VERSION 1.0 – FEBRUARY 2017

GUIDELINES FOR BUILDING LANGUAGE CORPORA UNDER GERMAN LAW

Guidelines by the DFG Review Board on Linguistics

Version 1.0, February 2017: This is an English-language translation of guidelines published by the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG) in March 2015, originally available at:

http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf

Translation from the German: Erik Ketzan, Julia Wildgans, John Weitzmann

This translation preserves the original text, but occasionally adds notes via [*2016 note*], for instance for updated citations.

Distributed by the CLARIN Legal Issues Committee (CLIC), http://clarin.eu

TABLE OF CONTENTS

Preliminary remarks

These recommendations were developed at two roundtable meetings of the German Research Foundation (DFG) in 2012 and 2013, which took place on the initiative of the DFG review board on linguistics, coordinated by Arnulf Deppermann and Mechthild Habermann in cooperation with Helga Weyerts-Schweda from the DFG. Within these roundtable meetings, working groups were formed which wrote these recommendations.

The roundtable meeting devoted to spoken corpora hosted by Arnulf Deppermann and Thomas Schmidt (IDS Mannheim) took place at the DFG's office in Bonn on November 9, 2012. The members of the working group on legal aspects of the use and provision of spoken corpora (part 1 of the recommendations) are: Jörg Bücker, Arnulf Deppermann, Sebastian Drude, Dagmar Jung, Paweł Kamocki, Erik Ketzan, Christoph Purschke, Angelika Redder, John H. Weitzmann and Thomas Schmidt (coordinator). Comments from the DFG legal advisor Mrs. Hagena-Schmedding and the DFG data protection officer Mr. Dörel were considered in draft versions.

The related roundtable meeting for written corpora hosted by Alexander Geyken (BBAW Berlin) and Marc Kupietz (IDS Mannheim) took place at the DFG on November 15, 2013. The members of the working group for formulating the information concerning legal aspects of the use and provision of written corpora (part 2 of these recommendations) were: Gerhard Heyer, Christian Mair, Roland Schäfer and Silke Schwandt. In addition, Dagmar Deuber, Richard Eckart de Castilho, Judith Eckle-Kohler and Iryna Gurevych contributed to the recommendations. Further parts of the text were written and edited by Pawel Kamocki, Erik Ketzan and John H. Weitzmann, who together performed a review of the legal scholarship, a process coordinated by Marc Kupietz.

Introduction

The possibilities of re-use and archiving of spoken and written corpora are affected by personality rights (depending on legal tradition also called: the right of publicity), copyright law and data protection / privacy laws. These recommendations include information about legal aspects which should be considered while creating corpora to ensure the greatest archivability and re-usability possible in compliance with current laws.

The information compiled here shall serve researchers who plan to create corpora or who are involved in evaluation of such measures as a guideline. This information is not exhaustive or to be considered as legal advice. Researchers should consult institutional legal departments and management before making legally relevant decisions. That said, further legal expertise should be sought if possible as early as project planning phases.

Part 1: Information on legal aspects of the use of spoken corpora

1.1. Data protection / privacy aspects

The data that data protection / privacy laws apply to, so-called "personal data", is data that refers to living[1] humans. This includes data that is traceable to individuals or very small groups of people.

For spoken corpora projects, the processing (i.e. collection, storage, modification, transmission, blocking, deletion and other uses) requires the consent and cooperation of the people who are to be recorded.[2] In this situation the interest of the recorded person to protect his/her personal data and the interest of the researcher to be able to use this data to the largest extent possible can run contrary to each other. The aim must be to negotiate and find a feasible, legally correct and ethically responsible compromise between the potentially opposing interests. This solution should on the one hand give full effect to the recorded person's right of data protection, on the other hand take account of the interests of the scientific community, e.g. not preclude ways of data use that do not impact onprivacy.

- Data protection regulations for responsible authorities which have their seat or a branch in Germany[3] can be found in EU law (Directive 95/46/EC and Directive 2002/58/EC), national laws (*Bundesdatenschutzgesetz* BDSG) and federal state laws (*Landesdatenschutzgesetze*, e.g. Hamburger DSG). The BDSG applies to public bodies of the Federal Republic of Germany and non-public authorities, e.g. companies. For universities and other public bodies of the federal states, the respective federal state data protection law applies.[4] These recommendations are for researchers at the above-mentioned institutions. Thus it should be emphasized that public bodies of the Federal Republic of Germany and the federal states of Germany must observe different laws of data protection (although the

[1] Data protection laws apply to deceased individuals only to a very limited extent.
[2] This primarily applies to data which was collected by the researcher him/herself. In cases of data from TV, radio, Internet, researchers are often not in the position to ask for the people's consent and their willingness to cooperate.
[3] We confine ourselves to information concerning the nationwide legislation throughout the Federal Republic of Germany. Due to the fact that the legislation in other EU member states follows the same EU directives, the regulations there are quite similar, with some differences. It should be also taken into account that the legislation of some federal states may contain additional regulations for data protection which are not mentioned here. The data protection legislation of non-EU-countries may differ considerably from the cases described below. If the collection of spoken data affects legislation of non-EU-countries (e.g. if data from abroad is recorded), additional legal advice should certainly be obtained.
[4] Landesdatenschutzgesetz Baden-Württemberg (LDSG BW), Bayerisches Datenschutzgesetz (BayDSG), Berliner Datenschutzgesetz (BlnDSG), Brandenburgisches Datenschutzgesetz (BbgDSG), Bremisches Datenschutzgesetz (BremDSG), Hamburgisches Datenschutzgesetz (HmbDSG), Hessisches Datenschutzgesetz (HDSG), Niedersächsisches Datenschutzgesetz (NDSG), Datenschutzgesetz Mecklenburg- Vorpommern (DSG M-V), Datenschutzgesetz Nordrhein-Westfalen (DSG NRW), Datenschutzgesetz Rheinland-Pfalz (DSG RLP), Saarländisches Datenschutzgesetz (SDSG), Sächsisches Datenschutzgesetz (SächsDSG), Datenschutzgesetz Sachsen-Anhalt (DSG-LSA), Landesdatenschutzgesetz Schleswig- Holstein (LDSG SH), Thüringer Datenschutzgesetz (ThürDSG).

content of the laws is largely the same, key differences exist).[5] [*2016 note: From May 25th 2018 onwards the EU General Data Protection Regulation will provide the main regulatory framework*.]

The aim of data protection laws including the General Regulation is to protect individuals against violations of their right of informational self-determination during the processing of personal data through authorities and other bodies. The *Bundesdatenschutzgesetz* states: "The purpose of this Act is to protect the individual against his/her right to privacy being impaired through the handling of his/her personal data." (§ 1 I BDSG). "Personal data" are understood in a comprehensive way as, "any information concerning the personal or material circumstances of an identified or identifiable individual (the data subject)" (§ 3 I BDSG). A similarly broad definition also applies to the term "handling [of personal data]". In the case of spoken corpora it includes the collection, storage, processing and publication of such data (§ 3 IV BDSG). EU law includes more requirements for the processing of personal data (Directive 95/46/EC, Art. 6, 7 etc.), notably that the consent of the person affected is needed for the processing of personal data (Directive 95/46/EC Art. 8 I).

The data privacy officer of the respective institution, the federal state or the Federal Republic is responsible for the control of the observance of the data protection laws. Researchers should thus resolve data protection issues of data management and processing with their data privacy officer.

As far as spoken corpora are concerned, data protection regulations apply at least to audio and video recordings, transcripts and metadata about speakers.

The most important instruments for meeting the data protection requirements while handling and using spoken corpora are privacy policies of the "responsible bodies" who deal with the data, relevant and informed declarations of consent and suitable anonymisation and/or pseudonymization of data. This is further discussed in the two subsections below.

[5] Each federal state law on data protection contains a chapter which is explicitly devoted to the use of data for scientific purposes, for scientific research or in research institutions. Although other parts of the LDSGs apply to researchers' work as well of course, researchers and institutions should pay special attention to these parts: § 35 BW LDSG, Art. 23 BayDSG, § 30 BlnDSG, § 28 BbgDSG, § 19 BremDSG, § 27 HmbDSG, § 33 HDSG, § 25 NDSG, § 34 DSG M-V, § 28 NRW, § 30 DSG RLP, § 30 SDSG, § 36 SächsDSG, § 27 DSG-LSA, § 22 LDSG SH and § 25 ThürDSG.

1.1.1. Declaration of consent

The processing of personal data is only permitted by law if the **consent** of the person affected is obtained (Directive 95/46/EC Art. 30).

- Before collecting personal data and recording spoken interactions, a **written declaration of consent** should be obtained from every person involved (so-called "informed consent"). The BDSG (which applies to public bodies of the Federal Republic of Germany and non-public bodies, see details above, [*2016 note: soon to be replaced to a large extent by the rules of the EU's General Data Protection Regulation, EU GDPR*]) states that the "consent shall [only] be effective when based on the data subject's free decision. Data subjects shall be informed of the purpose of collection, processing or use and, in so far as the circumstances of the individual case dictate or upon request, of the consequences of withholding consent. Consent shall be given in writing unless special circumstances warrant any other form.[6] If consent is to be given together with other written declarations, it shall be made distinguishable in its appearance." (BDSG § 4a I)

- Some LDSGs (which apply to universities, as discussed above) permit the **processing of personal data without consent** for special research projects if there are certain qualifications, e.g. "protection-worthy interests of the person concerned will not be affected because of the type of data, due to their obviousness, or because of the nature of the use," or "the public interest in the research projects outweighs the data subjects' protection-worthy interests that qualify for protection and the aim of the research cannot be reached in another way." [7] In such cases, however, there might be additional obligations for anonymization of data and notification of the respectively responsible *Landesdatenschutzbeauftragten*, see e.g. § 27 HmbDSG.

- If there is not such a privileged case and a declaration of consent is needed, the consenting person needs to be informed about the following:
    - the name of the research project
    - contact details of the person who is in charge of the project
    - aims of the research project
    - information about if and in which way personal data is collected, processed, used and for how long they are stored. Thereby, the ways of processing and using which may differ depending on the file type (audio-, video recordings, transcripts etc.) should be elaborated if necessary (see details below).

[6] For example, this could be the case in still existing oral communities, especially if signatures (by experience) are associated with negative consequences.In such cases an audio-visual documentation of the Aufklärung can be "any other form" of consent.
[7] § 35 Sect. 1 BW LDSG, § 30 Sect. 1 BlnDSG, § 28 Sect. 1 BbgDSG, § 19 Sect. 1 BremDSG, § 27 Sect. 1 HmbDSG, § 33 Sect. 1 HDSG, § 25 Sect. 2 NDSG, § 30 Sect. 2 SDSG, § 22 Sect. 4 LDSG SH.

    - especially within international research projects it needs to be considered that (if provided or at least not precluded) the **processing of personal data by third parties** as well as their **transmission to bodies outside EEA** must be mentioned explicitly.

- Quite often, so-called **special categories of personal data** are collected for spoken corpora. The German legislator defines them as "information on a person's racial or ethnic origin[8], political opinions, religious or philosophical convictions, union membership, health or sex life" (§ 3 Sect. 9 BDSG, § 5 Sect. 1 S.2 HmbDSG and superordinated Directive 95/46/EC Art. 8 Sect. 1). If spoken corpora consist of special categories of personal data, the declaration of consent needs to refer to these data explicitly (see e.v. § 4a Sect. 3 BDSG).

- The people affected must be thoroughly informed about **planned ways of processing and using the collected data** before signing the declaration of consent. It may be advisable to provide the information for the people affected in advance, using a written data protection declaration which they may internalize by stating a simple "Yes, I agree with that". In any case, the declaration of consent should include language that refers to if and how information about data protection was given.

- In the case of minors and otherwise not legally responsible persons, the declaration of consent must be signed by their legal guardians. Minors after the 6th year of life are granted a "veto right" against the declaration of consent given by their legal guardians (which should therefore be requested).

- If the declaration of consent is to be given together with other declarations, it and the corresponding privacy policy shall be made distinguishable in its appearance, e.g. from other Terms and Conditions.

- Oral declarations of consent must be confirmed in writing.

- Electronic declarations of consent must be documented.

- A description of how and for whom the declaration of consent may be **revoked/retracted** in the future should be included.

- In every data protection law there is the rule of data economy: both the extent of data collection and the planned manners of use should be as little as possible, i.e. limited to aims of research and education. A wider limitation - e.g. the use for a special research aim or by a limited number of people - may indeed prejudice the

[8] Especially within spoken data it needs to be considered that information about the language biography (e.g. information about the mother tongue, information about the dialect) often allows one to draw conclusions about ethnicity. Such information should be understood as "special categories of personal data" within the meaning of the law and should be addressed as such.

general scientific reuse, but should still be offered to people affected, accompanied by explanations about the negative consequences of such limitations for the research in a non-technical way.

- If **archiving and publishing data** during or after the project phase is intended, this purpose must be stated explicitly in the declaration of consent. The nature of publication (e.g. in a database on the Internet) should be described in a way so that future changes of the way of publication (e.g. because of technical changes in the archiving and publishing system) are covered by the consent. Apart from that, publishing or archiving of a subset of the collected data shall be covered by the consent.

- It is a current practice to **restrict the group of users who shall have access to the data**. In practice, data protection can be accomplished by a password which is allotted only upon request. Often this makes it easier for people affected to give their consent.

- As a counterpart of the declaration of consent data users should sign a written **use declaration** which binds them to: first, use the data only for aims stated in the declaration of consent; second, blacken personal data in a publication based on this data as far as possible, and; third, not give this data to third parties.

1.1.2. Anonymization / Pseudoanonymization

The BDSG (that applies to public bodies of the Federal Republic of Germany and non-public bodies, as discussed above) describes the necessity for anonymization and pseudoanonymization of personal data as follows: “Personal data is to be collected, processed and used, and processing systems are to be designed in accordance with the aim of collecting, processing and using as little personal data as possible. In particular, personal data is to be aliased or rendered anonymous as far as possible and the effort involved is reasonable in relation to the desired level of protection.” (§ 3a). “Rendering anonymous” here means the modification of personal data so that the information concerning personal or material circumstances can no longer -- or only through a disproportionate amount of time, expense and labour -- be attributed to an identified or identifiable individual.” (§ 3 Sect. 6 BDSG) “Pseudoanonymization” means to replace a person’s name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult.” (§ 3 Sect. 6a BDSG).

- All 16 LDSGs (that apply to universities, see details above) lay down rules for the case that data may be processed without consent of people affected, privileged because of scientificity of aims. Depending on the regulation in each federal state, the data must be rendered anonymous or aliased, once the research aim allows it,

and if necessary, features that allow an de-anonymisation should be kept separately (and deleted as soon as the scientific aim allows).

- Some LDSGs contain regulations for cases in which **anonymization and pseudoanonymization is not possible**.[9]

- If anonymization or pseudoanonymization of data is pledged before using or publishing, it should be set down in writing in the **declaration of consent**.

- **Different kinds of data types** will require different methods to anonymize or pseudoanonymize:

- With **metadata and transcripts,** an appropriate level of replacing information (i.e. pseudoanonymization) may usually be achieved by replacing the names of people, geographical locations, etc., so that speakers may not be identified, or identified only through disproportionate effort.

- Within audio data the identification of speakers may be hampered by Verrauschen or fading of parts in which names are stated. However without any further processing (e.g. alienation of the audio signal) the opportunity to identify the speakers by their voices still exists.

- Within video data an extensive rendering anonymous of image data (e.g. by pixelizing faces or cutting in black mattes) this would mean that the resulting video is re-usable only to a very limited extent.

While signing the declarations of consent, affected people must be as informed as possible about which ways of anonymization and pseudoanonymization apply for the collected data while ensuring reusability. If necessary, the provision of data may be restricted depending on the type and the ability to render anonymous (e.g. anonymised audio data available for a greater, related video data only for a strictly limited group of users). Even this needs to be regarded adequately within the declaration of consent.

[9] § 34 Sect. 2 DSG M-V, § 28 Sect. 2 NRW, § 22 Sect. 3 LDSG SH.

1.2. Copyright aspects

For spoken corpora, both copyright and related rights may be issues, especially when it comes to:

- audio and video recordings from radio and TV broadcasts, where authors, producers, broadcasting companies, and others own certain rights
- audio and video recordings from the Internet (streaming platforms and other sources) where the operators of the platform may own rights
- written material that belongs to a spoken corpus as supplementary material (e.g. powerpoint slides for a speech, coursebooks for the class, etc.) and
- pictures, graphics etc.

As soon as these materials are used in the course of research, the consent of relevant rightholders is necessary to perform the research legally. A general research and education law regarding copyright and other rights has not yet been implemented in Europe, although such a regulation has been, and is, continuously discussed. Currently, only the quotation exception (§ 51 of the German Act on Copyright and Related Rights (UrhG)) and some special regulations for building personal scientific archives allow very limited use of someone else's work at all.

The consent of the rightholders is usually given through an appropriate license agreement (or contract). In practice, it is a considerable problem when rightholders are not known or cannot be found. This is important because every right holder must give his / her consent before a use of the work which is otherwise only permitted for right holders is allowed (with the exception of films, and if there are no other special agreements). If more than one person created the work, the consent of each co-rightsholder must be obtained.

This also refers to transcripts of primary data protected by copyright law (e.g. spoken and song recordings), even if the transcript is technically the work of the scientist in the sense of copyright. In such cases this transcript is considered a simple copy of the work which is included in the primary data or a derivative work (e.g. translation). Both of these types of use are assigned to the original rightholder (except in above-mentioned copyright exceptions, e.g. the quotation exception).

Extra precaution is appropriate if the copyright-protectable material has not yet been, but will be published within a scientific work in a manner which cannot be avoided due to the best scientific practice in disclosing sources. This affects the authors' right of personality because it is their choice whether their works are disclosed to the public or not.

References for Part 1

Enke, Harry, Norman Fiedler, Thomas Fischer, Timo Gnadt, Erik Ketzan, Jens Ludwig, Torsten Rathmann, Gabriel Stöckle, and Florian Schintke (2013). Leitfaden zum Forschungsdaten-Management. Verlag Werner Hülsbusch.

Häder, Michael (2009): Der Datenschutz in den Sozialwissenschaften: Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland, [formerly] available at:
http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf

Metschke, Rainer/ Wellbrock, Rita (2002): Datenschutz in Wissenschaft und Forschung. Materi- alien zum Datenschutz Nr. 28, [formerly] available at:
http://www.datenschutz.hessen.de/down- load.php?download_ID=147

Part 2: Information on legal aspects of the use of written corpora

2.1. Copyright and related rights

2.1.1. Basics

In general, texts are protected by copyright in Germany[10] if they satisfy an originality standard and it has not been more than 70 years since the death of their authors. How the originality standard is defined, and therefore how it is met, is a controversial question and its answer may differ from case to case and from court decision to court decision. The requirements for meeting the originality standard for copyright protection have been set lower and lower by courts over the past decades. Texts such as simple statements of the news or plain business correspondence may still not be protected by copyright because they do not meet the originality standard. But there is the concept of "kleine Münze", a sort of everyday creativity of people in general which is fully protected by copyright.

There are also certain related rights that are especially relevant for texts:

Since 2013, there is a related right for publishers in Germany which sidesteps the originality standard, which grants protection to even the shortest paragraphs for a term of one year. This protection follows mere publication, and is limited to the right of making publicly available. It is, therefore, only invoked when the press content is placed online.

There are two related rights that have a wider protected domain but a smaller scope of application. These include scientific editions of works that are not protected by copyright, and one concerning posthumous works, i.e. works that are published after the death of their authors and as the case may be after the copyright term (70 years, see above). These rights protect all uses of these works (not only for online use) for 25 years.

Finally, there is the related right for the creators of databases. This term is 15 years and is not related to the contents of databases, but to the manner in which it is structured. This related right does not apply to unstructured data and requires substantial

[10] We only give information about the legislation in the Federal Republic of Germany. How works of German authors are protected in other countries and how foreign authors are protected in Germany is regulated in some international conventions. The most important ones are the Berne Convention for the Protection of Literary and Artistic Works (usually known as the Berne Convention) and the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). In Art. 5.1. the Berne Convention says that every state party must acknowledge the protection of works of citizens of other state parties as it acknowledges the protection of works of its own citizens. There are 168 countries that are parties to the Berne Convention (i.a. the EU, the USA, China, Japan, Russia and India).

investments of time and/or money. Right holders (i.e. those who have created a database through substantial investment) are protected from "substantial parts" of the database being reproduced or used further.

Related rights are distinguished from copyright especially in two key ways. First, related rights have a shorter term of protection. Second, related rights can protect the works created by a legal person, e.g. a company. Under copyright, companies may at most have exclusive rights in copyright-protected works, while authors may only be natural persons.

The rules of copyright law that are most relevant for written corpora affect the right of reproduction (§ 16), the right of distribution (§ 17), the right of making works available to the public (§ 19a), the related rights on scientific editions (§ 70) and posthumous works (§ 71), the related right of makers of a database (§ 87b) and the related right of press publishers (§ 87f). It is still a legal gray area whether Text and Data Mining (TDM)[11] and thus quantitative linguistic analysis are types of use with copyright implications which are not yet mentioned in § 15 UrhG but protected nonetheless. (More specifically, whether the act of performing analysis on the data falls within the scope of § 15 UrhG; the resulting digital copy undoubtedly falls under § 16 UrhG.) Court decisions clarifying this issues can perhaps be expected in the foreseeable future. Because there are clear parallels between TDM and a human reading a text, which is not a type of use relevant for copyright, it is easily conceivable that courts may rule that TDM is permitted by law even without permission of the right holder, similar to reading.

2.1.2. Copyright exceptions and their application to written corpora

Laws that balance of interests of authors and users are so-called copyright exceptions. These determine which types of uses are allowed without the consent of the right holders, and under which circumstances. The use of copyright protected material as research data is only broadly provided for. The so-called research exception (§ 52a UrhG) for example allows making available "small scale" works as well as "individual articles from newspapers or periodicals," and only if and insofar this is "necessary" for the respective research purpose and is "justified" for the "pursuit of non-commercial aims". The copies may be made available "exclusively for a specifically limited circle of persons" which may include a small research team whose members -- according to the legal commentators Dreier/Schulze (2013) -- may be of different research institutions, or a seminar, but not the whole scientific community. The limited circle must be limited

[11] We adopt the term Text and Data Mining because it is now frequently used in discussions by the international legal community. At the moment, there is no coherent system of definitions of the different terms which are used for scientific analysis of data, but many slightly different and partly overlapping nomenclatures. It can be argued that the meaning of TDM in any case includes quantitative linguistic analysis.

to people who access the materials for their own scientific purposes[12] and the measures taken must be effective considering the state of the art at the time.

The right of temporary acts of reproduction (§ 44a UrhG) allows a temporary caching of electronic data, although this right is often insufficient to legally cover the empirical methods and replicable results required by scientific research. The same can be said for the right of reproductions for private use, which are permitted by § 53 I UrhG, but allows a transfer only in private, i.e. not in work-related scientific field, and § 53 II UrhG, which allows a reproduction only for one's own personal scientific use (the possibilities of transfer are regulated in § 52a). The right of digital reproductions of complete books or magazines is further limited in § 53 IV UrhG. Concerning all the exceptions, one must keep in mind that they are subordinate to contrary license agreements. Additionally, § 52a IV UrhG states that an equitable remuneration shall be paid (guided by rates set out in the VG WORT case).[13]

Attention should also be paid to the fact that exceptions of copyright protection do not apply for related rights in the same way. The have their own respective protection exceptions that are named in the respective part of the UrhG.

As a conclusion it can be said that legal exceptions are typically not a sufficient basis for making available written corpora permanently. Making available a copy of the written corpus that has been the research object is not covered by any of the above mentioned research exceptions what may complicate the repeatability and thus the verification of respective research projects massively. Often enough even building up a corpus of texts for which no express permission was given, is unlawful because the digital copies produced in the process are not necessarily covered by copyright exceptions.

For building a corpus in conformity with the law, the consent of the right holders must be obtained, or it must be ensured that only texts are used:

- that are not protected by copyright, such as the text of laws, certain government documents, etc.
- where the term of copyright protection has expired, or
- where the texts do not meet the originality standard.[14]

A thorough checking/clarification of rights is therefore necessary. The costs for this may possibly be reduced by cooperating with other centers that that seek to use this data and therefore check its legal status.

[12] BT-Drucksache 15/38, S. 2

[13] In its decision of March 24, 2011 (file reference 6 WG 12/09) concerning the case VG Wort - Federal States, the higher regional court (OLG) of Munich considered a remuneration of 10 euros + VAT per work as equitable for scientific research within the scope of § 52a I Nr. 2 UrhG.

[14] Whether the originality standard applies to a succession of randomly assorted sentences is unclear. § 39 UrhG “Alterations of the work,” which belongs to the moral rights, is one argument that this method is not legally sound.

The situation tends to be considerably easier, if the intended use is covered by standard licenses which grant the necessary rights for the use in a corpus, to everyone, in advance. These are called "Public Licenses". In best-case scenarios, the author has already published his / her texts under a sufficiently liberal standard license. But often this is not the case. This means that individual license agreements with the respective right holders must be made, which requires time and other resources. In the case of texts published by presses / publishing houses, these may typically be contacted directly, because the publisher often obtains the right to license electronic uses in their contracts with authors. The same often applies to texts which are published on web portals, because operators are often granted the respective rights through "Terms and Conditions" agreements.

2.1.3. Adaptations (derivative works) and transformations

Adaptations in the meaning of the law are contents that are based on a previous work and meet the originality standards to qualify for protection (the law of copyright in adaptations), even if the previous work is not longer protected by copyright. If the previous work is still protected by copyright, adaptations may only be published with the consent of the author of the previous work. Transformations are, according to prevailing legal opinion,modified versions of previous works that do not meet the requirements of protection for copyright in adaptations. They also may also only be published with the consent of the author of the previous work.

The threshold to adaptation or transformation is reached if the an average observer's impression of a work is changed noticeably. Concerning pictures, this is for example the case if they are cropped or their sizes changed extremely. For films, e.g. if they are musically rendered. Texts are changed noticeably if they are shortened, amended, mixed with other texts or translated. A new layout or a transmission of a text from analogue to digital form is not an adaptation or transformation -- although usually a reproduction -- meaning generally when a text is removed from its original medium/context and remains recognizable as a discrete work. (In exceptional cases the change of the context of the work may result in an adaptation. For text corpora for research purposes, however, this is hard to imagine.)

When the original work is no longer recognizable by an average observer, no adaptation exists, but rather a new, independent work. Here, courts have said that the personal characteristics of the pre-existing work "fade away" from the new content.[15] The difference between an adaptation (§ 23 UrhG) and an independent work created in free use (§ 24 UrhG) is, however, fluid.[16] If, on the surface, the new content has nothing

[15] Federal Supreme Court of Germany in "Mecki-Igel I", GRUR 1958, 500, 502.
[16] See Dreier/Schulze, Urheberrechtsgesetz Kommentar, 4. ed., § 24 Marginal No. 1 and § 23 Marginal No. 4. [*2016 note: a newer 2nd edition was published in 2015*]

in common with the previous material, free use of the previous material is unproblematic (as far as the law of adaptations). Often, this results from the method which is used within Text and Data Mining. If a text, for example, is statistically analyzed or annotated, it can usually not be reconstructed from the emerging statistics or annotation. Thus both research results are not adaptations of the source text within the meaning of the law.

For source texts that are still protected, this does not solve the problem of contractual terms that prohibit temporary copies / caching that is technically necessary for the development of research results and making the texts permanently available only with consent, (see above). Apart from that, TDM may also be contractually prohibited because civil law largely allows contracting parties to agree on what they wish (the law of "private autonomy"). If an editor for example forbids TDM or the publication of TDM results, based on a text within a license agreement with a scientific institution that regulates the access to the material, this must be respected, even if the research results are independent and not adaptations or transformations and TDM should a priori not be regarded as a copyright protected type of use.[17] In this case, the basis for enforcement of the prohibition is not the copyright law, but the contract which was entered between the two parties. Such a contract, however, affects only the relevant parties.

It is possible to incorporate certain conditions for the use of the material in the agreement instead of a strict prohibition. This can be executed even by standard licenses, which are contracts. It is therefore conceivable that research or TDM results are made subject to copyleft terms.[18] Disregarding software licenses, however, it is absolutely not common that the conditions of standard licenses impose conditions independently of any existing legal position based on an absolute right (such as copyright or database protection). The six Creative-Commons licenses even explicitly state that they do not restrict anything that the licensee is allowed to do without the license anyway.[19] Their copyleft terms thus only apply under the pre-condition that there is a legal protection in the first place that requires permission of a rightsholder.[20] Thus copyleft and other limitations of CC licences would only be effective if TDM is regarded as a type of use within the meaning of the copyright law.

Since this question is not yet resolved everywhere in the world, the new CC license version 4.0 clarifies explicitly that the results of TDM should not be considered as an

[17] Whether TDM can be regarded as a type of use is currently being discussed by jurists and will certainly keep the courts busy. There are many reasons why TDM should be regarded as a kind of reading, which is as so-called *Werkgenuss* permitted without consent. See above 2.1.at the end.
[18] Meaning, that the licensee must offer the licensed content to the public under identical or similar condtions.
[19] See e.g. section 8.d. in the license CC-BY Version 4.0.
[20] The data bank licenses of Open Data Commons are an exceptional case, because these postulate their copyleft-conditions even for those regions of the world where no database protection law exists, e.g. the United States.

adaptation by the licensor. Thus neither the copyleft conditions of CC licenses[21] nor the other conditions "attribution," "no commercial use" and "no edits allowed" need to be taken into account, as far as TDM and its independent results are concerned.

If research results are still somehow considered adaptations or transformations within the meaning of the law, i.e. outside of TDM and without other licenses influencing the character of adaptation, the same recommendations apply for further use of these research results as for the use of independent works.

2.1.4. Collections and database works

According to § 4 UrhG, collections of works and databases are protected where the selection or arrangement of the elements constitute the author's own intellectual creation, regardless of whether the individual elements are protected or not. This may be relevant if collections of texts in the public domain are included in a corpus.

This protection of "databases works" should not be confused with mere databases, whose creators are additionally protected by §§ 87a - 87e UrhG (see above). The related right of the maker of a database only requires substantial investment; in contrast, a "database work" requires such an extraordinary arrangement of the content that the arrangement itself can be regarded as a creation (similar to authorship). Thus, the threshold for the (high) level of protection of a "database work" is much higher than those for a database protected in accordance to §§ 87a et seq. UrhG. The latter right of the maker of a database place may create restrictions of use if parts of a database are included in a corpus or such a corpus is made available.[22]

2.1.5. Orphan works

After § 61 UrhG was inserted into the Copyright Act in 2014, there are now some types of uses permitted by law concerning text works from collections of publicly accessible libraries, educational institutions, museums and archives, if they are already published and the respective right holders can not be found or identified even by a diligent search (defined in § 61a UrhG), and this research result was recorded in a central register. The permitted types of use concern making available to the public (§ 19a UrhG) and reproduction (§ 16 I UrhG). Since the right to create derivative works is not included, it may not be possible to rely on § 61 UrhG when using such works in corpora.[23] To take

[21] The name for the copyleft mechanism of CC licenses is "share alike", abbreviated as "SA".
[22] See the court decision of the European Court of Justice (October 9, 2008, Case C304/07) and of the Federal Supreme Court (August 13, 2009, file reference I ZR 130/04)
[23] At the copy deadline of the present document, this was still an open question. Any news on this point will be published on the CLARIN-D Legal Information Platform. [*2016 note: in general, it seems that adapted versions are not covered by § 61 UrhG*].

the path of least legal risk, orphan works should only be included in corpora in a way whereby no adaptation or transformation is carried out (see above).

There is still the unavoidable problem that the status of an orphan work may subsequently expire if the right holders appear and/or become known. From this point in time, the usual rules for the use of works apply again.

2.1.6. Software

The terms of use of commercial software are usually clearly laid out, in order to decide the terms under which it may be used and what implications may arise when such software is used to create independent and derivative works. Depending on the approach, the output of the software, i.e. the research result or document, remains independent in its legal status from that of the software.

Sometimes the legal status is more vague within software tools that were developed in an academic context, as they are often based on data (dictionaries or training corpora) which might be affected by third party rights.

For software developed in-house, it needs to be noted that the decision if and under which license the software will be released is reserved for the employer for whom the software was created (§ 69b UrhG).

2.2. Data protection issues

In compiling and making available written corpora, data protection rights may also be affected. This is particularly true within the use of texts which were not primarily intended for further use or disclosure, such as chat transcripts. In the case of such texts, it may be reasonable for ethical reasons and in order to avoid subsequent legal disputes to do a pseudoanonymization / anonymization (see part 1 : 1.1.2, above). Besides, the same legal requirements (adequate information and consent of those affected) apply, as they are explained in the first part concerning the oral corpora.

Generally it should be considered that after a release of a corpus, people who are affected because texts about them were made available may ask to be removed from the corpus. In such a case a weighing of interests of the people affected in the individual case needs to be done. Courts are gradually conceding[24] constitutional fundamental

[24] The case brought up by an offender released from prison against the University of Leipzig on deletion of his name from a corpus which was made available on the Internet (vocabulary project) was upheld at first instance before the District Court Hamburg (file reference 324 O 243/07). In the second instance it was rejected before the Higher Regional Court of Hamburg (file reference 7 U 123/09 ; the full texts of both decisions are not available on the Internet free of charge)

rights to scientific institutions on the basis of the criteria for legal admissibility of permanent online press archives.[25] However, it must be assumed that in case of doubt, the personality rights of individuals outweigh the interests of a particular researcher, unless it concerns a person of historical/journalistic interest who appears legitimately in the press. Whether the same is true with respect to the interests of an entire branch of research could only be decided by a court. Regarding the still existing problem of persistence of research data, there is a certain pragmatic consensus within the scientific community: text deletions because of personality rights should be considered acceptable also epistemologically, since the replicability of important and methodically valid research results does not depend on individual texts. What is probably more important is de facto the organizational effort that can be caused by individual deletions. It is recommended to factor this into project costs in advance, if possible. For basic information and further data protection issues, see Part 1, Section 2.

2.3. Best Practices

2.3.1 Recommendations for building corpora

- **In case of doubt, you should try to obtain licenses and consent.**
  Right holders are usually cooperative when it comes to non-commercial, scientific purposes and no economic or other interests are violated e.g. by an unrestricted distribution of copies.

- The attempt to get licenses should begin **as early as possible in the planning phase** of a project, since the negotiations may drag on over a long period of time and this is the only way to ensure that the necessary rights may be obtained before the project starts and therefore any **license fees** or other rewards may be included in the **calculation of the project costs**.

- Also as early as possible in the planning phase, **a center should be approached** that is experienced with licensing of the relevant type of resource. It may provide assistance or in some circumstances take care about obtaining the licenses, and at the same time ensure that the licensing terms are drafted so that the data and the results of the projects may be included into their own archives/projects after the duration of the project and made available for the long term.

- Recommendations for the draft of license agreements can be found in Perkuhn et. al. (2012, p. 53) and on the CLARIN-D Legal Information Platform.[26]

[25] Federal Supreme Court of Germany, court decision from December 15, 2009, p. 757 et. seq. with further references; Federal Constitutional Court, Court decision from June 5, 1973, BVerfGE 35, p. 202 et. seq.
[26] http://clarin-d.de/legalissues

- License agreements typically have a **limited term**, especially if they are associated with fees. Particularly in these cases, it is recommended to develop a strategy in cooperation with a center for making the content sustainably available. It also should be noted that **unintentional interpretations of the licensor** can prevent license renewals and additional licenses regardless of their legality.

- In cases where it is not possible to obtain sufficient rights to make available a text corpus to the scientific community permanently, but the reasons to build the corpus were nevertheless strong enough, the reasons should be documented and **compromise strategies** should be found on how a sustainable availability may be achieved at least rudimentarily. One possible model is e.g. to comprehensibly document how they may obtain the necessary rights themselves for subsequent users.

- Data protection issues should already be included in the planning phase of a project. If it is intended to collect personal data to a greater extent, an explicit document on the subject should be created and maintained (data protection concept). It must be captured which data is collected for which purposes. If necessary, appropriate consent declaration forms need to be developed and to be signed by the people affected by the data processing.

2.3.2. Recommendations for making written corpora available

- It is usually necessary and common practice to **limit the number of users** of corpora to people who identified and agreed with an End User License Agreement (see below) and, if necessary, additional data protection regulations. De facto this can be achieved by e.g. data access regulations via passwords which is allocated only on application and only in person or via a DFN-AAI-Authentication and web forms to request consent.

- As a general rule, rights and obligations which result from licensing agreements between right holders and corpus provider, need to be passed on to end-users via **end user license agreements** and **data privacy policies** (for example if a corpus provider undertakes an obligation to the licensor to document access to the corpus).

- With regard to personal data, anonymization and pseudoanonymization should be considered when making corpora available.

2.3.3. Recommendations for creating and making own works available: derivative works and databases

- Works that are **created by scientists themselves should always be released under license terms**, in order that subsequent users in the future may know if they can use the work for their own purposes. At the same time, contents that are (or become) free of copyright and on which the scientist did not acquire any other rights should not be portrayed as protected by law, and as far as possible explicitly marked as unprotected, e.g.with the help of "Public Domain Mark" (PDM).

- When selecting license terms, **existing, widely-used standard licenses that are as liberal as possible should be used** (e.g. one of the two Creative Commons licenses recognized in terms of the Open Definition[27], namely Creative Commons license versions BY and BY SA, or for software, a GNU license or BSD or Apache licenses which refrain from copyleft). So the result is most likely like the Open Access approach. The increasing trend is to publish scientific works with not more limitations than the Creative Commons license type "CC BY - Attribution," while pure data should be licensed entirely free of restrictions by "CC0". Even scientific publishers are increasingly open to such licenses.

- Particular attention should be paid to **indicating the license as accurately as possible and easy to find**.

- **Problems with derivative works** may be avoided in some cases, for example when annotations are published as an independent work from which the original work can not be reconstructed. If the license which is advised for a derivative work is roughly equivalent to the underlying, the same license should be used to facilitate the reusability. In any case, provisions of the license of the underlying work that sometimes allow only certain licenses for later processing (see e.g. the “Share-Alike” clauses in Creative Commons licenses[28]) should be noted .

2.3.4. Recommendations for the use of software when creating derivative works

- If no license terms are known, one should attempt to determine if and which restrictions apply to the use of the software.

[27] http://opendefinition.org/od/
[28] See the variety of content which is combined under different Creative Commons licenses, https://wiki.creativecommons.org/FAQ#Can_I_combine_material_under_different_Creative_Commons_licenses_in_my_work.3F

- Particularly with commercial annotation tools, it may be reasonable to clarify and set out in a supplementary agreement the extent that the outputs of the software may be distributed, because software license provisions often prohibit this altogether. Generally, however, only reverse engineering is to be prevented.

- Before using or licensing software, it should be clarified to what extent the outputs of the software may still be used after the license term expires.

References for part 2

Dreier, Thomas / Schulze, Gernot (2013): Urheberrechtsgesetz: UrhG. Urheberrechtswahrnehmungsgesetz, Kunsturhebergesetz, Kommentar. 4. Aufl. München: C.H.BECK [*2016 note: a newer 2nd edition was published in 2015*]

Kamocki, Pawel / Ketzan, Erik (2014): CLARIN-D Legal Information Platform, available at http://clarin-d.de

Kamocki, Pawel / Ketzan, Erik (2014): *Preparation of corpora from online and other resources: current state of German and EU law*, 7. Arbeitstagung des Empirikom-Netzwerks: "Social Media Corpora for the eHumanities: Standards, Challenges, and Perspectives", 20.02.2014.

Perkuhn, Rainer / Keibel, Holger / Kupietz, Marc (2012): Korpuslinguistik. - Paderborn: Fink, 2012. (UTB 18)

Participants list: DFG roundtable meeting "Spoken corpora"

November 9, 2012
DFG office, Kennedyallee 40, Bonn (Germany)

Professor Dr. Bernt Ahrenholz	Friedrich-Schiller-University in Jena
Dr. Jörg Bücker	Westfälische Wilhelms-University in Münster
Professor Dr. Kristin Bührig	University of Hamburg
Professor Dr. Arnulf Deppermann	Institut für deutsche Sprache, Mannheim
Dr. Sebastian Drude	Max Planck Institute for Psycholinguistics
Dr. Sigrun Eckelmann	DFG in Bonn
Dr. Oliver Ehmer	Freiburg
Professor Dr. Christian Fandrych	University of Leipzig
Professor Dr. Caroline Féry	Goethe-University in Frankfurt am Main
Professor Dr. Ulrike Gut	Westfälische Wilhelms-University in Münster
Professor Dr. Rüdiger Harnisch	University of Passau
Dr. Dagmar Jung	University of Cologne
Professor Dr. Roland Kehrein	Philipps-University in Marburg
Dr. Kerstin Kucharczik	Ruhr-University in Bochum
Dr. Christoph Kümmel	DFG in Bonn
Slawomir Messner	Philipps-University in Marburg
Dr. Gaia di Lucio	Bonn, PT-DLR
Professor Dr. Bernd Meyer	Johannes Gutenberg-University of Mainz
Ludger Paschen	Ruhr-University of Bochum
Professor Dr. Stefan Pfänder	Albert-Ludwigs-University of Freiburg
Dr. Christoph Purschke	Université du Luxembourg
Professor Dr. Uta M. Quasthoff	Technische University of Dortmund
Professor Dr. Angelika Redder	University of Hamburg
Dr. Ines Rehbein	University of Potsdam
Professor Dr. Christian Sappok	Ruhr-University of Bochum
PD Dr. Florian Schiel	Ludwig-Maximilians-University of Munich
Dr. Thomas Schmidt	Institut für deutsche Sprache, Mannheim
Professor Dr. Stavros Skopeteas	University of Bielefeld
Adriana Slavcheva	University of Leipzig
Jan Strunk	Köln
Dr. Vera Szöllösi-Brenig	Volkswagen-Stiftung
Professor Dr. Doris Tophinke	University of Paderborn
Dr. Helga Weyerts-Schweda	DFG in Bonn
Professor Dr. Heike Wiese	University of Potsdam
Dr. Kai Wörner	University of Hamburg

Participants list: DFG roundtable meeting "Text corpora"

November 15, 2013
DFG office, Kennedyallee 40, Bonn (Germany)

Dr. Noah Bubenhofer	Technische University of Dresden
Professor Dr. Arnulf Deppermann	Institut für deutsche Sprache, Mannheim (IDS)
Professor Dr. Dagmar Deuber	Westfälische Wilhelms-University of Münster
Dr. Eva-Maria Dickhaut	Akademie der Wissenschaften und der Literatur Mainz
Professor Dr. Mechthild Habermann	Friedrich-Alexander-University of Erlangen-Nürnberg
Dr. Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Professor Dr. Thomas Gloning	Justus-Liebig-University of Gießen
Professor Dr. Iryna Gurevych	Technische University of Darmstadt
Professor Dr. Ulrich Heid	Stiftung University of Hildesheim
Professor Dr. Gerhard Heyer	University of Leipzig
Professor Dr. Erhard W. Hinrichs	Eberhard-Karls-University of Tuebingen
Professor Dr. Martin Huber	University of Bayreuth
Professor Dr. Magnus Huber	Justus-Liebig-University of Gießen
Professor Dr. Wolf Peter Klein	Julius-Maximilians-University of Würzburg
Dr. Marc Kupietz	Institut für deutsche Sprache, Mannheim (IDS)
Professor Dr. Gerhard Lauer	Georg-August-University of Göttingen
Professor Dr. Christian Mair	Albert-Ludwigs-University of Freiburg
Professor Dr. Alexander Mehler	Goethe-University of Frankfurt am Main
Professor Dr. Roland Meyer	Humboldt-University of Berlin
Professor Dr. Manfred Pinkal	Universität des Saarlandes
Dr. Roland Schäfer	Freie University of Berlin
Professor Dr. Ingrid Schröder	University of Hamburg
Dr. Silke Schwandt	Goethe-University of Frankfurt am Main
Professor Dr. Manfred Stede	University of Potsdam
Professor Dr. Angelika Storrer	University of Mannheim
Professor Dr. Elke Teich	Universität des Saarlandes
Dr. Helga Weyerts-Schweda	DFG in Bonn
Dr. Stefan Winkler-Nees	DFG in Bonn
Professor Dr. Heike Zinsmeister	University of Hamburg