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.