during the 1967 revision of the Convention. Today, it can also be found in Article 13 of the TRIPS Agreement, Article 10 of the WIPO Copyright Treaty, and in Article 5(5) of the InfoSoc Directive.

The three-step test is the first "bottleneck" that every copyright exception must go through before it can be enacted in national law.

The first step of the test, i.e. limitation "to certain special cases" signifies that the exceptions must be clearly defined ("certain") and limited in scope and reach ("special"). For example, an exception allowing for all works to be used without permission from the rightholder for "personal development" purposes would fail this test, as it is not clearly defined (what exactly is "personal development"? Does it cover only spiritual, or also material enrichment? Whose "personal development" – only the user's or also his or her public's?, etc.) and above all it is too broad in scope (not "limited"); in fact, it could strip copyright of any practical significance. This also means that exceptions need to be periodically reviewed and adapted to the changing reality.

The second step of the test requires that copyright exceptions "do not conflict with a normal exploitation of the work". This chiefly means that the exceptions should not allow users to enter into economic competition with the rightholder. All uses that have economic importance, or are likely to acquire such importance, should in principle be reserved to the rightholders.

Thirdly, the exceptions should not "unreasonably prejudice the legitimate interests of the right holder". The ambiguity of the term "legitimate" makes this step particularly hard to evaluate, as it can mean both "lawful" (i.e., grounded in the law), and, more broadly, "valid" (as in: justifiable, reasonable). It seems that the narrower view, limited to protecting interests enshrined in the law, is more appropriate. Copyright exceptions should not "unreasonably prejudice" such legitimate interest. According to the WTO Panel, this precludes exceptions that cause or may cause "unreasonable loss of income" to the rightholder[15]. However, a contrario, a certain degree of prejudice ("reasonable prejudice", related to a "reasonable loss of income") remains acceptable. This is why some exceptions (such as the private copy exception) are accompanied by a compensation scheme (such as a special levy (tax) on purchases of recordable media) that is supposed to bring the prejudice suffered by rightholders back to an acceptable level.

As demonstrated, the three-step test offers considerable freedom of interpretation and "sufficient breathing space for social, cultural and economic needs" (Senftleben, 2010)[16]. However, the views on the practical significance of the three-step test differ. In particular, the way in which the test is incorporated in the InfoSoc Directive indicates that its role, according to the European legislator, is to "narrow down" the scope of statutory exceptions: Article 5 in its (1), (2) and (3) offers a list of rather narrowly defined uses that should (Article 5(1)) or may (Article 5(2) and (3)) be covered by statutory exceptions in national laws of the Member States. This list

15 WTO, Report of the Panel, WT/DS160/R, 15 June 2000, para 6.229.

16 M. Senftleben, The International Three-Step Test: A Model Provision for EC Fair Use Legislation, 1 (2010) JIPITEC 67, para. 1, https://www.jipitec.eu/archive/issues/jipitec-1-2-2010/2605/JIPITEC%202%20-%20Senftleben-Three%20Step%20Test.pdf.

includes, for example, quotations, uses in research and teaching, as well as the private copy exception (see below). Then, the same article 5 (in its (5)) provides that all the above exceptions (whether affecting only the reproduction right, or also the right of communication to the public) should *in addition* be subject to the three-step test. As a result, exceptions stemming from the InfoSoc Directive tend to be defined in a very narrow way, which severely limits their practical significance, especially for research activities.

National laws also seem to attribute different roles to the three-step test. The German Copyright Act does not mention the three-step test, which indicates that it is a "high level" requirement stemming from international law and addressed primarily to the national legislator, and not to those who apply the law. Simply put, if an exception makes its way into the German Copyright Act, this means that the legislator assumes it has passed the three-step test and the role of the test is limited to serving as an aid in the interpretation of German copyright regulations (unless, of course, an international body such as the WTO Panel decides otherwise and forces the national legislator to reform the law).

Some legal scholars advocate that the three-step test should be elevated to the rank of an "open norm" in the EU, similar to the American fair use doctrine[17]. In this approach, every use that passes the three-step test should be allowed, and the existing exceptions should only serve as examples of "specific cases". Although this approach is tempting and may seem justified, this deliverable is not a place for a debate on law reform.

b) Exception for TDM for Scientifc research purposes (§ 60 d UrhG)

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:

17 See esp. M. Senftleben, The International Three-Step Test: A Model Provision for EC Fair Use Legislation, 1 (2010) Journal of Intellectual Property, Information Technology and E-Commerce Law, Vol. 1, No. 2, pp. 67–82; see also Gervais, Daniel J., Towards a New Core International Copyright Norm: The Reverse Three-Step Test, SRRN, https://ssrn.com/abstract=499924.

- **Research organisations**, **cultural heritage institutions** and **citizien scientists**[18]
  - are allowed to make **copies** of content
  - that they have **lawful access** to
  - in order to carry out **TDM**
  - for **scientific research purposes**.

The question of whether the creation of DTFs is covered by the TDM exception will be addressed below. An analysis of § 60d beyond this question will not be conducted at this point.[19]

TDM in German copyright law is defined in § 44b UrhG very broadly as an "automated analysis of individual or several digital or digitised works for the purpose of gathering information, in particular regarding patterns, trends and correlations" (§ 44b, very close, but not identical to the definition in art. 2 of the DSM Directive).

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. For example, a Statistical DTF can be seen as gathering information about patterns of occurrence (e.g., the frequency of use of certain words or expressions), and language-model based DTFs can be seen as representing morphological and semantic information about tokens that are based on patterns of co-occurrence and stored as dense numerical vectors. In fact, producing a DTF is very similar in nature to applying methods like topic modelling or stylometry, 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 be a goal of a TDM process. For example, finding any keyness analysis supposes, in a first step, to collect the frequency and/or dispersion information about all the different words found in a corpus, which is essentially information about how the frequency of a given word evolved over time may require creation of Statistical DTFs. Certain research questions may also entail creation of Transformational DTFs, for example:

1. In order to compare the semantic similarity of words in texts, a static word embedding model can be trained and then applied for the comparison.

18 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.

19 For a more detailed analysis of § 60 d UrhG, G. Iacino, P. Kamocki, P. Leinen, Assessment of the Impact of the DSM-Directive on Text+.

2. In order to explore the hidden thematic structure of a corpus, a topic model can be trained and the text data is transformed to two probability distributions: a topics-in-documents distribution and a words-in-topics distribution, which can be used to analyse the occurrence of semantic/thematic word clusters in text collections.

But, deriving DTFs can also 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.[20] 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.[21] 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.

Whether a goal in itself or merely a means towards another goal, all reproductions necessary to create DTFs are exempted under the TDM exception (§ 60 d UrhG). Importantly, German law also expressly allows modifications of the source material when it's technically necessary to carry out TDM. Necessary changes of format or encoding of the source material for building DTFs are therefore expressly allowed (§ 23 (3) UrhG).

c) Possible alternative: Temporary acts of reproduction (§ 44 a UrhG)

Technically necessary acts of reproduction, such as those made in the process of automatically creating DTF, might be covered by § 44a UrhG.[22]

According to the CJEU[23], an act of reproduction can be covered by the exception for temporary acts of reproduction (Art. 5.1 of the InfoSoc Directive, § 44a UrhG) if it meets 5 cumulative conditions:

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

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

22 See Recital 9 DSM Directive.

23 CJEU, judgement of 26 April 2017, Stichting Brein, Case C-527/15, ECLI:EU:C:2017:300: https://curia.europa.eu/juris/document/document.jsf?text=&docid=190142&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=8019316.

(1) the copy is temporary (i.e., not intended to last),
(2) it is transient or incidental (i.e., its lifetime is limited to what is necessary, and the copy is then deleted automatically without human intervention[24]),
(3) it is an integral and technical part of a technological process (the reproductions have to happen within and not outside the process; the process can be manually activated[25]),
(4) the sole purpose of that process is to enable a transmission in a network between third parties by an intermediary or a lawful use of a work or subject matter (i.e., a use that was authorised by the rightholder or is not restricted by copyright law[26]), and
(5) that act does not have any independent economic significance (i.e. the copies do not enable the generation of an additional profit going beyond that derived from the lawful use of the protected work and do not lead to a modification of that work[27]).

Classically, this exception is invoked mostly in the context of web browsing (to exempt the cache copies made in the process from the obligation to obtain authorisation).

The first four requirements are unquestionably met in the creation of derived text formats. The necessary acts of reproduction are a technical and integral part of the automated process of creating a DTF. They will be automatically deleted at the end of the creation process. If the DTF itself no longer contains copyright-protected content, then the reproductions made while creating the DTF do not manifest in the DTF and these reproductions are therefore, temporary and incidental.

Since a use should be considered lawful where it is authorised by the rightholder or not restricted by law, the key question it boils down to is: is creating DTFs restricted by copyright law? As already discussed above, creating a DTF is allowed under the TDM exception and therefore, not restricted by copyright law.

It seems particularly unclear whether copies made in the process of deriving DTFs meet the 5th condition, the "lack of independent economic significance" – a DTF may indeed have such significance (cf. above about the second step of the three-step test). While in the academic context there is no intent to enable additional profit for researchers, rightholders might argue that the creation of DTFs disables additional profit on their side.

Some rights holders (in particular in news and scientific domains) are likely to argue that DTFs may indeed harm their business models, inasmuch as these involve selling data for Large Language Models training (cf. e.g., Le Monde's partnership with Open AI), or providing dedicated TDM platforms (cf. Elsevier's). Potentially, this

24 CJEU, Infopaq, para 64.
25 CJEU, order of 17 January 2012, Infopaq II, Case C-302/10, ECLI:EU:C:2012,16,para 39, https://curia.europa.eu/juris/document/document.jsf?text=&docid=118441&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=8021836.
26 Recital 33 of the InfoSoc Directive; CJEU, Infopaq II, para 42.
27 CJEU, Infopaq II, para 54.

may indeed be an obstacle to making DTFs under the exception for temporary acts of reproduction, or even fail to pass the three-step test (which would prevent it also under other exceptions). This risk seems much less present in case of literary works that are clearly intended for close reading (e.g. novels, short stories etc.) and where mining presents limited commercial advantages, compared to scholarly literature (e.g. in medicine) where mining may yield commercial benefits.

It seems that even though the use of § 44a UrhG to enable the creation of DTFs cannot be completely excluded, it is not necessary to rely on it. The requirements of § 44a UrhG regarding the creation of a DTF are only fulfilled when and if the requirements of a TDM exception (§ 60d UrhG or § 44b UrhG) are fulfilled as well. Therefore, relying on § 44a UrhG for creating DTFs is redundant. Indeed, the TDM exceptions (see above) provide a much more convenient and secure basis for this activity.

4. DTF and the integrity right

Although rarely discussed in the context of new technologies, moral rights of authors should not be absent from the discussion on DTFs. The so-called integrity right is recognised by Article 6bis of the Berne Convention. In Germany, it can be found in § 14 UrhG. The texts read as follows:

Article 6bis(1) of the Berne Convention:

(1) Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to (...) object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation.

§ 14 UrhG:

The author has the right to prohibit the distortion or any other derogatory treatment of his or her work which is capable of prejudicing the author's legitimate intellectual or personal interests in the work.

It seems possible that some DTFs, in particular Transformative DTFs with shuffled sentences or added noise, can be perceived by the author (or his or her heirs) as prejudicing his or her legitimate interest, or even to his or her honor or reputation. This could be the case e.g. when sentences or paragraphs are shuffled, which distorts the message of the work, or when added noise deteriorates the artistic quality of the work. Theoretically, the recipient of such Transformative DTFs could then have a false impression about the opinions of the author or his or her writing skills.

On the other hand, Transformative DTF are not intended for close reading, and therefore are not to be read and understood by humans. It can therefore be argued that such DTFs are not meant to give any impression about the opinions or writing skills of authors of source texts.

Nonetheless, transformative DTFs might carry a non-negligeable risk of infringing the integrity right.

IV. 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 any constraints of copyright law. Therefore, this section will address the question of the criteria by which a text can be assessed as (still) being protected by copyright. Additionally, this section will determine whether the DTF itself meets the originality threshold and might therefore be protected by copyright in favor of the creator of the DTF.

1. DTFs as (partial) reproductions of source texts?

The very broad definition of the right of reproduction in Article 2 of the InfoSoc Directive (see above) can easily be reduced *ad absurdum*. Interpreted strictly, *partial reproduction* would include a situation in which only the smallest meaningful part of a work (in case of literary works: one word of text) is copied. This, however, would mean that individual words are protected by copyright, which would not only be unworkable in practice, but also devastating for freedom of (artistic) expression. In order to avoid this, the CJEU was confronted with the task of defining the limits of partial reproduction.

In the 2009 Infopaq case, the CJEU ruled that parts of works should receive protection under the reproduction right (Article 2(a) of the InfoSoc Directive) as long as they "contain elements which are the expression of the intellectual creation of the author of the work".[28] In other words, under EU law, extraction of a snippet from a copyright-protected text is a copyright-relevant act (reproduction in part) only if the snippet meets the originality threshold (see section 1) below).

In the same case, the Court also ruled that in certain contexts, the use of unoriginal snippets of a literary work may still amount to an act of reproduction when "the cumulative effect of those extracts may lead to the reconstitution of lengthy fragments which are liable to reflect the originality of the work in question"[29]. This condition can be referred to as "reconstructability" – section 2) below)

More recently, in the Pelham case, the CJEU ruled that the use of a very short sample from a phonogram (in the facts of the case: an approximately 2-second rhythm sequence) in another phonogram is an act of reproduction, "unless that sample is included in the phonogram in a modified form unrecognizable to the ear"[30]. Although it is not clear whether and how this "recognisability" condition applies to literary works (texts), it is presented and discussed in section 3) below.

28 CJEU, Infopaq, para 39.
29 CJEU, Infopaq, para 50.
30 CJEU, Pelham, 29 July 2019, para 39.

a) DTFs as potentially containing "partial reproductions" of works

It is crucial to determine whether in the light of CJEU's and national case law, the snippets of source material that are still present in a DTF meet the originality criterion, and therefore can be regarded as "partial reproductions" of the source texts. In such cases, their use (further copying and sharing) would either require permission from rightholders, or it would need to enter within the scope of an existing statutory exception. In both situations, the purpose of a DTF would not be fully achieved.

The Court also ruled (also in the Infopaq case) that individual words considered in isolation are not intellectual creations of their authors, and therefore they are not protected by copyright[31]. However, originality of snippets of 11 words (11-grams) extracted from newspaper articles cannot, according to the CJEU, "be ruled out"[32].

One could expect that in Germany, the country of origin of the *kleine Münze* (small coin) doctrine (under which, briefly put, works that only exhibit a low level of originality can still be protected by copyright)[33], very short texts would enjoy copyright protection. Indeed, in 1964 the appellate court of Düsseldorf declared a four-word slogan "Ein Himmelbett als Handgepäck" ('A canopy bed as hand luggage') to be original[34]. This decision is, however, both dated and isolated.

More recently, German courts have systematically refused copyright protection for slogans such as: "Für das aufregendste Ereignis des Jahres" ('For the most *exciting event of the year*')[35], "Thalia verführt zum Lesen" ('Thalia tempts into reading')[36] or "Wenn das Haus nasse Füße hat" ('When the house has wet feet')[37]. These decisions were mostly motivated by lack of originality (i.e. not meeting the required *Schöpfungshöhe*), which cannot manifest itself in a very short work. Originality of a literary work resides in the selection and presentation of its content ("individuelle Auswahl oder Darstellung des Inhalts"); a longer text allows for more free and creative choices[38], whereas very short texts (like slogans or catalogue descriptions[39]) can only exceptionally meet the required threshold[40].

Probably the most representative and relevant for the subject of this deliverable is the 2021 ruling of the German Federal Court of Justice (*Bundesgerichtshof*, BGH) in case I ZR12/08. The administrator of the website perlentaucher.de, containing (among others) summaries of book reviews from various newspapers (which included

31 CJEU, Infopaq, para 45–46.
32 CJEU, Infopaq, para 47–48.
33 First formulated in A. Elster, Gewerblicher rechtsschutz, umfassend Urheber- und Verlagsrecht, Patent- und Musterschutzrecht, Warenzeichenrecht und Wettbewerbsrecht, De Gruyter, 1921.
34 OLG Düsseldorf, 28.2.1964, 2 U 76/63.
35 OLG Frankfurt, 4.8.1986, 6 W 134/8.
36 LG Mannheim, 11.12.2009, 7 O 343/08.
37 OLG Köln, 8.4.2016, 6 U 120/15.
38 OLG Köln, 30.9.2011, I-6 U 82/11.
39 LG Frankenthal, 3.11.2020, 6 O 102/20.
40 OLG Köln, 8.4.2016, 6 U 120/15.

passages from the original reviews), granted two Internet bookstores (amazon.de and bucher.de) a license to use these summaries. In the aftermath, perlentaucher was sued for copyright infringement (alongside trademark infringement and unfair competition) by the publishers of the Frankfurter Allgemeine Zeitung and the Süddeutsche Zeitung, where some of the reviews had originally been published. The Federal Court referred the case back to the Frankfurt Court of Appeals, but in doing so clearly confirmed the Court of Appeal's position according to which individual words and short snippets (*knappe Wortfolgen*) are not protected by copyright.

Furthermore, the Federal Court also criticized the fact that the Court of Appeals did not take into account the proportion of "borrowed" text in the summaries and did not assess the "distance" between the original reviews and the summaries.

This requirement of "sufficient distance", based on the now abandoned doctrine of free use (*freie Benutzung*) of copyright-protected works (former § 24 of the German Copyright Act, repealed in June 2021), relates in an interesting way to the criteria of reconstructability and recognizability, formulated by the Court of Justice of the European Union (see below). Eventually, only some of perlentaucher's summaries that were not distinct enough from the original reviews were found to be infringing.

It seems, therefore, that German courts in general refuse copyright protection of short texts; their originality was only admitted exceptionally, and in older case law. However, the use of multiple short snippets from a copyright protected text, although such snippets individually are not protected by copyright, may still amount to copyright infringement if the resulting compilation of snippets is not sufficiently different ("distant") from the original work.

It seems that cases where a snippet as short as 11 consecutive words is original are relatively rare. However, this 11-words threshold results exclusively from the facts of the *Infopaq* case, and should not be interpreted as a minimum length for a snippet to be found original. Arguably, even shorter snippets can be found original; however, such cases are probably so rare that they can generally be ignored in large-scale text mining projects.

Briefly put, the shorter the snippet, the less likely it is to be found original. The "risk" of originality of very short snippets (shorter than 11 words) appears so low that it's almost negligible. Longer snippets, however, come with non-negligeable risks. The exclusion of very rare snippets from a DTF can significantly reduce this risk.

The inclusion of a snippet that meets the originality threshold in a DTF constitutes a reproduction in part which means that such a DTF can only be shared with permission of the rightholder or within the scope of a statutory copyright exception.

b) DTFs and the "reconstructability" criterion

As mentioned above, 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]"[41].

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

If one considers the sentence "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua" as an example, one can argue that seemingly it can be reconstructed e.g. from the following series of 3-grams: "Lorem ipsum dolor", "dolor sit amet", "amet, consectetur adipiscing", etc., because of the overlapping words. This is probably true in a small corpus, but not necessarily in a larger one, where numerous 3-grams starting with "dolor", which can be a grammatically coherent continuation of "Lorem ipsum dolor", may occur (such as, for example, "dolor in reprehenderit", which actually does occur in the famous "Lorem ipsum" text). On the other hand, if the sentence is divided into non-overlapping 3-grams: "Lorem ipsum dolor", "sit amer, consectetur", "adipiscing elit, sed", etc. it does not necessarily mean that the original text is not reconstructible, even though it certainly becomes more difficult to do so.

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.[42] 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.[43]

Whether it is possible to do so, has to be evaluated separately for each individual DTF. The fact that one DTF (e.g. containing only words no. 1, 3, 5, 7…) can be used to reconstruct the source text if combined with another DTF (e.g. containing only words no. 2, 4, 6, 8 …) developed independently from the same source text, should not influence its copyright status. If, however, several DTFs were built in a coordinated manner (e.g. by one team, or by members of the same consortium), and jointly they can be used to reconstruct the source text, this may have an influence on a judge who could find them infringing.

In any case, 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 sen-

41 CJEU, Infopaq, para 50.

42 K. Du (2024), Rekonstruierbarkeit von abgeleiteten Textformaten, https://events.gwdg.de/event/607/contributions/1408/.

43 K. Kugler, S. Münker, J. Höhmann, A. 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.

sible to restrict it to reconstructions possible with a "reasonable effort."[44] 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.

c) DTFs and the "recognisability" criterion

The third condition that may impact the legal status of short snippets of texts was relatively recently formulated by the CJEU. In 2019 Pelham decision[45] the Court ruled that the use of a very short sample from a phonogram (in the facts of the case: an approximately 2-second rhythm sequence) in another phonogram is an act of reproduction, "unless that sample is included in the phonogram in a modified form unrecognizable to the ear".

Formally, the ruling only addressed the letter (c) of Article 2 of the InfoSoc Directive (i.e., the part relating only to the right of reproduction of phonogram producers), and it is not clear whether this conclusion can (and should) be extrapolated to all rights mentioned in Article 2 (i.e., copyright and certain related rights). Martin Senftleben (2020)[46] argues that "the CJEU can give the copyright concept of "partial reproduction" a meaning that corresponds to the approach taken in Pelham". However, as Senftleben[47] also notices, this does not mean that for literary works the recognizability criterion (the "Pelham test") will replace the criteria put forward in Infopaq (i.e., originality and reconstructability).

Furthermore, it is not clear how the recognizability criterion can relate to literary works. It is possible to interpret the criterion as referring to the identifiability of the work, i.e. the possibility to indicate which literary work the excerpts were taken from. Such an interpretation could lead to most undesirable consequences. It would push the frontiers of copyright protection beyond reason, as it is easy to imagine that some unoriginal elements like statements of (generally unknown) facts, uncommon names or simple yet grammatically flawed expressions could indeed pass the recognisability test.

If this interpretation of the Pelham criterion for partial reproduction were to apply to n-grams used in language technologies, it seems that the use of certain anonymization techniques taken from the realm of data protection could provide some relief.

Arguably, the deletion (or other alteration such as obfuscation) of proper names (called "named entities" in computer science) could make some excerpts less recognizable to an average reader. For example, a recent study demonstrated that OpenAI

44 B. Raue, C. Schöch, Zugang zu großen Textkorpora…; cf. in the context of anonymisation of personal data Recital 26 of the GDPR.

45 CJEU, C-476/17, Pelham, 29 July 2019, para 39.

46 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.

47 Ibid.

models memorised a large collection of copyrighted material by having models perform a task to predict the identity of a single name in a passage of text that contains no other named entities.[48] Named entities can be automatically detected in a corpus and anonymized or otherwise pre-processed.

Moreover, understanding the criterion of recognisability as mere identifiability would lead to a clash with the need to publish metadata about the text. Not only could it be argued that the publication of identifiers about the work or the title could be problematic, but also other basic data for analysis, such as the year of publication or the author. Applying this criterion in this way would result in the inability to publish metadata relating to the DTF, making it unusable for research. In this approach, a library catalogue (which essentially is a list of books' metadata) would infringe copyright, as it would allow to identify the works it refers to.

The concept of recognizability in relation to literary works can and should be understood differently: It is not about the reader identifying the specific work, but rather about recognizing the elements that constitute the creative value of the text, or the individual style of the author. According to this interpretation, the criterion of recognizability is not fulfilled merely by the use of the name of a famous literary character or by the publication of metadata.

2. DTFs as adaptations or transformations of source texts?

The right of adaptation is mentioned in Article 12 of the Berne Convention, according to which "Authors of literary or artistic works shall enjoy the exclusive right of authorizing adaptations, arrangements and other alterations of their works". This right is not harmonised at the EU level, unlike the rights of reproduction and communication to the public.

In Germany, the right of adaptation is safeguarded by § 23 UrhG, according to which "Adaptations (Bearbeitungen) or other transformations (Umgestaltungen) of a work (...) may be published or exploited only with the author's consent". The same section continues: "If the newly created work maintains sufficient distance to the work used, this does not constitute adaptation or transformation (...)".

In the creation of a derived text format, the original text undergoes numerous modifications. The legislator took into account the technically necessary changes involved in creating a corpus for text and data mining (TDM) when introducing the TDM exception. These changes do not constitute adaptations under § 23 of the German Copyright Act (UrhG) but are explicitly excluded (§ 23 (3) UrhG). However, whether the outcome of the TDM process constitutes an adaptation has not been explicitly addressed by the legislator. It can be assumed that the legislator believed that the result of a TDM process is no longer protected by copyright, which is why this issue was not

48 K. Chang, M. Cramer, S. Soni, D. Bamman, (2023, December). Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4 (2023), in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7312–7327).

regulated. However, it appears that in certain cases TDM outputs may still contain copyright-protected material. In such situations, the question arises whether these constitute adaptations. This article, however, aims to define the conditions under which no copyright protection remains, and for this reason, this question will not be delved into.

3. DTFs as original compilations?

Prima facie, DTFs could potentially be regarded as compilations, i.e. collections of various elements (original or not) that can be protected by copyright by virtue of the original "selection or arrangement" of those constitutive parts.

Typical examples of compilations within the meaning of copyright law include dictionaries, encyclopedias, catalogues, collections of poems, "best of" lists, etc. If the exercised free and creative choices (e.g. "My favorite quotes from Woody Allen") in the process, the resulting compilation can be original and therefore protected by copyright.

However, creation of scientifically valuable DTFs does not involve free and creative choices. The compiler does not eliminate certain elements (e.g., certain n-grams) from the DTF on subjective grounds (e.g., just because he or she does not "like" them). On the contrary, DTFs are mostly automatically generated according to predefined criteria and rules of the art. Any subsequent modifications (e.g., deletion of rare expressions) is also based on objective, rather than subjective grounds.

Therefore, the hypothesis according to which DTFs can be original compilations should be excluded.

V. 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 (§§ 44b, 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. Especially Statistical DTFs

and Transformative DTFs, can contain partial reproductions of source texts; this risk can be mitigated by avoiding reproducibility and recognisability of source texts e.g. through avoidance of longer n-grams, especially n-grams longer than 10 words and – depending on the adopted perspective – by avoiding rare n-grams.

However, there are many types of DTFs which clearly fulfil the above-mentioned criteria and can therefore be used freely and risk-free.

VI. Bibliography

https://zfdg.de/2020_006

https://www.nomos-elibrary.de/10.5771/2699-1284-2020-2-118/zugang-zu-grossen-textkorpora-des-20-und-21-jahrhunderts-mit-hilfe-abgeleiteter-textformate-versoehnung-von-urheberrecht-und-textbasierter-forschung-jahrgang-1-2020-heft-2?page=1

https://www.nomos-elibrary.de/10.5771/2699-1284-2020-2-143.pdf

https://www.nomos-elibrary.de/10.5771/2699-1284-2020-2-128.pdf

Keli Du & Christof Schöch: "Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning" (long presentation). In: Karajgikar, J., Janco, A., & Otis, J. (2024). DH2024 Book of Abstracts. Zenodo. https://doi.org/10.5281/zenodo.13761079.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7312–7327, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.453.

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.

Y. Lin, J.-B. Michel, E. Aiden Lieberman, J. Orwant, W. Brockman, S. Petrov, Syntactic Annotations for the Google Books NGram Corpus, in: Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 169–174. URL: https://aclanthology.org/P12-3029.

S. Bhattacharyya, P. Organisciak, J. S. Downie, A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features, Interdisciplinary Science Reviews 40 (2015) 61–77. URL: http://www.tandfonli

ne.com/doi/full/10.1179/0308018814Z.000000000105. doi:10.1179/0308018814Z.00000
0000105.

J. Jett, B. Capitanu, D. Kudeki, T. Cole, Y. Hu, P. Organisciak, T. Underwood, E. Dickson Koehl, R. Dubnicek, J. S. Downie, The HathiTrust Research Center Extracted Features Dataset (2.0), 2020. URL: https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId=79069329. doi:10.13012/R2TE-C227.

C. Schöch, F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. Hinzmann, J. Röpke, Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen (2020). URL: http://zfdg.de/2020_006. doi:10.17175/2020_006.

P. Organisciak, J. S. Downie, Research access to in-copyright texts in the humanities, in: Information and Knowledge Organisation in Digital Humanities, Routledge, 2021, pp. 157–177.

P. Organisciak, B. Capitanu, T. Underwood, J. S. Downie, Access to Billions of Pages for Large-Scale Text Analysis, iConference 2017 Proceedings Vol. 2 (2017). URL: https://hdl.handle.net/2142/98873.

N. N. Parulian, R. Dubnicek, G. Worthey, D. J. Evans, J. A. Walsh, J. S. Downie, Uncovering Black Fantastic: Piloting A Word Feature Analysis and Machine Learning Approach for Genre Classification, Proceedings of the Association for Information Science and Technology 59 (2022) 242–250. URL: https://onlinelibrary.wiley.com/doi/10.1002/pra2.620. doi:10.1002/pra2.620.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team,... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176–182.

Richey, S., & Taylor, J. B. (2020). Google Books Ngrams and political science: Two validity tests for a novel data source. PS: Political Science & Politics, 53(1), 72–77.

El-Ebshihy, A., El-Makky, N. M., & Nagi, K. (2018). Using Google Books Ngram in Detecting Linguistic Shifts over Time. In KDIR (pp. 330–337).

Hessel, J., & Schofield, A. (2021, August). How effective is BERT without word ordering? implications for language understanding and data privacy. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 204–211).

G. Iacino, P. Kamocki, P. Leinen, Assessment of the Impact of the DSM-Directive on Text+

F. Barth, J. Calvo Tello, K. Du, P. Genêt, L. Keller, J. Knappen, Liste der Abgeleiteten Textformate (working title), forthcoming, 2025.

J. Baudry (2023). Non-consumptive research use, an analysis of the legal situation, on Couperin.org, https://www.couperin.org/le-consortium/actus/non-consumptive-research-use/.

WTO, Report of the Panel, WT/DS160/R, 15 June 2000, para 6.229.

Martin Senftleben, The International Three-Step Test: A Model Provision for EC Fair Use Legislation, 1 (2010) Journal of Intellectual Property, Information Technology and E-Commerce Law, Vol. 1, No. 2, pp. 67–82.

Gervais, Daniel J., Towards a New Core International Copyright Norm: The Reverse Three-Step Test. Available at SSRN: https://ssrn.com/abstract=499924.

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

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

First formulated in A. Elster, Gewerblicher rechtsschutz, umfassend Urheber- und Verlagsrecht, Patent- und Musterschutztrecht, Warenzeichenrecht und Wettbewerbsrecht, De Gruyter, 1921.

N. Carlini et al., "Extracting Training Data from Large Language Models" in: Proceedings of the 30th USENIX Security Symposium (August 11–13, 2021), USENIX Association, 2021, pp. 2633–2650.

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.

Calvo Tello, José, Mathias Göbel, Ubbo Veentjer, Stefan E. Funk, Nanette Rißler-Pipika, and Keli Du. 2025 (accepted). 'FAIR Derived Data in TEI and Its Publication in the TextGrid Repository'. *jTEI*.

Burnard, Lou. 2004. 'Metadata for Corpus Work'. In *Developing Linguistic Corpora: A Guide to Good Practice*, edited by Martin Wynne. Oxford: AHDS Literature, Languages and Linguistics. https://ota.ox.ac.uk/documents/creating/dlc/chapter3.htm.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. Scientific Data 3 (March). https://doi.org/10.1038/sdata.2016.18.

Zusammenfassung: Text- und Data Mining (TDM)-Methoden werden angewandt, um große Textmengen für die wissenschaftliche Forschung zu analysieren. Handelt es sich bei dem analysierten Text um urheberrechtlich geschütztes Material, hat die Nutzung solcher TDM-Methoden urheberrechtliche Implikationen. Die bestehenden Urheberrechtsschranken ermöglichen TDM innerhalb eines engen Rahmens, der die Speicherung, öffentliche Zugänglichmachung und Wiederverwendung von Datensätzen einschränkt. Dieser Beitrag untersucht den rechtlichen Rahmen für die Umwandlung des Ursprungsmaterials in ein abgeleitetes Textformat (ATF), das keinem Urheberrechtsschutz mehr unterliegt, um die Nutzung von TDM ohne Einschränkungen zu ermöglichen. Zunächst wird die Erstellung des ATF selbst untersucht: Auch sie umfasst urheberrechtlich relevante Handlungen, die durch die TDM-Ausnahmen abgedeckt sind. In einem zweiten Schritt muss der urheberrechtliche Status des erstellten ATFs anhand von drei Kriterien bewertet werden: Das ATF darf keine Elemente enthalten, die die Schöpfungshöhe des Ursprungsmaterials begründet haben, das Ursprungsmaterial darf nicht anhand des ATF rekonstruierbar sein und das Ursprungsmaterial darf nicht wiedererkennbar sein.

Summary: 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 exceptions. 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.