Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation

Timm Lehmberg

Dr. Georg Rehm

Dr. Andreas Witt

Felix Zimmermann

Abstract

Comprehensive data repositories are an essential part of practically all research carried out in the digital humanities nowadays. For example, library science, literary studies, and computational and corpus linguistics strongly depend on online archives that are highly sustainable and that contain not only digitized texts but also audio and video data as well as additional information such as metadata and arbitrary annotations. Current Web technologies, especially those that are related to what is commonly referred to as the Web 2.0, provide a number of novel functions such as multiuser editing or the inclusion of third-party content and applications that are also highly attractive for research applications in the areas mentioned above. Hand in hand with this development goes a high degree of legal uncertainty. The special nature of the data entails that, in quite a few cases, there are multiple holders of personal rights (mostly copyright) to different layers of data that often have different origins. This article discusses the legal problems of multiple authorships in private, commercial, and research environments. We also introduce significant differences between European and U.S. law with regard to the handling of this kind of data for scientific purposes.

Introduction

This article approaches the topic of digital book repositories from an unusual angle—that of an applied research project in the field of computational linguistics that is concerned with sustainably archiving and providing access to large and heterogeneous collections of language data. At first

LIBRARY TRENDS, Vol. 57, No. 1, Summer 2008 ("Digital Books and the Impact on Libraries," edited by Peter Brantley), pp. 52–71

glance these linguistic corpora and digital book repositories do not seem to have many things in common, but, at the end of the day, both are collections of digital texts, generally delivered to the user by some kind of Web-platform. This common ground and the experiences we had in the above-mentioned project allow us to speculate about several functional aspects future digital book repositories need to take into account, especially with regard to current trends in Web-based content creation and aggregation.

Linguistic corpora are databases used for applied and empirical research in linguistics and related areas. A corpus is a collection of digital texts that are annotated with linguistic analysis information using XML-based markup technologies (see, for example, Burnard & Bauman, 2007). It is common practice to use transcribed speech or dialogues, digital books, scientific papers, or newspaper articles as the source material or primary data of such a corpus. Custom tools are employed to process these sets of primary data by adding linguistic information, for example, phrase, sentence, and paragraph boundaries, part-of-speech and morphological data for every single word, or syntactic trees that describe the grammatical structure of all or only selected sentences. Nowadays, added information such as these is usually stored in separate files that reference the primary data contained in yet another file (in contrast to embedded annotation, its predecessor, this approach is called stand-off annotation), for example, you can have one file each for the text structure layer, part-of-speech layer, the morphology layer, and the syntax layer. As the annotation of a corpus is an extremely complex and time-consuming task, it is not uncommon for one research group to extend a corpus, initially created by another group, by adding further annotation layers. From a legal point of view, a most interesting situation emerges: typically, the primary data is copyrighted material used by some kind of agreement with the publishing house or other copyright holder; while the annotation layers refer to the primary data, they are independent works on their own. As a consequence, each research group that creates one such layer has the right to decide its terms of distribution, that is, to apply individual licenses.

This situation can be directly transferred into the domain of digital book repositories. Let us imagine, for example, a scenario in which such a repository provides two digital books on a certain topic. Book A is freely available under a Creative Commons license, book B is commercially available for a given fee. Future digital book repositories will have novel functions, especially with regard to currently popular Web 2.0 approaches such as social networking, blogging and probably microblogging, and content aggregation. Future digital book platforms will allow us to excerpt and to rearrange pieces of the available books using methods that essentially work like copy and paste, perhaps in order to quickly assemble a collection of important notes, quotes, and diagrams on a specific topic. The legal aspects of such functions are both interesting and complex. While access

to the sections taken from book A could be given under no specific terms of use, the system could only grant access to the sections taken from book B if the user is able to provide proof of purchase of the original book.

The following section provides additional details with regard to linguistic resources such as corpora and treebanks. We briefly present three case studies that describe one such corpus, a well-known tool for the transcription and analysis of spoken language data and critical editions. One of the conclusions of this discussion is that, from a legal point of view, linguistic resources can be compared to mashups. We then describe the Web 2.0-related phenomenon, and characterize the complex legal situation. Finally we present the implications of this discussion for our work with regard to linguistic resources and linguistic mashups.

Sustainability of Linguistic Resources—Three Case Studies

It is the goal of linguistic sustainability initiatives to archive and to make available heterogeneous sets of linguistic resources, that is, not only corpora but also linguistic software, so that interested parties are able to access them (Dipper et al., 2006). Nowadays researchers predominantly work with empirical data, they use and they create corpora, normally with a linguistic theory and a specific research question in mind. When a project is finished it can be very difficult to gain access to its corpus. In an ideal world, academics can turn to a sustainability initiative (also referred to as preservation projects) in order to archive their datasets (Trilsbeek and Wittenburg, 2006) and to make the data available to other researchers, for example, by means of a Web-based corpus repository.

The joint project Sustainability of Linguistic Data in which three of the four authors work, processes the language data from the research centers SFB 538 ("Multilingualism"), SFB 441 ("Linguistic Data Structures") and SFB 632 ("Information Structure"), each funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG). These three centers have collected vast amounts of data over a period of several years. The collection contains a total of about sixty-five linguistic corpora that, among others, consist of written and spoken language, synchronic and diachronic data, hierarchical and timeline-based markup, as well as lexical resources. According to rough estimates it took more than one hundred person years to collect and to annotate these resources.

Our goal is to convert this collection into a comprehensive, homogeneous, and sustainable linguistic format that can be easily imported into a Web-based platform to be accessible and usable by researchers and applications for at least five decades. In addition to the implementation of the platform, our work is concerned with several related areas, for example, appropriate data and metadata formats (Schmidt et al., 2006, Wörner et al., 2006; Witt et al., 2007) and the incorporation of an ontology of lin-

guistic concepts into a user-friendly search interface (Rehm, Eckart et al., 2007). Legal aspects are important because our linguistic resources usually comprise multiple layers of primary data and annotated information (see Lehmberg, Chiarcos, Rehm, et al., 2007; Zimmermann & Lehmberg, 2007; Newman, 2007). It is not uncommon for these individual data layers to have *multiple origins*—these directly translate to *multiple licenses* that we have to take into account for a *single resource*, which, in turn, means that we have to provide highly specialized access restrictions so that access to data layers can be controlled separately.

Case Study 1: EXMARaLDA

Spoken language corpora are collections of authentic spoken language to be used for linguistic research. An integral part of this type of corpus usually are literal transcriptions of recorded audio or video data that are annotated with linguistic information, for example, prosody, discourse, morphology, or syntax. Several tools exist for the creation of transcriptions, each using different types of data formats and transcription standards. In our sustainability initiative we use EXMARaLDA (Extensible Markup Language for Discourse Annotation, see http://www.exmaralda.org), a collection of applications and XML-based data formats that provide features for creating, analyzing, and exchanging not only single transcriptions but entire corpora of spoken language. The tools—an editor for transcriptions, a corpus manager for administrating corpus metadata, and a concordance tool—are being developed at SFB 538 "Multilingualism" and freely available. The main objectives in the development of EXMARaLDA are (Schmidt & Wörner, in press): to facilitate the exchange of spoken language corpora between researchers and technological environments (e. g., different operating systems, different software tools); to make best use of the multimedia and hypertext capabilities of modern computer systems while working with video or audio data and their transcriptions (e.g., to develop ways of synchronizing the navigation in the recording with the navigation in the transcript); to pave the way for long-term archiving and reuse of costly and valuable language resources (e.g., to ensure the compatibility of corpora with existing or emerging standards for digital archiving).

Transcriptions that are created with the help of the EXMARaLDA editor use a musical score notation of data and annotations that can be aligned to their respective primary audio or video files (Schmidt & Wörner, 2005). Additionally, the use of transcribed spoken language data for linguistic purposes requires the collection of extensive metadata containing speaker information such as age, nationality, linguistic and social background, as well as information on situational contexts, date, and location. Hence, spoken language corpora created with EXMARaLDA can be seen as a composition of multiple layers of textual as well as multimedia

data that come from multiple sources and locations. These characteristics raise a number of questions regarding copyright (of the transcription and annotation layers that normally is carried out by multiple researchers) and data protection (concerning the large amount of personal information contained within the primary data).

Case Study 2: TüBa-D/Z

A corpus consists of two parts: (a) one or more authentic source texts of a single or multiple genre, and (b) one or more layers of annotation that refer to linguistic properties of the texts (morphology or part-of-speech information, document structure, etc.). The linguistic properties are annotated manually by academics or automatically by software tools; the source text collection (STC) has been acquired beforehand from third parties such as websites or publishing houses. In practically all cases the STC is a copyrighted property that is subject to access restrictions. Ultimately, it is up to this copyright holder to decide if, and under which conditions, the complete linguistic resource—a crucial part of which is the STC—can be made available to the public or research community.

TüBa-D/Z (Tübingen Treebank of Written German; Telljohann, Hinrichs, & Kübler, 2004, 2006) is a treebank, that is, linguists analyzed all sentences in terms of their syntactic structures and added syntax trees. The corpus is based on a commercially available CD-ROM that contains an archive of the newspaper *die tageszeitung* (*taz*). TüBa-D/Z currently consists of about 27,000 sentences.

If a researcher (the licensee) wants to obtain TüBa-D/Z, available for academic purposes free of charge, he or she has to sign a license agreement with the Linguistics Department at Tübingen University (the licenser). It states that the licenser is the copyright holder of the linguistic annotation and that the STC, as published on the CD-ROM, is copyrighted by the company contrapress media GmbH. The licensee therefore has to certify that he or she or the institution the person works for has a valid licence of the CD-ROM; furthermore, a copy of the CD-ROM invoice has to be submitted as additional proof.[1] Only if the licenser receives the signed agreement and a copy of the invoice, can the licensee be sent the access information for the password-protected TüBa-D/Z download site.

Rehm, Witt, Zinsmeister, and Dellert introduce the masking of linguistic resources, in order legally to bypass licensing restrictions such as the ones described above. The idea is to mask the STC, but not the layers of linguistic annotation. This approach practically removes the STC, so that the original licensing and copyright restrictions no longer hold for the new resource. The advantage is that the information that is most crucial and most interesting to other linguistics researchers, the annotation itself, can be made available *without* any restrictions (see Rehm, Witt, Zinsmeister et al., 2007a, p. 166).

This solution is possible because the institution that created the linguistic annotation is its copyright holder. Therefore, it is up to the institution to decide the conditions under which a masked linguistic resource is to be made available to third parties, because different licenses apply to the different constituent parts of the corpus. Usually, a research institution tries to make a resource available online at no cost. Nevertheless, modern corpora may be comprised of *multiple* annotation layers that have been created by more than one research group. As each group can be considered the creator of their own annotation layers, it can decide their terms of distribution. The common practice of taking an existing standard corpus and adding another layer of analysis not only extends a corpus with more linguistic analysis information, it also adds another layer of legal restrictions. Commercially available software tools used in annotating the corpus might restrict its terms of distribution as well.

There are two aspects of corpus masking that we would like to emphasize. First, a tool was developed that is able to mask corpora on the fly and can be integrated into a Web-based corpus delivery platform (Rehm, Witt, Zinsmeister et al., 2007a, 2007b). Should someone who is interested in a corpus not have a valid license for the STC, he or she can still receive the corpus, albeit in masked form. Second, a linguistic corpus potentially can be associated with *several* accessibility regulations. For example, full access to the TüBa-D/Z treebank requires the licensee to have a valid license of the *taz* CD-ROM, whereas the masked version of TüBa-D/Z can be placed under, say, the GNU Free Documentation License. As a consequence, not only sustainability initiatives but also digital book repositories have to come up with very flexible systems of representing the relationships and dependencies between the digital books or the source texts respectively, as well as the different layers of annotation and their corresponding license restrictions: if one or more layers whose license regulations are very restricted are removed from a corpus that is about to be delivered, the next restrictive license of the remaining part of the corpus needs to be applied. This representation should be included in the metadata records of any corpus and a corresponding process logic should be integrated into the platform (Rehm, Eckart et al., 2007).

Case Study 3: Critical Editions

The final case study demonstrates that the legal questions that arise when dealing with linguistic corpora are much more common than they may appear. Classic and influential literary works are often not only published as regular books, but also as critical editions. These scholarly editions have their origin in literary studies and complement the original text with additional information. Their preparation is a large-scale and time-consuming endeavor and culminates in a publication that includes, in addition to the original literary work, a multitude of supplementary facts, for

example, notes or letters written by the author, relevant sources to enable the reader to understand old metaphors or idiomatic expression, and explanations concerning the impact of the original piece. Usually these editions are the result of long-term research projects financially supported by universities, research foundations, or publishers. Critical editions are often based on previous scholarly editions.

To give an example, the most recent Norton Critical Edition of *Moby Dick* not only includes the Northwestern-Newberry text of the work, but also biographical information compiled by Hershel Parker, prose and graphics on whaling by John B. Putnam, reactions on the first publication of Melville's book, a chapter titled "Posthumous Praise and the Melville Revival: 1893–1927" including a text by William Faulkner, and several other resources. Due to the fact that the texts in such an edition as well as the edition itself are created by multiple authors, every single text—unless the duration of copyright protection has not expired—as well as the entire edition are capable of being protected by copyright law. Because of its age the original text of *Moby Dick* is no longer copyrighted, while the copyright of additional texts included in the Norton Edition is held by their respective authors. The Norton Critical Edition as a whole is copyrighted by its publishing house, W. W. Norton and Company. The publisher usually makes sure that the individual copyright of texts included in the edition is kept. In a number of countries, for example, in Germany, there are specific regulations for the protection of critical editions and derivative works but they all have in common that editions, inasmuch as their creation requires a special amount of creativity, are protected by copyright law.

The production of a printed book is an expensive process. Especially lesser known authors only sell in small figures—critical editions of these books are not very attractive to publishing houses. This will, inevitably, lead to scholarly projects that create critical editions which will be available in digital form only, ideally free of charge for the research community and independent from any publishing houses.

These future versions of critical editions are very similar to the linguistic resources described above. Primary sources that are, in most cases, protected by copyright law, are annotated by a group of researchers. Additionally, supplementary texts written by different authors are included in these editions. Moreover, the compilation of the data generally also has to be considered as creative work being capable of protected by copyright law. We arrive at a situation in which different copyrights (plus privacy issues, if letters are included also) come into play at the same time. Since these editions will be primarily distributed by digital archives and digital libraries, these institutions have to ensure that the dissemination of digital critical editions does not infringe any copyright.

Conclusions

The data collections described previously seem to be very specialized and hard to categorize from a legal point of view due to their heterogeneity with regard to authorship, source, and media type. In common practice, researchers usually think that the corpus layer that contains the primary data is the only one capable of being protected by law. This layer often originates from multiple holders of personal rights such as authors or subjects. However, the same protection may also apply to the transcription of spoken language as well as to the annotation layers that, again, are created by multiple persons. While digital versions of critical editions also consist of a set of primary data and several layers of analysis information, the legal situation is, to put it simply, identical to linguistic resources, but it becomes even more evident, because the analysis information added to the primary source are proper texts that have a genuine author and proper bibliographical information.

This complexity and heterogeneity of an intricate data system such as a large linguistic resource that, to the end-user, appears as an atomic piece of data is a phenomenon that is not too uncommon, especially with regard to the World Wide Web. If we take a look at Web 2.0 platforms and applications that embed multimedia content from various sources and creators (such as blogs that embed videos from YouTube, or sites that offer new functionality based on services such as Google Maps that can be accessed by third party resources using an API (application programming interface), it becomes obvious that linguistic data collections have a lot in common with hybrid Web applications that are nowadays commonly referred to as *mashups*. For the evaluation of the legal situation of this type of applications and data repositories it makes sense to take a closer look at the legal status of mashups within the framework of the World Wide Web.

Mashups—The Legal Point of View

Mashups have become an essential and defining part of the Web, version 2.0. Application programming interfaces and functions to embed external data allow millions of users to merge digital content such as texts (usually via RSS feeds), photos, audio, and video that is physically stored in multiple different Web-based repositories, each providing users with their own individual services but also sharing their data for automatic access. A multitude of free and easy to use tools such as, for example, Yahoo Pipes, and Dapper, that require a minimum of technical knowledge enable users to create mashups. (See http://pipes.yahoo.com and http://www.dapper.net.)

The related concept *user-generated content* has to be distinguished from the term *mashup*, which describes any type of publicly available Web content produced by users. Unlike user-generated content, to be characterized as the simple contribution of content by Web users, mashups are based on

the principle of modular design, turning users into Web-DJs who produce a sort of *user-remixed content.*

Currently the most frequent types of mashups are video and website mashups. Video sharing services such as YouTube, Yahoo Video, and Google Video are renowned for a large number of parodies that are meticulously composed of video remixes and dubbings. For instance, there is a trailer for a—nonexistent—sequel of the movie *Titanic* that was assembled using clips from various existing movies (see mrderekjohnson, 2006). Another popular video mashup shows a montage of public appearances of U.S. President George W. Bush and the former prime minister of the United Kingdom, Tony Blair, that has been lip-synchronized to the song "Endless Love" by Mariah Carey (see locopolitico, 2007). Unlike video mashups which, in the strict sense, already existed in the early 90s in the form of multimedia CD-ROMs, website mashups are a novel phenomenon. The free availability of Google Maps, for instance, has led to a large number of mashup services that combine their own data with Google's cartographical material. Two examples are the "Chicago crime map" (see http://www.chicagocrime.org) that provides a visual overview of all reported crimes in the Chicago area and the "Rentometer" (http://www.rentometer.com) that visualizes the average cost for renting a flat or a house across the United States. But website mashups have also found their way into scientific Web applications. The search engine "Ispecies" (http://www.ispecies.org) allows users to retrieve information about animal species and to extract photos as well as map data from a number of different resources.

The extensive use of current technologies in Web 2.0 scenarios quickly leads to a number of questions concerning the legal restrictions of using, remixing, and distributing online content. Legal conflicts with regard to mashups are often varied and cover multiple areas of law. Following we give a brief introduction into the expected fields of legal conflicts and we describe the consequences for mashup providers and end-users.

Contract Law

An important area to be considered when looking at mashups from a legal point of view is contract law. It is of no importance if the third-party data can only be accessed for a certain fee, a provider who intends to create mashups (*mashup provider*) first needs to sign a license agreement with the provider of the data and API he intends to use (*data provider*). This license agreement specifies that the mashup provider is strictly bound to the data provider's specifications concerning duration and modality of using their data. To give an example, eBay's API license agreements require that "the eBay Content is segregated from non-eBay content, and the eBay Content must be presented in such a way that the eBay Content is visually separate (as with lines or color changes) from non-eBay Content" (see *API license,* 2007).

In the case of Google Maps the data provider explicitly excludes his data from commercial use: "For individual users, Google Maps, including local search results, maps, and photographic imagery, is made available for your personal, non-commercial use only" (see *Google Maps,* n.d. §1). This raises the question under which circumstances a mashup that uses data from Google Maps has to be considered "commercial." Does, for instance, a mashup created by a private individual become a commercial application if a banner advertisement is displayed right next to an embedded Google map? In European law "commercial use," as mentioned in the Google license agreement, is defined by the existence of a financial transaction between end-users and mashup providers in connection with the use of the mashup. Hence, the inclusion of a banner advertisement next to a mashup would be judged as unobjectionable.

Copyright Law

Most legal problems concerning mashups occur in association with copyright law. It is part of the intellectual property rights, that provide the legal protection of nonmaterial goods, that is, any kind of intellectual property of a third party. This includes, among others, literary works as well as databases, software, and utility patents. In this context copyright law provides exclusive protection for authors (and other creators of intellectual property) who have the exclusive exploitation rights to their own work. In the European Union (EU) the copyright law of the member states is standardized by the "Directive on the harmonisation of certain aspects of copyright and related rights in the information society" (2001/29/EC).[2] Two essential articles of this directive that have a direct impact on the handling of mashups concern "Reproduction right" (Art. 2) and the "Right of communication to the public of works and right of making available to the public other subject-matter" (Art. 3). Both are affected if content is made available to be viewed or downloaded via the Internet.

Usually each act of use and publication of third-party content needs to be permitted by its copyright holder. This, however, is not necessary if the content is not capable of being protected[3] or if "exceptions and limitations" as defined in Art. 5 Dir. 2001/29/EC are affected, that provide exceptions to the rights provided in Art. 2 and 3 and allow the use of copyrighted data even without the author's permission. As an example, Art. 5 lists the "use for the sole purpose of illustration for teaching or scientific research" (Dir. 2001/29/EC Art. 5 (3) (a), see also section 4) or "quotations for purposes such as criticism or review" (Dir. 2001/29/EC Art. 5 (3) (d)).

In U.S. law, 17 U.S.C. Art. 108–122 limits the exclusive copyright (see Copyright Act of 1976, §§108–122). Furthermore, 17 U.S.C. Art. 107 incorporates the "Fair Use Doctrine" that allows a limited use of protected material without permission from the authors "for purposes such as criti-

cism, comment, news reporting, teaching [...], scholarship, or research" (see also Grimmelmann, 2007). The criteria to be considered are, among others, "the purpose and character of the use" (commercial, or nonprofit educational purposes), "the nature of the copyrighted work" as well as "the amount and substantiality of the portion used in relation to the copyrighted work as a whole and the effect of the use upon the potential market for or value of the copyrighted work." Currently it is being discussed whether the Fair Use Doctrine should be adopted for European law (Nimmer, 2006). Though the Fair Use Doctrine would provide more flexibility to European law, the legal certainty on the part of mashup providers would not be improved, for, in case of a legal dispute, the respective court could only decide *a posteriori* if fair use exists.

A frequently discussed issue concerns the question at which degree of revision a work that is based on another copyrighted work becomes a subject of protection of its own. An independent work created by this "free use" of another person's work may be published and exploited without the consent of the used work's author. In German copyright law, free use is regulated by Art. 24 UrhG.[4] However, it is required that the copyrighted work used is no longer identifiable as an essential part of the new work. In legal practice this question can only be decided in single individual cases, but due to the fact that in mashups the works used usually appear as clearly visible parts, mashup providers can hardly claim to have created a completely new work in terms of free use.

An important aspect of copyright law that is highly relevant when looking at mashups is the protection of databases that in the European Union is provided by Dir. 96/9/EC. Article 1.2 of this directive defines a database as a "collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means." This directive makes two significant stipulations. First, it offers copyright protection to databases that, based on the selection or arrangement of their contents, constitute the author's own intellectual creation. Thereby the author owns the exclusive right to carry out or authorize its reproduction, modification, and distribution. Second, the directive creates an exclusive right protection *sui generis* for makers of databases, independent of the degree of innovation. This protection of any investment allows the makers of databases to prevent unauthorized extraction and/or reutilization.

In consideration of the fact that data transferred via an API usually originates in databases, this directive has a big impact on the legal situation of mashups. The use of data from third-party databases either requires the completion of a license agreement (see section 3.1) or it must occur without prejudice to the rightholder's *legitimate interest.* Such prejudice would occur if the entire pool of information included in the source

database were provided by a mashup so that users would not have to visit the data provider's website anymore. However, as long as mashups do not affect the data provider's financial success in a negative way, an infringement of database law is not to be expected.

Databases are protected by U.S. copyright law as *compilations*. In 17 U.S.C. Art. 101, compilation is defined as "a collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship" (see Copyright Act of 1976, §101). Though since the late 90s there has been a discussion about incorporating an exclusive protection of databases into law, in 2004 the Database and Collections of Information Misappropriation Act (H.R. 3261) failed to respond to the fundamental concerns of its many diverse opponents (see Database Misappropriation Act of 2003): "Proponents of the bill, largely database producers and publishers, have not been able to identify a gap in existing law that needs to be filled or to demonstrate that their businesses have suffered because of any lack in the law" (see ALA, 2006).

Further Aspects of Copyright Law

Another area of intellectual property rights to be considered is *trademark law*.[5] Trademarks play an important role in identifying the sources of products or services with respect to business organizations. Due to the fact that mashups by their very nature embed data from third-party sources, in case of doubt the third-party trademarks have to be identified to avoid users mistaking the data provider's original service for the mashup provider's service. It is up to the owner of a trademark to specify how to refer to it. Google, for instance, defines a number of "Rules for Proper Use" for people who use Google trademarks as well as "logos, web pages, screen shots or other distinctive features ('Google Brand Features')" (see Guidelines, n.d.). Among others, these rules give precise instructions on how to write and distinguish the trademark from the surrounding text, and how to place Google logos on a website.

Under certain circumstances mashups can be confronted with issues of patent law. While in European law computer programs are excluded from patent protection, in U.S. law the patenting of software is admissible (see European Patient Office, 2007, Art. 52(2)(c)). Thus, mashup providers in the United States must pay attention not to use-protected methods for ranking and accessing content without holding a license.

Another relevant and widely-discussed issue is the liability of webmasters concerning illegal content such as defaming speech or pictures. In case of self-created data, only the webmaster is liable for what he or she has put online. If unlawful third-party content has been embedded into a website by means of a mashup, the situation becomes very complicated. In European law the liability for this case is regulated by Art. 14 and

15 Dir. 2000/31/EC (Directive 2001/29, EC, 2001). According to this directive, webmasters are not liable to check and to control third-party content embedded into their website. However, in a number of member states this regulation is interpreted to the effect that webmasters are not liable for any third-party content until there is an infringement of law. In that case they have to prevent similar future infringements proactively by filtering it from the third-party content. Thus, the dilemma of mashup providers is as follows. On the one hand, after recognizing an infringement of law mashup providers become liable to check and to filter the third-party data in matters of potential further infringements. On the other hand, the licencs agreements of most data providers do not allow for any manipulation of the transferred data. Against this background the interpretation of the above-mentioned article needs to be reconsidered in the face of Web 2.0 technologies.

Conclusion: Legal Evaluation of Mashups

The legal situation that arises from creating and using mashups in private as well as in commercial environments appears to be very complex (see section 3). However, with regard to mashups in the context of digital humanities (see section 2), especially in linguistics, two fundamental differences exist that are a result of their purpose and application in scientific research.

Unlike typical mashups that usually are in a permanent state of flux, mashups that contain linguistic data are rather static. In regular mashups users can add new content to arbitrary data layers—probably even additional data layers—whenever they want, while the only purpose of existence of linguistic mashups is to provide the basis for empirical research, that is, information can only be added by the persons who work in the project that creates or maintains the respective corpus. Enabling arbitrary users to modify or to extend a linguistic corpus would lead to unstable data collections and, therefore, would have to be considered bad scientific practice, because modifications would inevitably lead to a situation in which the conclusions drawn from a previous version of the corpus would no longer hold. If they were modified unsystematically, research results based on the original data could neither be proven again nor could they be compared with more recent results that are also based on the same resource. This inherent property of linguistic data enables as well as forces researchers to implement access strategies that accommodate both the current law and the specific nature of the data.

This strategy makes a second distinctive property of linguistic data collections relevant. Due to the fact that linguistic resources are collected, processed, and published for scientific purposes exclusively, their legal situation differs from regular mashups in a number of points, because several national as well as supranational laws and regulations provide specific

rules for the use and distribution of scientific data (see also section 3). Below, we provide a brief overview of the most important regulations and their impact on *researcher-generated mashups*.

In U.S. law, limitations on the exclusive copyright that allow the reproduction or distribution of single copies or phonorecords of copyrighted work by libraries and archives can be found in 17 U.S.C. Art. 108 (see Copyright Act of 1976, §108). However, unlimited online publications of scientific mashups to be used exclusively for research purposes is not covered by this article. Paragraphs (a)–(d) limit the number of legal copies of a protected work to at most three copies. Indeed, restrictions such as these could be implemented by means of a digital copyright or digital rights management system (DRM), but this cannot be considered a practical solution in a research environment, because if one or more researchers, probably spread around the entire world, need to process or to analyze a data set, it has to be fully accessible to these persons for their work. If they could access only one copy of a data collection at any given time, the immense amount of effort and expense necessary to plan and to create such an archive could not be justified.

From all paragraphs in Art. 108, (e) is the one that comes closest to the requirements of digital text repositories such as linguistic corpora. It allows the reproduction and distribution of an entire copyrighted work (or substantial parts) from a library or archive to be used for scientific purposes "if the library or archive has first determined, on the basis of a reasonable investigation, that a copy . . . of the copyrighted work cannot be obtained at a fair price." This paragraph does not enable the unlimited reproduction and distribution of copyrighted primary data from linguistic archives. Resulting from the indefinite terms "fair price" as well as "reasonable investigation," neither of which is defined clearly in this regulation, Art. 108 carries with it a high liability risk for the providers of data archives. Actually, with regard to linguistic archives, it seems to be impossible to specify a fixed price for research material as the variation in the availability of linguistic corpora is too big—some corpora are available for free, for others proof of purchase of the primary data has to be submitted (see section 2.2), yet other resources are available as commercial products and bear a very high price.

The only applicable regulation for linguistic data to be distributed and reproduced for scientific purposes in U.S. law is the Fair Use Doctrine (see Copyright Act of 1976, §107; see also section 3.2). A major problem is that the requirements for a specific usage situation to be called "fair use" are not very well defined by this article, therefore, it is affected by legal uncertainty, too. As a consequence, there have been discussions in several areas about the application of fair use (see also Stanford University Libraries, 2005–2008 and http://www.librarylaw.com). In practice, this uncertainty has led to a situation where a number of scientific institutions

have created their own fair-use guidelines for the use of their own copyrighted work. In the case of linguistic archives that contain multiple layers of copyrighted data this approach may provide a solution for the use of data structures and annotated data created by the researchers associated with an archive, but it would not solve the problems that arise when dealing with copyrighted primary data such as newspaper texts or other literary compositions.

In European law there is no generalized restriction to copyright for special purposes that would be comparable to the Fair Use Doctrine. Instead, Art. 5 Dir. 2001/29/EC contains a number of possible exceptions and limitations to the exclusive rights of copyright holders concerning the reproduction and distribution of their work (section 3.2). As the prescriptions in this article have an optional status only, member states are free to enact them as they see fit, more restrictively or less restrictively, when they transpose Art. 5 into national law. To give an example, the above mentioned Paragraph 3 (a) provides member states with the opportunity to allow the reproduction and distribution of copyrighted work for noncommercial research purposes. Unfortunately, the immense leeway this regulation gives to the member states has not been exploited yet. Instead, in German copyright law, Art. 52 (a) UrhG only allows a *limited subgroup of persons* copying or distributing *parts* of a copyrighted work for noncommercial research or educational purposes. Moreover, Art. 52, Paragraph 4 states an obligation to pay remuneration to the copyright holders. If we compare the situation of linguistic data in European and in U.S. law, the latter provides more options for the distribution of copyrighted data for research purposes despite a higher amount of legal uncertainty.

With regard to the use of data for noncommercial scientific purposes there are regulations and directives in different national as well as supranational laws. However, these regulations do not have anything to do with the respective data collections being mashups (as described in section 2). In practice, every single layer of these data collections has to be checked and cleared individually with respect to these regulations and appropriate decisions as to their level of accessibility have to be made.

Mashups in the Digital Humanities and in Computational Linguistics

In our project we apply several processing techniques to approximately sixty-five highly heterogeneous linguistic resources in order to archive them in a sustainable way and to make them available for search and query purposes in a web-based platform. First, we use several custom-made tools to split these digital text collections from their monolithic native state into multiple linguistic annotation layers that each contain the primary data. There are technical, conceptual, as well as legal reasons for splitting up the individual annotation layers: as *multiple* copyrights and licenses can potentially apply to the mashup as a whole (see section 2.2), we divide

these layers into physically separated data files (Witt et al., 2007). Before this processing we need to inform ourselves about the legal situation concerning every single one of the above-mentioned corpora. For this reason we collect metadata information with regard to the legal situation of the primary data and the annotation layers. We ask the project staff that is or was responsible for building a corpus to fill out a Web-based questionnaire that encapsulates all the necessary legal questions for these further processing stages (Lehmberg, Chiarcos, Hinrichs et al., 2007). One of the main components of the Web platform is a comprehensive database of metadata records that we create from the answers given to the questionnaire so that we can specify access restrictions with regard to the individual layers of primary data and linguistic annotations (Rehm et al., 2008). These metadata records enable us to specify in a very detailed way that, for example, a certain user is authorized to inspect all data layers of the TüBa-D/Z corpus (see section 2.2) while another user who does not have a valid license for the CD-ROM on which the corpus is based, can inspect the annotations only without having access to the primary data, that is, the newspaper articles, themselves. This user could apply the Corpus Masker tool in order to download the TüBa-D/Z corpus with the original text masked in a random fashion so that he or she can inspect the linguistic annotations in context and, for example, to apply them for linguistic experiments (Rehm et al., 2007a, 2007b).

New Challenges

In the final chapter of his well-received book about the differences between traditional scientific methods of structuring our world by means of taxonomies and formal classification schemes and the ubiquituous construction of meaning with the help of tagging articles, music, photos, and videos in the Web 2.0, David Weinberger gives an example that illustrates the enormous potential of publicly sharing metadata. As soon as digital books have gained widespread popularity, a multitude of novel functions will be available. According to Weinberger, "[e]very time a student highlights or annotates a page, that information will be used—with permission—to enhance the public metadata about the book" (p. 222). He continues listing additional features such as, for example, sharing which pages are read, reread, and which pages are skipped, and highlighting the passages marked by certain user groups, such as poets, students, professors, or priests. Using location-based services, digital books would know where they are read, so that playlists for specific environments, towns, or regions could be compiled. Weinberger thinks that what has happened in the Web 2.0 with music, photos, and video and—a point overlooked by the author—that has already been started years ago with value-adding services such as customer reviews and ratings offered by nearly all online book shops will also happen to books themselves. Reading, it is con-

cluded, will become a social activity, the enormous amounts of metadata will enrich the way in which people try to make sense of what they read and learn.

This example is compelling and exciting, but it is also a few years ahead of us. While at the end of 2007 Amazon.com launched their new e-book reader dubbed Kindle to quite some success, there are still multiple challenges in terms of both software and hardware to realize the scenario described by Weinberger. Nevertheless, one of the most crucial prerequisites is that of how to handle digital rights, copyright, and data access properly—in addition to contract law and copyright law these aspects also touch upon privacy of use. The phrase "with permission," inserted in the sentence quoted in the previous paragraph using dashes, almost seems to be an afterthought on Weinberger's part, as there are a plethora of legal questions associated with his scenario. How can students make sure that their teachers are unable to access which passages of their textbooks they have read and how often? Can all or only some references and quotes be realized as hyperlinks to their respective sources? What about the multi-source material commonly found in critical editions or, to a lesser extent, similar genres such as, for example, lecture notes? For this scenario to become a reality, an open protocol that enables users, mashup providers, and data providers to regulate access and to initiate as well as to control license agreements transparently and on a fine-grained level is needed.

Notes

1. The *die tageszeitung* CD ROM costs about 50 Euros. Licences for other (newspaper) corpora are often, if available at all, much more expensive.
2. See http://eur-lex.europa.eu/LexUriServ/LexURIServ.do?uri=CELEX:32001L0029:EN:HTML. EU directives are legislative acts of the European Union that require member states to transpose them into national law without dictating the means of achieving this. Due to the fact that the deadline of transposal for directive 2001/29/EC has expired, it is assumed that it already has been transposed.
3. In German law, for instance, a certain level of creativity and individuality of a work is required for copyright protection. To give an example, specific headlines or standardised business letters do not qualify with regard to this distinction.
4. Urheberheberrechtsgesetz, the German Copyright Act.
5. The most important international regulations concerning trademark law are the Paris Convention for the Protection of Industrial Property (http://www.wipo.int/treaties/en/ip/paris/trtdocs_wo020.html) and the Trade-Related Aspects of Intellectual Property Rights (TRIPS, http://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm).

References

American Library Association (ALA). (2006, March 28). HR 3261: The Database and Collections of Information Misappropriation Act. *Database protection legislation.* Retrieved September 3, 2008, from www.ala.org/ala/washoff/woissues/copyrightb/dbprotection/databaseprotection.cfm#status

API license agreement. (2007, January 10). Retrieved September 3, 2008, from http://developer.ebay.com/join/licenses/individual

Burnard, L., & Bauman, S. (Eds.). (2007). *TEI P5: Guidelines for electronic text encoding and interchange.* Text Encoding Initiative Consortium. Retrieved July 11, 2008, from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Copyright Act of 1976, 17 U.S.C. 101 *et seq.* (n.d.) Retrieved September 3, 2008, from http://www4.law.cornell.edu/uscode/html/uscode17/usc_sec_17_00000101—000-.html

Database and Collections of Information Misappropriation Act of 2003, HR. 3261, 108th Cong. (2003). Retrieved September 3, 2008, from http://www.copyright.gov/docs/regstat092303.html

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., & Witt, A. (2006). Sustainability of linguistic resources. In Erhard Hinrichs, Nancy Ide, Martha Palmer & James Pustejovsky (Eds.), *Proceedings of the LREC 2006 satellite workshop merging and layering linguistic information* (pp. 48–54). Genoa, Italy. Paris: European Language Resources Association (ELRA).

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. (2001, June 22). Retrieved September 3, 2008, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:HTML

European Patent Office. (2007). Revision of the European Patent Convention (EPC 2000) [Special Issue]. *Official Journal of the European Patent Office.* Retrieved September 3, 2008, from http://www.european-patent-office.org/epo/pubs/oj007/01_07/special_edition_1_epc_2000.pdf

Google maps terms and conditions. (n.d.) Retrieved September 3, 2008, from http://maps.google.com/intl/en/help/terms_maps.html

Grimmelmann, J. (2007). The structure of search engine law. *Iowa Law Review, 93*(1), 1–63. Retrieved July 11, 2008, from http://www.nyu.edu/projects/nissenbaum/papers/Grimmelmann_StructureOfSearchEngineLaw.pdf

Guidelines for third party use of Google brand features. (n.d.) *Google permissions.* Retrieved 2008, September 3, from http://www.google.com/permissions/guidelines.html

Lehmberg, T., Chiarcos, C., Hinrichs, E., Rehm, G., & Witt, A. (2007). Collecting legally relevant metadata by means of a decision treebased questionnaire system. In Sara Schmidt, Ray Siemens, Amit Kumar & John Unsworth (Eds.), *Digital Humanities 2007* (pp. 164–166). ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Lehmberg, T., Chiarcos, C., Rehm, G., & Witt, A. (2007). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Georg Rehm, Andreas Witt & Lothar Lemnitzer (Eds.), *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen—Data structures forlLinguistic resources and applications: Proceedings of the Biennial GLDV Conference 2007* (pp. 93–102). Tübingen: Gunter Narr.

locopolitico. (2007, February 9). Bush and Blair's endless love [Video file]. Video posted to http://www.youtube.com/watch?v=w8rr6fz1hQQ

mrderekjohnson. (2006, April 5). Titanic: The sequel [Video file]. Video posted to http://www.youtube.com/watch?v=vD4OnHCRd_4

Newman, P. (2007). Copyright essentials for linguists. *Language Documentation & Conservation, 1*(1), (pp. 28–43). Retrieved July 11, 2008, from http://scholarspace.manoa.hawaii.edu/html/10125/1724/newman.html

Nimmer, R. (2006). Google print project—Unfair use of copyright. *Computer und Recht* CRI, 1–6.

Rehm, G., Eckart, R., & Chiarcos, C. (2007). An OWL and XQueryBased mechanism for the retrieval of linguistic patterns from XMLCorpora. In Galia Angelova, Kalina Bontcheva, Mitkov Kalina, Ruslan Mitkov, Nicolas Nicolov & Nicolai Nikolov (Eds.), *International conference recent advances in natural language processing (RANLP 2007)* (pp. 510–514). Borovets, Bulgaria. Shoumen: Incoma.

Rehm, G., Witt, A., Lehmberg, T., Dellert, J.; Eishold, F.; Evang, K.; Leshtanska, M., & Stark, M. (2008). The metadata database of a next generation sustainability web platform for language resources. *Proceedings of LREC 2008 (Language Resources and Evaluation Conference).* Marrakech, Morocco: European Language Resources Association (ELRA).

Rehm, G., Witt, A., Zinsmeister, H., & Dellert, J. (2007a). Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In Sara Schmidt, Ray Siemens, Amit Kumar & John Unsworth (Eds.), *Digital Humanities 2007* (pp. 166–170). ACH, ALLC, Urbana Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana Champaign. Urbana-Champain, IL: Graduate School of Library and Information Science, University of Illinois.

Rehm, G., Witt, A., Zinsmeister, H., & Dellert, J. (2007b). Masking treebanks for the free distribution of linguistic resources and other applications. *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen, Norway, number 1 in Northern European Association for Language Technology Proceedings Series, (pp. 127–138).

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., & Hinrichs, E. (2006) Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards—The State of the Art*. East Lansing, Michigan. Retrieved July 11, 2008, from http://linguistlist.org/emeld/workshop/2006/papers/schmidt.html

Schmidt, T., & Wörner, K. (2005). Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung, 6*, 171–195. Retrieved July 11, 2008, from http://www.gespraechsforschung-ozs.de/heft2005/px-woerner.pdf

Schmidt, T., & Wörner, K. (in press). EXMARaLDA—Creating, analysing and sharing spoken language corpora for pragmatic research. *Proceedings of the 10th International Pragmatics Conference (Göteborg, 8–13 July 2007)*.

Stanford University Libraries. (2005-2008). *Copyright & fair use*. Retrieved September 3, 2008, from http://fairuse.stanford.edu/Copyright_and_Fair_Use_Overview

Telljohann, H., Hinrichs, E., & Kübler, S. (2004). The TüBaD/Z treebank—Annotating German with a contextfree backbone. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal. Paris: European Language Resources Association (ELRA).

Telljohann, H., Hinrichs, E., Kübler, S., and Zinsmeister, H. (2006). Stylebook for the Tübingen treebank of written German (TüBaD/Z). Technical Report, Seminar für Sprachwissenschaft, Universität Tübingen. Retrieved July 11, 2008, from http://www.sfs.uni-tuebingen.de/resources/sty.pdf

Trilsbeek, P., & Wittenburg, P. (2006). Archiving challenges. In Jost Gippert, Nikolaus P. Himmelman & Ulrike Mosel (Eds.), *Essentials of Language Documentation* (pp. 311–335). New York: Mouton de Gruyter.

Weinberger, D. (2007). *Everything is miscellaneous—The power of the new digital disorder*. New York: Times Books.

Witt, A., Schonefeld, O., Rehm, G., Khoo, J., & Evang, K. (2007). On the lossless transformation of singlefile, multilayer annotations into multirooted trees. In B. Tommie Usdin (Ed.), *Proceedings of Extreme Markup Languages 2007*. Montréal, Canada. Retrieved July 11, 2008, from http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2007/Witt01/EML2007Witt01.pdf

Wörner, K., Witt, A., Rehm, G., & Dipper, S. (2006). Modelling linguistic data structures. In B. Tommie Usdin (Ed.), *Proceedings of Extreme Markup Languages 2006*. Montréal, Canada. Retrieved July 11, 2008, from http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2006/Witt01/EML2006Witt01.pdf

Zimmermann, F., & Lehmberg, T. (2007). Language corpora—Copyright—Data protection: The legal point of view. In Sara Schmidt, Ray Siemens, Amit Kumar, and John Unsworth (Eds.), *Digital Humanities 2007* (pp. 162–164). ACH, ALLC, Urbana-Champaign, IL, USA: Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

After graduating in 2005 from the University of Hannover, Timm Lehmberg became a research associate in the projects Evidential Markers in German (University of Hannover) and Sustainability of Linguistic Resources (at SFB 538 "Multilingualism"). At the beginning of 2008 he started working in the project Language Variation in Northern Germany at the University of Hamburg. His main interests are in data modelling and corpus-based research in association with language change and grammaticalization theory as well as legal issues in corpus linguistics.

Georg Rehm works in Tübingen University's collaborative research center Linguistic Data Structures in a project that is developing a Web-based sustainability platform for

linguistic resources. He holds a doctorate in applied and computational linguistics and a masters in computational linguistics and artificial intelligence. In addition to specific legal aspects of linguistic data his main research interests are text linguistics, novel applications of XML-based markup languages in computational linguistics and natural language processing on the Web, especially text structure parsing of Web documents and the automatic identification of Web genres.

After graduating in 1996, Andreas Witt started working as a researcher and instructor in computational linguistics and text technology at Bielefeld University and received a doctorate in computational linguistics and text technology in 2002. In 2006 he moved to the University of Tübingen, where he works on the sustainability of linguistic resources. Witt's main research interests deal with questions on the use and limitations of markup languages for the linguistic description of language data. He is a member of several research organizations, among them the TEI Special Interest Group on overlapping markup, for which he wrote parts of the latest version of the chapter "Multiple Hierarchies," included in version P5 of the TEI-Guidelines.

Felix Zimmermann graduated in 2006 from the law faculty of the University of Hannover and became a research associate at its Institute of Legal Informatics. Since 2007 he has worked at the Institute of IT-Security and Security Law at the University of Passau. Felix Zimmermann's main research interests are IT-specific areas of law such as liability on the Internet, telecommunications law, and domain name law. His dissertation is about the legal impact of IT-procurement procedures.