Originally published in: Lüngen, Harald/Kupietz, Marc/Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Pisetta, Ines (Eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event). – Mannheim: Leibniz-Institut für Deutsche Sprache, 2021. Pp. 10-19.
DOI: *https://doi.org/10.14618/ids-pub-10467*

Corpus Reusability and Copyright – Challenges and Opportunities

Markus Gärtner[1] Felicitas Kleinkopf[2] Melanie Andresen[1] Sibylle Hermann[3]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Center for Applied Legal Studies (ZAR), Karlsruhe Institute of Technology
[3]University Library, University of Stuttgart
markus.gaertner@ims.uni-stuttgart.de, felicitas.kleinkopf@kit.edu,
melanie.andresen@ims.uni-stuttgart.de, sibylle.hermann@ub.uni-stuttgart.de

Abstract

Making research data publicly available for evaluation or reuse is a fundamental part of good scientific practice. However, regulations such as copyright law can prevent this practice and thereby hamper scientific progress. In Germany, text-based research disciplines have for a long time been mostly unable to publish corpora made from material outside of the public domain, effectively excluding contemporary works. While there are approaches to obfuscate text material in a way that it is no longer covered by the original copyright, many use cases still require the raw textual context for evaluation or follow-up research. Recent changes in copyright now permit text and data mining on copyrighted works. However, questions regarding reusability and sharing of such corpora at a later time are still not answered to a satisfying degree. We propose a workflow that allows interested third parties to access customized excerpts of protected corpora in accordance with current German copyright law and the soon to be implemented guidelines of the Digital Single Market directive. Our prototype is a very lightweight web interface that builds on commonly used repository software and web standards.

1 Introduction

In several fields of text-based research with corpora such as corpus linguistics, digital humanities, and computational literary studies, researchers have for a long time been faced with the precarious situation that text corpora cannot be published for reuse due to legal issues. While the Fair Use doctrine of United States law and similar legal systems expresses a rather usage-friendly idea for copyrighted material, other legislatures take approaches that focus primarily on the rightholders instead. One of these is the German copyright law[1] (as substantially determined by EU law) as the legal context in which this work is situated.

Using copyrighted material as the basis of public text corpora was generally not possible under German copyright until recently. Naturally the option of making special arrangements with individual rightholders always existed and has been used by larger projects and institutions, often focusing on data from the news domain. But given the time investment needed to reach such agreements and the relatively short lifespan of most (smaller) research projects, those cases remain exceptions. As a direct result, large portions of the corpora created from German texts outside the news domain are based on material that has already been in the public domain.[2]

With recent changes in German and European copyright, protected material is now available for non-commercial research (see Section 2). However, the question about archiving and public availability of research data and corpora created from copyrighted material after the official end of associated projects is still not solved to a satisfying degree.

To address this issue, we present an architecture concept and its prototypical implementation that allows researchers to make excerpts of otherwise non-publishable copyrighted text corpora available for (scientific) reuse. For this approach, the intelligent choice of excerpts tailored to the user's needs is key, because having only a randomly or statically selected part of a corpus available is of limited benefit for some research questions. Therefore, the system additionally integrates a dedicated query component. In order to maximize the utility, users can express their interest based on available annotations in the corpus and as such receive excerpts of higher relevance for them.

[1]For the remainder of this text "copyright" or "local copyright" refers to German copyright law, unless explicitly stated otherwise.

[2]Usually due to the original author being dead for a sufficiently long period of time.

DOI: *https://doi.org/10.14618/ids-pub-10470*

The approach is tailored to the current legal situation in Germany, but can easily be transferred to other legal frameworks that contain regulations of similar setup. With the upcoming implementation of the DSM-Directive (see Section 2) into national laws, the copyright situation for text and data mining (TDM) within the EU becomes more homogeneous. As such the concept in this paper can serve as a blueprint for corpus reusability in this shared legal sphere.

We discuss the relevant legal framework in Section 2 and contextualize our work in Section 3. Sections 4 and 5 describe the XSample approach and the current prototype implementation and finally Section 6 concludes.

2 Legal Framework

In order to discuss why copyright problems arise in relation to the reuse of TDM corpora, the legal framework of TDM is first presented below in Section 2.1. Of enormous importance in this respect is recent European law, the Directive on Copyright in the Digital Single Market (Section 2.2), which, although not directly applicable to national law, had to be implemented by the member states by June 7th, 2021. Finally, we discuss why the reusability of corresponding corpora remains legally unclear and what approach should be considered to address this problem under European (Section 2.3) and German law (Section 2.4) .

2.1 Text and Data Mining and Copyright Law

Copyright law must only be observed when practicing text and data mining if text and data are protected under copyright or related rights. The preconditions on a protection depend partially on national law and partially on European law. According to the case law of the European Court of Justice (ECJ), a work protected by copyright exists if it is the author's own intellectual creation.[3] However, the necessary level of creation is low: It can already be reached by a part of a work that consists of eleven words.[4] In research areas that deal with text-based resources, a protection by copyright must be assumed in most cases. Moreover, databases are protected under a so-called sui generis right in case of being a qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents, Article 7 (1) of the directive 96/9/EC on the legal protection of databases (Database-Directive).

In fact, the analyses performed in text and data mining processes as such do not violate intellectual property. However, in terms of preparing research data it is necessary to copy and to make works and related rights available to the public within research groups, see also Raue (2018, p. 381) and Geiger et al. (2018, p. 817 f.).[5]

By Articles 2 and 3 of the Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society (InfoSoc-Directive), these acts of exploitation are exclusively entitled to the holders of the copyrights and related rights as the reproduction right and the right of communication to the public. Therefore, these acts require the rightholder's permission or an exception or limitation provided for by law.[6]

2.2 A Developing Legal Framework within the European Union

Although research on copyrighted works is still a rarity in the digital humanities, it has already been allowed in several member states of the European Union for a few years: The first member state to implement a national regulation that allows text and data mining research has been the United Kingdom in 2014[7], followed by France in 2016[8], Estonia in 2017[9] and Germany in 2018[10] (Geiger et al., 2018, p. 830 f.). In introducing those limitations or exceptions to copyright and related rights, these states referred to Article 5 (3a) InfoSoc-Directive, an authorization to grant additional rights for non-commercial scientific research.

By Articles 3 and 4 of the new Directive 2019/790 on copyright and related rights in the Dig-

[3] ECJ – Infopaq, ECLI:EU:C:2009:465, no. 30 ff.

[4] ECJ – Infopaq, ECLI:EU:C:2009:465, no. 38.

[5] DSM-Directive, recital 9. However, after the case law of the ECJ, the requirement of being publicly available "refers to an indeterminate number of potential listeners, and, in addition, implies a fairly large number of persons", e.g. ECJ, ECLI:EU:C:2012:140, Società Consortile Fonografici (SCF)/Marco Del Corso, no. 84. Therefore, only large research groups can be considered public.

[6] Contrary to the literal meaning, an exception or a limitation does not limit usage but enables it by permitting particular usages, for instance reproducing a copyrighted work or making it publicly available.

[7] Regulation 3 of the British Copyright and Rights in Performances Regulations 2014, No 1372, Article 29a

[8] Article 38 of the French law no. 2016-1321, 7th October 2016

[9] Estonian Copyright Act 19 (3)

[10] BT-Drs. 18/12329 (printed matters of the German parliament), Regulation 17, § 60d

ital Single Market (DSM-Directive), all EU member states are now obliged to provide for mandatory exceptions or limitations for text and data mining for the benefit of non-commercial scientific research and also for other purposes. Both permissions are subject to the condition that practitioners of TDM already have lawful access to the protected works. Rightholders can prevent TDM in non-commercial contexts only to the extent absolutely necessary by invoking the operability and safety of their systems, Article 3 (3) DSM-Directive, whereas it is possible to express a reservation in a machine-readable manner within commercial contexts, Article 4 (3) DSM-Directive. According to the European legislator, the interests of rightholders are affected by TDM research only to a minor extent: In any case, the member states are explicitly not to provide for compensation of rightholders for the acts of exploitation carried out to prepare corpora.[11]

Due to this obligation it must be legally possible in all EU member states to carry out research in terms of text and data mining on copyrighted works or databases in any case from June 2021 onward. By missing this deadline, member states risk infringement proceedings, Articles 258 ff. Treaty on the Functioning of the European Union (TFEU).

2.3 Remaining Legal Uncertainties: Reusing Copyrighted Corpora

It was and still is not only the legal uncertainty in terms of research on copyrighted works that was and is slowing down scientific progress, but also the uncertainty regarding the fate of research data after completion of the respective research, e. g. the scientific review at a later stage, the storage of research data and the reusability of corpora (Kleinkopf et al., 2021): According to Article 4 (2) DSM-Directive, corpora may be retained in commercial contexts only for as long as is necessary for the purposes of the TDM. In contrast, Article 3 (2) DSM-Directive does not provide for a time limit for retention in non-commercial contexts, but limits retention to the purposes of non-commercial scientific research (and also requires appropriate safeguarding). In this respect, it is up to the member states to implement this into national law in the most research-friendly way possible.

Regarding the question of a legal option to reuse the corpora, recital 15 of the DSM-Directive should find attention. This recital refers to Article 5 (3a) InfoSoc-Directive that authorizes member states to allow acts of reproduction of protected works and communicating them to the public for the purposes of non-commercial scientific research. Therefore, the national exceptions or limitations for those purposes could be applied (Kleinkopf et al., 2021). According to Article 25 of the DSM-Directive, member states are also explicitly allowed to go beyond the requirements of the DSM-Directive and grant extended authorizations on the basis of the InfoSoc-Directive. This includes that other national exceptions or limitations in favor of non-commercial scientific research are still applicable.

The combination of different copyright limitations is not new in European law: In "Eugen Ulmer/TU Darmstadt", the ECJ decided that it is possible to combine different exceptions and limitations under the InfoSoc-Directive, provided that the requirements of each are met.[12] The specific case concerned the combination of copyright exceptions and limitations under Article 5 (2b) and Article 5 (3n) of the InfoSoc-Directive. Because the exceptions and limitations of the InfoSoc-Directive continue to apply under the DSM-Directive, the idea behind this approach is to transfer this case law to Article 3 of the DSM-Directive and Article 5 (3a) of the InfoSoc-Directive (Kleinkopf et al., 2021). In addition, recital 15 of the DSM-Directive assumes the cumulation of exceptions and limitations under copyright law.

One limit is the three-step test under Article 5 (5) of the InfoSoc-Directive, which must be observed both in the context of legislation and in the context of judicial interpretation of the law (Stieper, 2009, p. 73 with further evidence).[13] The three-step test states that copyright limitations must be limited to certain special cases (step one), they must not interfere with the normal exploitation of the work (step two) and they must not unreasonably prejudice the legitimate interests of the author (step three). The scientific reuse of TDM corpora must be regarded as such a special case, furthermore, the primary market is not affected if only parts of the corpora are reused (Kleinkopf et al., 2021). The requirements for unreasonableness for rightholders as stated in the third stage tend to be regarded as high (Senftleben, 2004, p. 210 f.) and, in view of

[11] DSM-Directive, recital 17

[12] ECJ - Eugen Ulmer, ECLI:EU:C:2014:2196, No. 50 ff.

[13] German federal High Court of Justice (BGH), judgment of 11th July 2002 - I ZR 255/00, GRUR 2002, 963 – Elektronischer Pressespiegel

the worthiness of protection of scientific interests under Article 13 Charter of Fundamental Rights of the European Union, are not met at least in the case of an obligation to pay remuneration (Kleinkopf et al., 2021). This remuneration is often granted in general permissions of non-commercial scientific research, see e. g. § 60h of the German Copyright Law (Urheberrechtsgesetz).

2.4 Reusability of Copyrighted Corpora under German Law

The German legislator recently updated the German permission to use copyrighted works in favor of non-commercial, scientific text and data mining research, § 60d Urheberrechtsgesetz.[14] While the permission of acts of reproduction of protected works has been extended to commercial purposes, § 44b Urheberrechtsgesetz, the permission for scientific research has not been significantly extended: Although it is permitted to make the corpora accessible for peer review procedures and to retain them for research purposes, it is legally unclear whether the corpora may also be archived by third parties (Kleinkopf and Pflüger, to appear). Moreover, the German legislator missed the chance to add the possibility to make corpora explicitly reusable. The more general permission of exploiting copyrighted works under German Copyright Law is § 60c Urheberrechtsgesetz that implements Article 5 (3a) InfoSoc-Directive in national law. By applying § 60c on § 60d Urheberrechtsgesetz, it is possibly to reuse corpora at least partly. In detail, § 60c allows the usage of extracts that do not exceed 15% of the total work. It also allows to use individual short publications such as journal articles that are no more than 25 pages long completely.[15]

3 Related Work

The official opening of copyrighted works for research has only been implemented quite recently and copyright regulations have also changed several times. As such, much of previous work has been designed under different legal conditions. However, development of infrastructure or support software was generally focused on circumventing copyright in legal ways.

Research projects working with German corpora have established different ways of dealing with the legal situation. The Institute for German Language (IDS) hosts the German Reference Corpus DeReKo that comprises more than 50 billion words.[16] Access to the data is regulated by more than 200 individual license agreements with rightholders (Kupietz et al., 2018). However, access to DeReKo is still restricted: It is available "for non-commercial, scientific research by registered users and strictly within the query-and-analysis-only framework" (Kupietz et al., 2018, p. 4353). Consequently, researchers can never access full texts but only get results for specific queries with limited context. The situation is similar for the second large reference corpus for German, the DWDS corpora[17] (Geyken et al.). While the effort of individual license agreements is possible for large and long-term funded institutions as in these examples, this is not feasible for most individual projects with few employees on short-term contracts. Many projects therefore fall back to historic data that are already in the public domain or do not publish their corpora at all.

One recent suggestion that is implementable for small projects and individual researchers is the concept of derived text formats ("abgeleitete Textformate") by Schöch et al. (2020b), see also Schöch et al. (2020a). The authors propose methods to obfuscate the original text in a way that the result is no longer covered by the original copyright and may be published and shared freely. This includes, for instance, the publication of word (or n-gram) frequency lists or a text version with scrambled word order. These derived text formats allow for some types of analysis that are popular in the digital humanities, like stylometry or topic modeling.

Whether derived text formats or results for a specific query only are useful or access to more context is required depends, of course, on the research question. In the XSample project, we explore the different needs via two use cases from the humanities. The first use case is the project CAUTION[18] in literary studies that explores the phenomenon of unreliable narrators that are, for instance, lying or do only have limited knowledge about the narrated world. It can easily be seen that this phenomenon cannot be captured in word frequency lists, but requires a lot of textual context. The second use

[14] Bundesgesetzblatt (German federal law gazette), Bgbl. 2021 Teil I Nr. 27 p. 1204 ff.

[15] BT-Drs. 18/12329, p. 35

[16] https://www.ids-mannheim.de/digspra/kl/projekte/korpora/

[17] https://www.dwds.de

[18] https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-Caution/

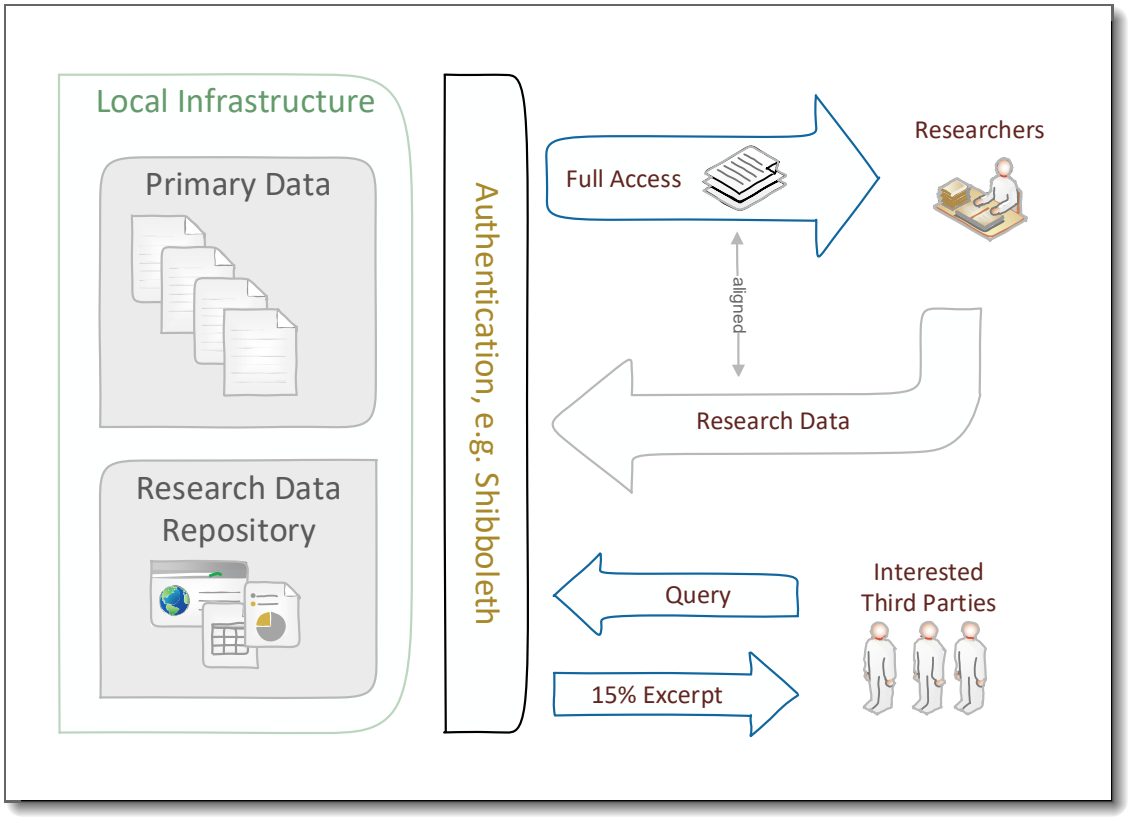

Figure 1: Architecture overview for XSample with the main parties involved and the general data flow.

case follows a linguistic research question about the academic language of linguistics and literary studies, replicating Andresen (to appear). The core of the analysis is based on a comparison of frequencies that can be performed on derived text formats. However, the interpretation of the quantitative results requires that findings can be recontextualized in the original texts, making full text access highly desirable, if not mandatory.

In sum, most qualitative approaches require as much context as possible right away, some quantitative approaches can be performed with little context, but their interpretation and evaluation has to rely on context as well. We therefore think that the given possibilities for making corpus data available can be complemented by a more flexible, individual approach. Our workflow is based on the right to distribute excerpts of texts and will be described in the following section.

4 The XSample Workflow

Our approach is based on combining § 60c and § 60d Urheberrechtsgesetz. Those regulations allow researchers the use of copyrighted material and libraries the passing on of excerpts of protected material up to a certain limit, respectively. With

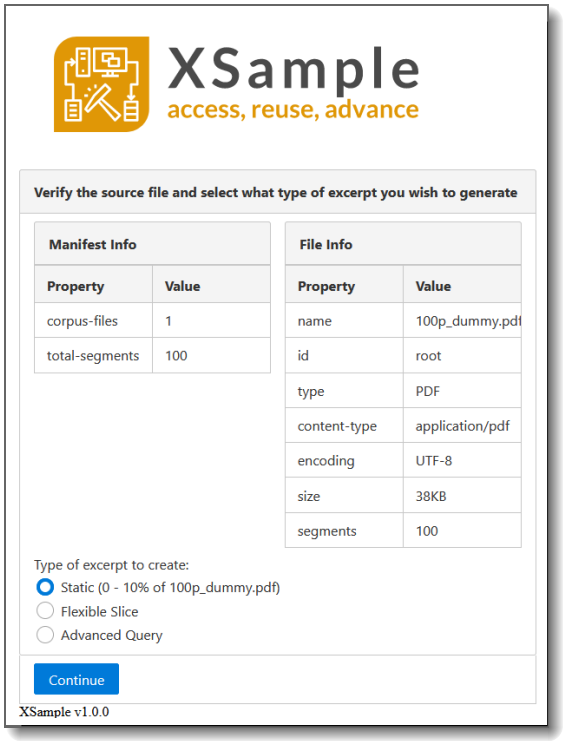

Figure 2: Screenshot of the XSample landing page.

libraries or other archiving institutions as central actors, the architecture depicted in Fig. 1 supports the entire workflow from copyrighted primary data to individualized excerpts for end users.

In an initial ingest step (upper loop in Fig. 1) researchers in an active project process copyrighted texts and deposit research data (intermediate annotations or finalized corpus resources) into the repository. The "aligned" label in Fig. 1 signifies that there has to be a way to map annotations or linguistic units back to the segments (pages) of the primary data they appear on or refer to. This reverse mapping is a crucial prerequisite for the query-driven excerpt generation described in Section 5.2.

Unsurprisingly, both use cases in our project highlighted the fact that typical corpus generation processes need to be adjusted in order to keep or restore this kind of mapping information. In both cases, text was originally extracted from PDF and EPUB documents either directly or via OCR, cleaned and then transformed into formats specific to the use case, losing the page mapping in the process. Subsequent modifications of the processes led to manual restoration of mappings as annotations in one case and automatic preservation as external (tabular) mapping files in the other.

Both the primary data and generated annotations[19] are stored in the shared or private domain, making them not directly available to the public. Additionally, special metadata following the XSample schema in JSON-LD[20] format is added to the repository in the public domain. This metadata makes the protected data findable and serves as entry point for end users that wish to receive excerpts of the corpus for inspection or evaluation (lower loop in Fig. 1). Actual end users of the XSample concept can be any kind of interested third parties, but are primarily expected to be other researchers that wish to evaluate the data for either reproducibility or suitability in the context of their own projects.

During the excerpt generation process users are subsequently redirected to the XSample web interface, visible in Fig. 2 in a horizontally compacted layout. There they are able to further specify exactly how the excerpt should be generated. Available options at this time include the following:

1. A **static** excerpt generation configurable by the original corpus creators within the XSample metadata file.
2. A user-defined continuous section or **slice**. Selection of this slice is achieved via a simple GUI with the same double-knob slider also used in the query approach in Fig. 4.
3. Filtering of (linguistic) annotations in the corpus based on user interests expressed in a formal **query**. This approach is explained in greater detail in Section 5.2 and also showcased in Fig. 4.

After successful completion of the XSample workflow, users are presented with a zip archive for download. Contained within this archive are the actual pages of the primary data that represent the excerpt itself alongside with annotations for those parts. At present the prototype implementation supports annotations in the tabular format of the CoNLL 2009 Shared Task (Hajič et al., 2009) and an extension for a TEI[21] subset is being worked on.

To conform with current law, the system must not give access to more than 15% of any particular resource[22] to individual users and therefore needs to track quotas. In order to minimize integration footprint and authentication overhead, the prototype implementation does not manage users itself, but relies on information provided by the Dataverse repository for identification and tracking of individual users.

5 Architecture

The XSample prototype is implemented completely web-based[23] and only consists of a few components in order to keep it lightweight and minimize the need for adjustments when integrating it into existing infrastructure. It is still under active development and while the current version can already serve a large portion of the basic XSample workflow, it is not yet feature complete. The source code is publicly available on GitHub[24] under an open source license. Sections 5.1 to 5.3 describe the integration into an existing Dataverse repository, excerpt generation and the handling of composite corpora in more detail.

[19] Strictly speaking this only concerns annotations which are still covered by copyright, i. e. those dissimilar to the derivational approach described by Schöch et al. (2020b).

[20] https://json-ld.org/

[21] Text Encoding Initiative https://tei-c.org/

[22] See Section 5.3 for details on how special cases are handled with respect to excerpt size limits.

[23] Using the Jakarta Server Faces (JSF) framework.

[24] https://github.com/ICARUS-tooling/xsample-server

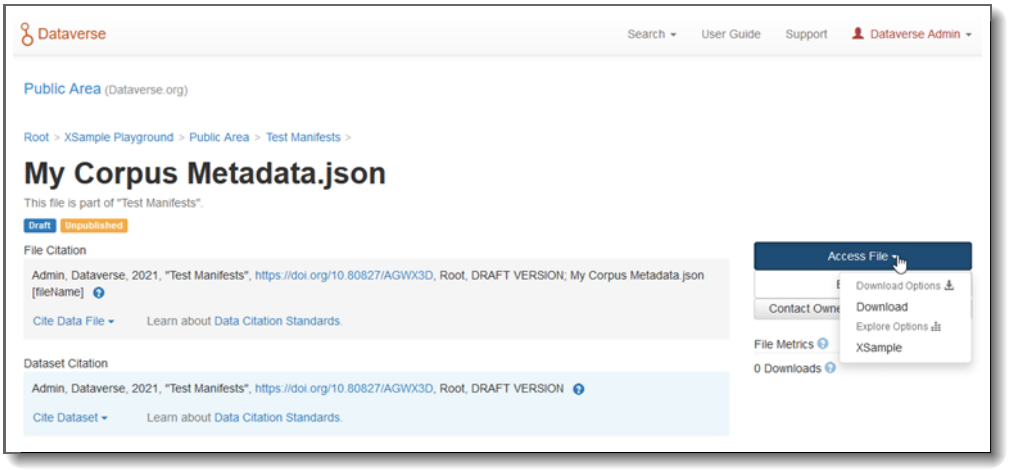

Figure 3: Screenshot showcasing the integration of XSample into the Dataverse user interface nd a metadata file as entry point for the XSample workflow. XSample is added as an *External Tool* for files of a specific content type and therefore appears as additional access option, visible in the dropdown menu to the right in the screenshot.

5.1 Dataverse Integration

The XSample prototype implementation is geared towards interfacing with a Dataverse[25] repository instance. Dataverse is an open source repository software built on JSF that is widely used for research data management and offers the granularity in access control required for the XSample workflow outlined in Section 4. Since Dataverse is also able to interface with existing authentication providers of the university or institute the system is deployed on, we can already rely on identification of unique users for excerpt quota tracking.

For integration, Dataverse's *External Tools API*[26] is used. It allows to register external web services for datasets[27] or files of specific content types in a way that does not require code modifications for the repository. External tools registered that way are then added as menu items when interacting with the Dataverse web interface. When used, they can send the user to a predefined server or service and also transmit various additional parameters, depending on their configuration. Possible parameters (all of which are used for XSample) are, among others, the resource ID, the public URL of the Dataverse repository or the user's API token.

Figure 3 shows an example snippet of the Dataverse interface for a metadata file[28] that serves as entry point to the XSample workflow. The "Access File" menu to the right contains the link to the external XSample server, usable to initiate the excerpt generation process.

5.2 Query-Driven Excerpt Generation

Depending on the use case, composing the excerpt of static (e. g. the first 15% of a corpus) or random elements might be of little benefit as there is no guarantee that passages or phenomena relevant to a user's particular interests are covered. In order to optimize excerpt generation, XSample includes a corpus query interface in the excerpt step (lower loop in Fig. 1) of the workflow.

In this interface, users can express their interest in a formal query language which the query backend evaluates on the annotation contents of the corpus to produce excerpt candidates. Candidates are determined by mapping the raw hits of a query result, for instance sentences when searching for a specific syntactic phenomenon, to actual segments in the primary data used for excerpt generation. In the case of primary data being in PDF format, the segments and candidates will be individual pages.

The distribution of candidates and their underlying raw hits over the entire corpus is subsequently visualized (cf. Fig. 4) to give users a preview of the expected size of their excerpt and to allow them to further refine the query. This visualization does, however, not contain the raw text or annotations

[25] The Dataverse Project, https://dataverse.org/

[26] https://guides.dataverse.org/en/latest/api/external-tools.html

[27] Within a Dataverse repository "datasets" and are used to organize file resources into logical groups.

[28] Due to an inconsistency in Dataverse 5.3, the version currently used for the XSample prototype, API tokens of users are not transmitted to external tools for public files. This issue has been raised in the Dataverse developer community and is being worked on. As a temporary workaround XSample metadata files in the test setup are therefore required to be private/drafts (cf. Fig. 3) until the inconsistency is fixed.

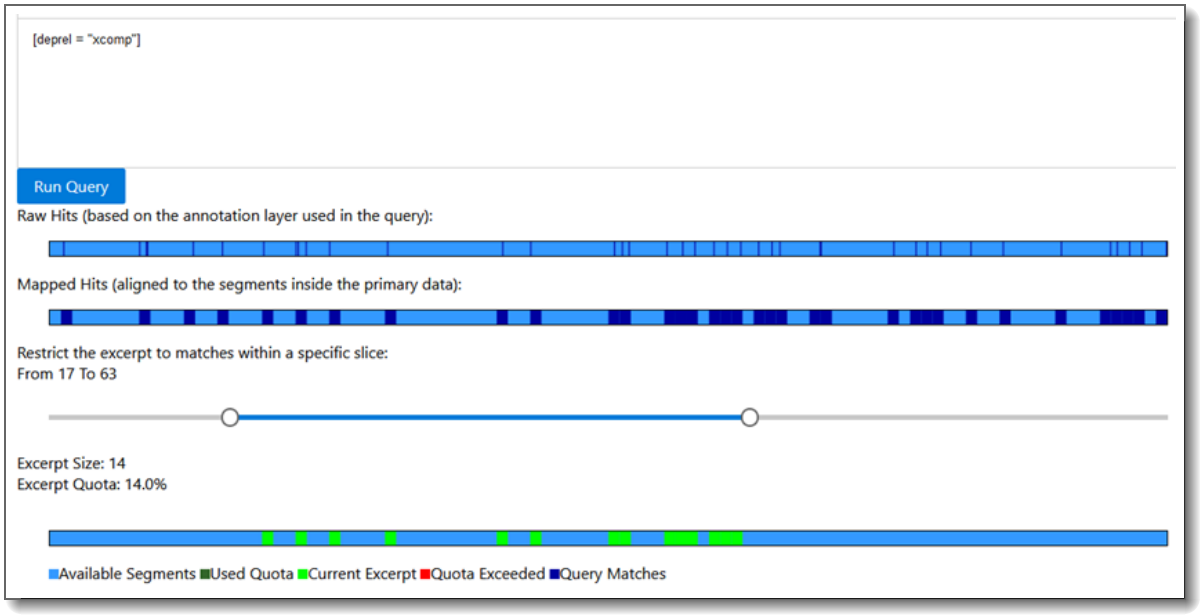

Figure 4: Screenshot snippet of the graphical interface for query-based excerpt generation. From top to bottom the screenshot shows components for query definition, raw hit visualization, mapped hit visualization and a control component with final excerpt visualization similar to the slice excerpt generation mentioned in Section 4.

of the corpus, as access to those is only granted when downloading a finished excerpt. Once satisfied with the result of their query, users can restrict the excerpt to be composed of candidates in a selectable region (similar to the "slice" option in Section 4) or request a random subset of candidates to be used. This way, the potential relevance of excerpts for individual users can be maximized.

The dependence on available annotations and specific formats in the corpus poses a major challenge for the implementation of the query-driven excerpt generation: Both the query backend and the component responsible for splitting annotation files for the excerpt must be able to handle a given set of corpus files to make this approach viable: The former to evaluate the query in the first place and the latter to split the annotation files when they are requested to be part of the excerpt.

For rapid prototyping we initially chose the query component of ICARUS (Gärtner et al., 2013) as evaluation backend, as it readily supports the CoNLL 2009 format and provides a simple bracket-style query language. In parallel, an alternative based on a more general middleware solution (Gärtner and Kuhn, 2018) is being worked on. Since the interface between the query backend and both the user interface and excerpt generation component is rather slim, plugging in a new implementation to support additional formats or query languages can be done fairly easy.

5.3 Composite Corpora

For simple corpora that consist of only a single copyrighted work, applying current regulations and size limits to the excerpt generation process is pretty straightforward. They apply directly to the entire corpus and in special cases such as certain journal articles or *small-scale* works the corpus is completely exempt from the 15% limit (cf. Section 2.4).

The situation becomes much more complicated when dealing with composite corpora, that is, corpora composed of a collection of individually copyrighted works: In such cases all rules and exceptions refer to contained works rather than the corpus as a whole.[29] As a direct result, the XSample server cannot deliver a blanket 15% excerpt for a composite corpus, but takes measures to ensure that the 15% limit is adhered to for each individual work. The server is informed of the actual corpus composition by the metadata (see Section 4) that serves as entry point for the XSample workflow. While the metadata schema allows for arbitrarily complex corpus compositions, the current server implementation is more limited: On the backend a nesting depth of one[30] is supported and the user

[29] While corpora could be viewed as databases themselves, researchers interested in making them available for reuse are typically consenting to copyright uses.

[30] A corpus may consist of multiple copyrighted works without further subdivision.

interface is only able to handle single-work corpora at this time but is currently being adjusted to match the backend capabilities. Especially the inclusion of works that are exempt from the 15% limit in composite corpora poses a serious challenge when developing the user interface while also aiming for a high degree of usability and intuitiveness.

6 Conclusion

In this paper we analyzed the evolving legal situation in Germany regarding copyright in the context of the European Digital Single Market, highlighting the shortcomings for research in text-based disciplines. We then proposed the XSample workflow as a concept for providing excerpts of copyrighted (text) material in order to support reproducibility and reusability. Our prototype implementation is web-based and initially designed to interface with Dataverse repositories only. It also features a query component to guide the excerpt generation process to more relevant samples based on a user's interests. However, having a very small integration footprint for both the repository and query components, it can also be adjusted to work with other systems. In the future we intend to widen the support for different corpus or annotation formats and also explore the possibility to apply the concept to material beyond text, such as audio or video resources.

Acknowledgments

This work was funded by the Ministry for Science, Research and the Arts in Baden-Württemberg (MWK) via project XSample through the funding program "BW-BigDIWA – Wissenschaftliche Bibliotheken gestalten den digitalen Wandel".

References

Melanie Andresen. to appear. *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen*. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP). Narr Francke Attempto.

Markus Gärtner and Jonas Kuhn. 2018. A Lightweight Modeling Middleware for Corpus Processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1087–1095, Miyazaki, Japan. European Language Resources Association (ELRA).

Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. 2013. ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Sofia, Bulgaria.

Christophe Geiger, Giancarlo Frosio, and Olekasndr Bulayenki. 2018. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data? Legal Analysis and Policy Recommendations. *Int. Review of Intellectual Property and Competition Law (IIC)*, pages 814–844.

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand, and Lothar Lemnitzer. Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45(2):327–344.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA.

Felicitas Kleinkopf, Janina Jacke, and Markus Gärtner. 2021. Text- und Data-Mining: urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora und ihre Bedeutung für die Digital Humanities. *MMR: Zeitschrift für IT-Recht und Recht der Digitalisierung*, pages 196–200. Open Access version available at http://dx.doi.org/10.18419/opus-11445.

Felicitas Kleinkopf and Thomas Pflüger. to appear. Digitale Bildung, Wissenschaft und Kultur – Welcher urheberrechtliche Reformbedarf verbleibt nach Umsetzung der DSM-RL durch das Gesetz zum Urheberrecht im digitalen Binnenmarkt? *Zeitschrift für Urheber- und Medienrecht (ZUM)*.

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

Benjamin Raue. 2018. Free Flow of Data? The Friction between the Comission's European Data Economy Initiative and the Proposed Directive on Copyright in the Digital Single Market. *Int. Review of Intellectual Property and Competition Law (IIC)*, pages 379–383.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis,

Maria Hinzmann, and Jörg Röpke. 2020a. Abgeleitete Textformate: Prinzip und Beispiele. *RuZ - Recht und Zugang*, 1(2):160–175.

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

Martin Senftleben. 2004. Grundprobleme des urheberrechtlichen Dreistufentests. *GRUR International Journal of European and International IP Law*, pages 200–211.

Malte Stieper. 2009. *Rechtfertigung, Rechtsnatur und Disponibilität der Schranken des Urheberrechts*. Ph.D. thesis, Tübingen.