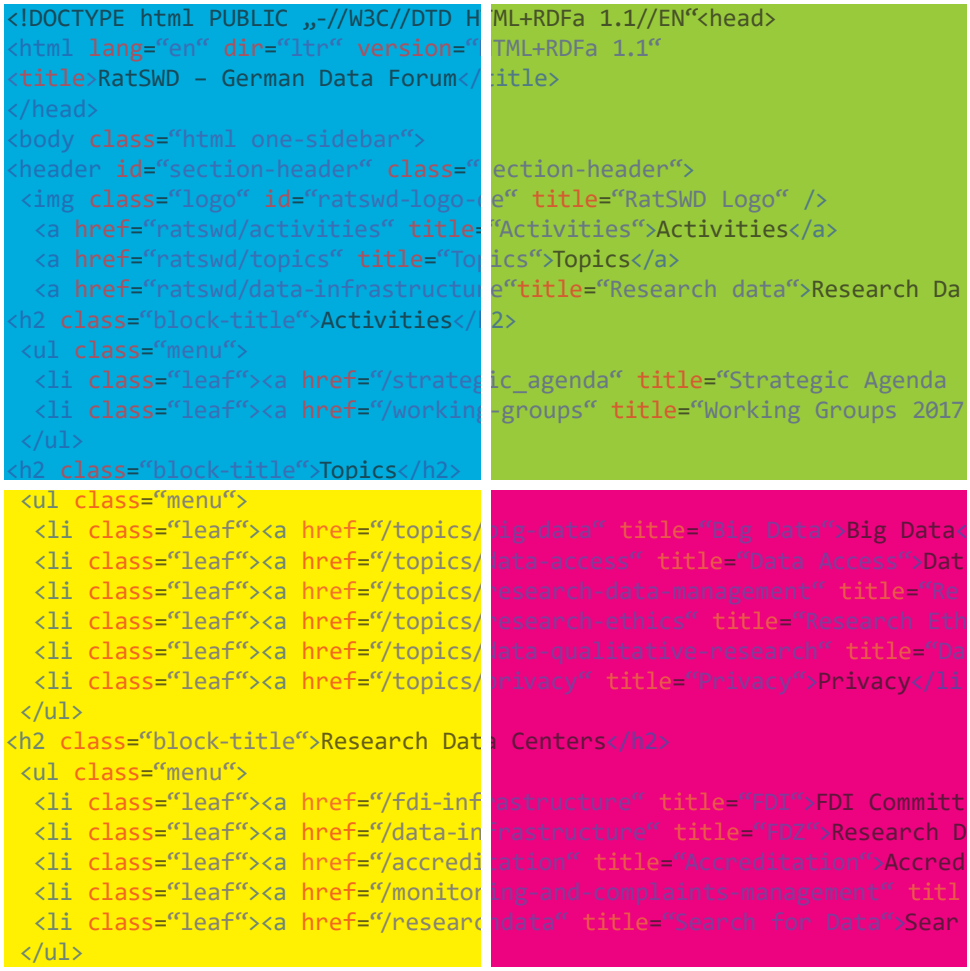
4 | Output
6th Appointment Period

RatSWD.
German Data Forum

Big data in social, behavioural, and economic sciences

Data access and research data management

Including an expert opinion on
“Web scraping in independent academic research”



SPONSORED BY THE

Federal Ministry
of Education
and Research

German Data Forum (RatSWD)

Big data in social, behavioural, and economic sciences

Data access and research data management

Including an expert opinion on
“Web scraping in independent academic research”

Contents

Executive Summary

The amount of data stored worldwide is growing exponentially. According to the study "Data Age 2025 – The Digitization of the World" (Reinsel/Gantz/Rydning 2018), the global data volume is expected to grow from 33 zettabytes (ZB) in 2018 to 175 ZB in 2025. In the social, behavioural and economic sciences, there is a growing interest in utilising data generated through increased digitisation, often referred to as big data. This interest is based on the fact that these data are huge in numbers, offer options for real-time analysis, allow non-reactive collection methods, and the possibility to observe interactions between individuals.

With this in mind, the German Data Forum (RatSWD) recognises the need to systematically tap into the potential of these data for academic research. The restrictions that currently exist – de jure and also de facto – on research involving big data sources in Germany must be addressed. When researchers are granted access to big data sources owned by corporations, they are often denied access to variables or observations that are critical for corporate strategies. Often, they cannot disclose the data to other researchers and are confronted with the risk that their access to the data could be unilaterally terminated before their research project is completed. This can result in conflicts of interest or restrictions regarding the publication of research findings. In addition, processing big data is considerably more demanding than processing survey data or administrative data. Therefore, researchers using big data should thoroughly examine the processes used to generate data and the algorithms employed to collect them.

- Users should should familiarise themselves with the legal requirements that apply to data gained from big data sources. If, for example, an Application Programming Interface (API) is used to collect the data, researchers must first read and understand the API's terms of service before using the data.
- Many researchers use web scraping to extract data from the internet. The German Data Forum (RatSWD) is aware that the legal implications are often unclear when this method is used and therefore obtained a legal opinion. This legal opinion addresses the legal issues surrounding data access, data usage, and data archiving in connection with web scraping. The German Data Forum (RatSWD) aims to ensure legal security for researchers when working with big data sources.
- To strengthen the German research infrastructure in the social, behavioural, and economic sciences, the German Data Forum (RatSWD) suggests to provide standardised access to anonymised (micro)data[1] from public or private big data sources. This can be accomplished by establishing independent research institutions or by commissioning existing institutions which operate as fiduciary agencies. The infrastructure required for this could be a key topic for the National Research Data Infrastructure (NFDI) currently in development. Furthermore, in view of potential uses, open data should generally be provided in machine-readable form.

1 Data including detailed information, e.g., on individuals or companies.

1 Introduction

Due to digitisation, public institutions and especially private enterprises are generating growing volumes of social, behavioural, and economic data. Researchers are increasingly interested in using these data. Inter alia, this interest is based on the

1) large number of cases: as opposed to traditional data compilations, digitally recorded data allow for a more detailed and comprehensive representation of social and economic relationships
2) time factor: behaviour can be observed in real-time and over longer periods of time
3) non-reactive collection methods[2]: the collection process as such usually does not affect the observed behaviour
4) network character: Not only individual behaviour can be observed, but also the interactions between people

However, the availability of these data for research in the social, behavioural, and economic sciences and other user groups, such as official statistics agencies, is limited, due to problems in data access, utilisation and archiving

The Big Data Working group of the German Data Forum (RatSWD) examined whether and how researchers can have secure access to data and replicability of research findings can be ensured. The working group also discussed which data privacy and ethical issues need to be considered for big data research. This report provides an overview of the experiences made when using big data for research in specific fields and what the conditions for accessing the data access were. This report details several specific barriers faced by researchers, especially with respect to data access.

In line with Adjerid's and Kelley's (2018) definition, the working group understood big data as large amounts of data generated for purposes other than research that are available at short notice, often in an unstructured form. Mass administrative data such as social insurance data were intentionally not taken into consideration because the issues in this report, as they apply to these kinds of data, have already been subject to extensive academic and political discussions.

Presently, private enterprises collect and store large amounts of data that were not originally intended for research but can nevertheless be used for this purpose. This includes data generated in large volumes as by-products to production processes or collected in connection with electronic communications (e-mail, text messages, messenger services), in web applications (information and communication in social media, searches in search engines, purchase and sales behaviour on online platforms), via personal end devices (personal computers, smart meters, movement data, location data, data from apps, and communication information), or data on the purchasing behaviour of individuals collected via debit, credit, or discount cards.

2 Non-reactive methods collect data that result from everyday behaviour, i.e., without any connection to possible uses in research. Besides digital behavioural data, non-reactive data also include process-induced data.

In most cases, it is up to private enterprises to decide whether or not they make data publicly accessible. While companies sometimes offer researchers structured and dedicated access to data through contracts or binding agreements, the majority of publicly accessible data is intended for the commercial use by end customers (mainly via websites) or other third parties.

A special form of exploiting big data sources is web scraping. With this method, big data can be extracted from the internet and made available for research. However, it is not yet clear under which circumstances web scraping is legal. As this method has already been extensively used by researchers and therefore considered to be particularly relevant, the expert opinion on "Web scraping in independent academic research", which was requested by the working group, is provided as an appendix to this report. Web scraping includes both the use of Application Programming Interfaces (API) and the extraction of data from websites designed to be read by human end users. To differentiate between these two terms, the latter is referred to as "screen scraping" (cf. our definition under section 2.3).

Section 2 describes the different access paths to big data that are available to researchers, thereby differentiating between individual and institutional access to big data and the collection of data via unregulated access paths, specifically using screen scraping. Finally, this section looks at the risks and opportunities of using open data. Section 3 provides a summary of the findings in the legal opinion initiated by the working group with respect to data collection via web scraping. Section 4 presents data trusts as a possible solution to the problems with data accesss that have been pointed out. Section 5 contains a glossary of the technical terms used in this paper.

2 Access paths to big data for researchers

This chapter begins by presenting a brief overview of the three access paths to big data that are widely used in research. These access paths are classified by the degree to which the data producer (usually a private corporation) has consciously consented to the use of their data (individual data access, section 2.1), one-to-many agreements (institutionalised access, section 2.2), or lack of any agreements with the data retriever (screen scraping, section 2.3). Section 2.4 discusses open data and freedom of information as potential alternative sources for big data.

2.1 Individual access to non-public enterprise data

Many researchers enter into agreements with private enterprises to obtain individual access to data for specific research projects (see Edelman 2012 or Einav and Levin 2014 for an overview of relevant research projects).

Example 1: Cooperation between RWI – Leibniz Institute for Economic Research and ImmobilienScout24

Within the framework of this cooperation (an de Meulen/Micheli/Schaffner 2014), ImmobilienScout24.de provided all the information collected on their internet platform to RWI for academic research (see Boelmann and Schaffner 2019 for a description of the dataset). The data include all advertisements posted on the platform since 2007. ImmobilienScout24.de is currently the leading online platform for the real estate market in the German-speaking region. Besides being up-to-date and containing a wealth of information, the geographical information of these data is what makes them stand out. Since all of the properties advertised on ImmobilienScout24.de are georeferenced, observations can be made on wide-scale developments. Furthermore, the data can be merged with a wide range of information on the respective regional aggregation level.

Example 2: Cooperation between the Federal Statistical Office and mobile phone companies

The Federal Statistical Office (Destatis) cooperates with enterprises to obtain access to their digital data. Within the framework of a feasibility study to explore the potential of using mobile phone data for official statistics, Destatis started a cooperation project with two subsidiaries of Deutsche Telekom AG in September 2017, namely T-Systems International GmbH and Motionlogic GmbH (Hadam 2018). The goal of this project was to examine the extent to which mobile phone data can accurately map and estimate daytime populations and residents, commuter flows, and the distribution of tourists.

Data access as laid out in examples 1 and 2 is often only granted if the companies benefit from the specific research projects. For example, if companies can gain knowledge about the potential of their own data or the behaviour of their customers, they can develop new products or services as a result of the project findings or optimise their own organisation (for example, their internal data management). However, such project-based individual access can lead to various problems; researchers should carefully consider the risks that could arise.

Notes and recommendations:

- Data-generation is not often completely transparent. This means that user behaviour can be influenced by machine-generated recommendations that may change over time. Information must be obtained from the data generator before beginning the cooperation project.
- Researchers are often not provided with all the data that could be potentially relevant for research projects, and important variables and observations (for the company's strategy) are often withheld.
- It must generally be assumed that researchers who were granted access to data are not allowed to share these data with other researchers. This limits the ability to reproduce and replicate research results and thus, potentially, the possibility of publishing them. It also precludes the secondary use of the data.
- Since data access is granted under a contract, which can usually be terminated unilaterally, the successful completion of research projects is at risk.
- Finally, if the data are provided to the researchers within the framework of consultancy agreements, conflicts of interest can occur.

2.2 Institutionalised data access via open interfaces

In addition to companies that enter into agreements with researchers for individual data access, there is a growing number of companies – among them many internet platforms – that make parts of the digital data they generated available via publicly accessible web interfaces (Application Programming Interfaces or APIs) for use by third parties. With these APIs, companies voluntarily grant structured insights into – and access to – the data structures behind the respective services. These insights can range from being very limited to extensive. Although APIs are primarily intended for commercial use (e.g., in third-party mobile applications), they are promising resources for future research projects. Researchers are therefore increasingly using the options APIs provide to collect extensive data onhuman behaviour, economic indicators, and other information for their research.

To use these interfaces, users must consent to the company's terms of service. Consent is often given implicitly by using the interface or in a prior registration step. However, the provisions regularly preclude the re-utilisation of the obtained data for specific purposes and sometimes impose strict requirements for data management.[3]

3 For example, Twitter demands that any tweets that were deleted from the platform also be deleted from datasets that were collected and stored by third parties (https://developer.twitter.com/en/developer-terms/agreement-and-policy.html, last accessed: 23.08.2019).

Example 3:

One example for the use of these data is a service offered by Google Trends that illustrates the quantitative development of search terms users enter into Google's search engine. These data have often been used for research (for example, Carrière-Swallow and Labbé 2013; Hyunyoung and Varian 2012; McLaren and Shanbhogue 2013; Bug 2015: 86; Rieckmann and Schanze 2015).

Schmidt and Vosen (2011, 2012, 2013) used the information provided in the Insights for Search application by Google Trends to forecast private consumption. Based on random samples, this application provides weekly insights into searches starting with the year 2004, classifying them into 605 categories and subcategories. The authors used 45 categories that are relevant for private consumer spending and compatible with the private consumption figures according to the national accounts maintained by the Federal Statistical Office to estimate a weighted consumption indicator. One problem here was that when a query was sent to the Insights for Search application, the underlying software selected a random sample from all registered Google searches on a given day. Queries sent to Google Trends on the same day delivered the same results. However, the results on different days varied due to the different random samples selected by Google Trends from all searches. The researchers solved this problem by using average values from 52 random samples selected on different days.

Example 4:

The RWI consumption indicator resulting from this research project provides a good picture of current developments in consumption. Schmidt and Vosen (2012) demonstrated that economic forecasts can be improved with the help of Insights for Search applications by Google Trends by considering special policies such as scrapping bonuses that were introduced by many countries during the 2009 recession.

Besides the legal uncertainties, examples 3 and 4 indicate that it is also important to closely examine how the retrieved data were generated; this applies to all APIs. The problems that may arise with these data were demonstrated, for example, by the former Google Flu Trends service that Google had developed and published as a tool to forecast influenza activity. Initial evaluations indicated the accuracy of the forecasts was very high. Later, it was noted that because of problems related to the accuracy and, above all, transparency of the data, the measurement methods, and relatively simplistic forecast models, Google Flu Trends did not produce consistently reliable forecasts (Lazer et al. 2014). Such poor long-term reliability is rarely evident to researchers from the outset, making it difficult for them to assess the quality of the results (in particular, if they are based solely on a single dynamic data source) (Di Bella/Leporatti/Maggino 2018).[4]

4 The Robert Koch Institute, for example, uses a broader methodical approach: https://www.rki.de/DE/Content/Infekt/IfSG/Signale/Projekte/Forschungsantrag_Demis_Signale.pdf?__blob=publicationFile (last accessed: 07.05.2020).

The Twitter[5] APIs are now often used for empirical research projects (for an overview, see e.g. McCormick et al. 2017). In research fields related to computer science, such as computational social science, Twitter data are now even considered part of the "de facto core datasets" (Pfeffer/Mayer/Morstatter 2018). A 1 % sample can be retrieved free of charge via the Twitter API. A fee is charged for larger random samples of tweets. Based on the assumption that these samples are a random selection from the platform's overall content, many research projects naturally have relied on the 1% access. When making inferences, it should first be taken into account that Twitter users are inherently not represenative of the general population.[6]
Additionally, recent findings based on reverse engineering show that the free 1% sample is not selected completely randomly from the entire platform's content and can be manipulated (Morstatter et al. 2013, Pfeffer/Mayer/Morstatter 2018). With such limited data, various other faults can occur when operationalising measurement techniques and making inferences about target populations, which must be considered in the research design (Sen et al. 2019).

Besides Twitter, Wikipedia also provides APIs with granted public access to the data on which their services are based. These APIs can be used, for example, to record visitor numbers for individual pages or topics, or changes to the texts. These data are also increasingly used in research (see e.g., Moat et al. 2013; Slivko 2018).

Comparable APIs are offered by many internet platforms and service providers from YouTube to Instagram, or specialised platforms like Stackoverflow.[7]
In addition, there are many APIs that can be used to analyse mobility patterns. An example for this are APIs by companies that rent out vehicles.[8] A wide range of related literature exists for bike rentals in particular (see e.g. Fishman 2016, Yongping and Zhifu 2018).

Static data downloads that are made available by some companies offer access similar to that provided by APIs. They typically consist of collections of files from exported databases or database elements (so called database dumps). The use of these downloadable files – as with interfaces – is subject to the provider's terms of use and copyright terms. Examples include the Wikimedia Foundation – which offers comprehensive database dumps for free[9]– as well as the Internet Movie Database[10] and the statistics platform FiveThirtyEight[11].

5 https://developer.twitter.com/en/docs/api-reference-index.html (last accessed: 07.05.2020).
6 In general, young male users are overrepresented. This also applies in Germany, where only 4 % of the population uses Twitter at least once a week (Frees and Koch 2018). In this case, "use" often means passive reading without writing or interacting.
7 A discussion forum for software development.
8 See, for example, https://github.com/ubahnverleih/WoBike or https://github.com/CityOfLosAngeles/mobility-data-specification/blob/dev/agency/README.md#vehicle-events
9 https://dumps.wikimedia.org (last accessed: 07.05.2020).
10 https://www.imdb.com/interfaces (last accessed: 07.05.2020).
11 https://data.fivethirtyeight.com (last accessed: 07.05.2020).

Notes and recommendations:

- Researchers should thoroughly review the terms of service, terms of use, or general terms & conditions of the APIs and document them, if possible. Some companies have made exceptions for academic researchers in their terms of use that provide greater liberties for data retrieval and use.
- Company regulations usually address not only data access, but also put restrictions on data storage, sharing, and publishing. Potential conflicts with good academic practice[12], especially with replicability, must be taken into consideration on a case-by-case basis.[13]
- Compliance with the terms of use does not discharge researchers from their obligations under law and owed to research ethics. For example, researchers need to protect the right to data privacy of the users whose behavioural data are the basis of the datasets. These obligations may go beyond what the data provider's terms of use require. This could also include considerations under copyright law and, rarely, competition law (see also chapter 3.2).
- Researchers who access data via APIs (or file downloads) should thoroughly analyse the data generation processes and the algorithms used for data acquisition if they are documented or can be otherwise determined. Note that data-generation processes may change quickly (e.g., through implementation of a new platform software). The validity and reliability of metrics derived from these data must be critically reviewed.

12 See, for example, "Guidelines for Safeguarding Good Research Practice" by Deutsche Forschungsgemeinschaft (DFG): https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp (last accessed: 07.05.2020).

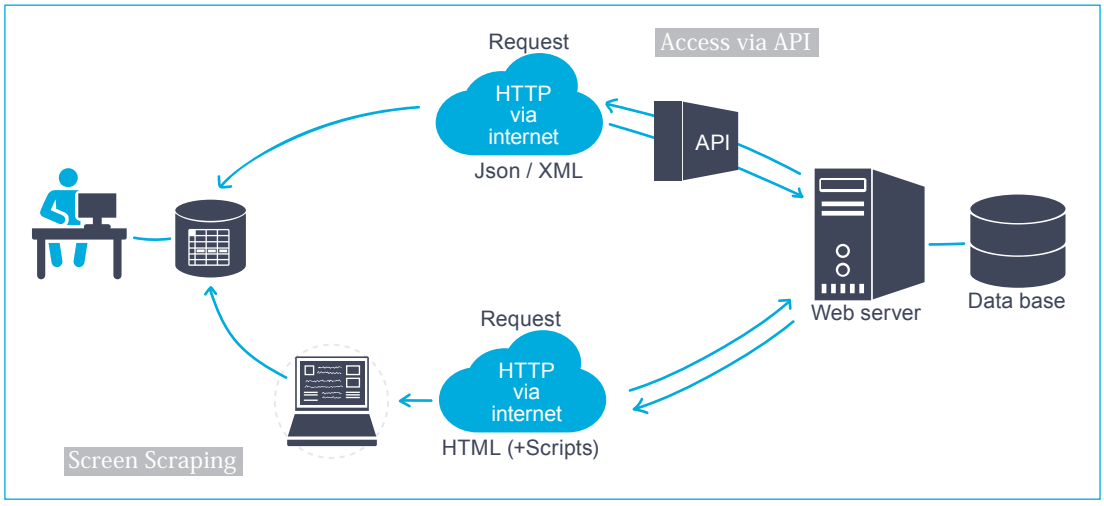
13 Conflicts between the initial data provider's terms of use and those of archiving platforms (resource permanence) or national laws could arise.

2.3 Collecting web data via screen scraping

Often, entities on the internet provide neither dedicated access points to retrieve or download their data nor any options to enter into bilateral agreements for access. Internet platforms provide data on websites for human users, but these data are not always fully accessible via special interfaces or downloads.

In these cases, researchers often use a method called screen scraping. While in practise there are two different definitions for web scraping and screen scraping – the two terms are sometimes used interchangeably– this paper defines screen scraping as the automated retrieval and extraction of information from the unstructured part of the internet intended for human users (von Schönfeld 2018: 20). Accordingly, a screen scraping program simulates human behaviour to gain access to websites, in order to retrieve and then analyse the information they contain (von Schönfeld 2018: 58).[14] In contrast, web scraping is understood to be an umbrella term for an algorithm-based method for retrieving publicly accessible information from the World Wide Web by technological means (ibid.). Besides screen scraping, this also includes other methods (especially programming interfaces, see section 2.2). The difference between screen scraping and programming interfaces is illustrated in figure 1.

Fig. 1: Two web scraping methods in comparison: accessing data via API and screen scraping



Screen scraping uses an application or a script to send queries to one or more predefined URLs[15] or servers. The attached metadata then emulate a typical web browser, for example Mozilla Firefox. A static document is returned that is usually written in HTML or CSS, or sometimes in human-readable file formats such as PDF files. Next, the information that is interesting for the requesters is extracted using the user's predefined heuristics.[16] In another step, linked images or other multimedia contentscan additionally be downloaded and saved.

14 Exempt from this are the usually higher access rates of automated procedures that naturally outpace those of human users.

15 The URLs used for the requests can be explicit URL lists or may contain only vague samples for the requests, for example all pages of a defined domain or all websites with specific search terms in their titles.

16 Example: the term "price" is used to search a HTML tree within an online shop site for HTML elements that contain the prices of the listed products, or the term "description" to search for descriptions. These elements are then transferred, as variables, into a structured format.

If content cannot be retrieved simply by requesting the URL, techniques are used that, after the website has loaded, simulate entries into fields, and mouse-clicks on buttons by human users to request more information. The subsequent dynamic loading of page elements allows the retrieval of content that can only be created through special sequences of user behaviour. By entering information into the appropriate fields (partly also by directly requesting the URL), screen scraping applications can log in as users – if the login data is available – and gain access to data that is classified as "non-public" by the queried website. By doing this (in addition to prior registration) the user usually accepts the terms of use – either implicitly or explicitly by clicking on the corresponding fields. By using interactive elements, clicking on hyperlinks (also referred to as crawling or spidering), or generating URL variations, the majority of the services offered on internet platforms can be explored and retrieved without knowing the correct URL or page lists from the start.

Many research projects in the social, behavioural, and economic sciences now use screen scraping methods. This is often necessary, e.g., for researching products and prices, because the relevant internet platforms only rarely offer public interfaces. Such research includes, among others, price developments on hotel or holiday apartment booking platforms (Gyódi 2017), rental trends on online classified ad platforms (Boeing and Waddell 2016), algorithmic pricing on shopping platforms (Chen/Mislove/Wilson 2016), or price index forecasts for consumer goods on the websites of large retail chains (Powell et al. 2017).

Web scraping – and particularly screen scraping – is popular in many other research fields as well. Topics that have been studied include:

- Ethnic segregation, studied by web scraping reviews on restaurant review sites (Davis et al. 2019)
- Variations in qualification requirements on job platforms (Verma et al. 2019)
- Inequality on freelance job websites (Hannack et al. 2017)
- Dynamics of political reporting and consumption on the internet (Ørmen 2019)
- New indicators, for example for innovation (Kinne and Axenbeck 2018)

Today, web scraping is of special importance in practical research. Its relevance for research is presented in chapter 3, using a specific case as an example, along with the central points of the legal opinion obtained by the German Data Forum (RatSWD), including the resulting recommendations for action.

2.4 Open data and the Freedom of Information Act

In addition to the access paths to big data that were discussed earlier, open data and the German Freedom of Information Act potentially allow access to political and administrative data. Generally, data of this kind are not compiled for research – as is the case with big data. The volumes of data need not necessarily be huge. However, special restrictions apply when using the data for academic research, similar to the data access path discussed above.

Open data

The term "open data" applies to machine-readable and structured data that allow free use, re-use ,and sharing by everyone. According to the definition of open data, the only conditions for using open data are to name the source, or note that the information is open source, in order to ensure transparency. Open data must not include any personal data or data that is subject to data protection regulations.

Different from processed data, which often are legally protected, open data usually include text and images, as well as tables, maps, or databases. Such data are also referred to as raw data and are used as the basis for processed information. They can originate from very different areas of society: geodata, cultural data, data from science and research, or weather and environmental data.

Definitions

Open data

Open data are all data assets made available in the public's interest without restrictions for free use, dissemination, and re-publishing. This includes, for example, teaching materials, geodata, statistics, transport information, academic publications, medical research findings, or radio or TV broadcasts. Open data are not exclusively understood to be data assets from public administrations, because private enterprises, universities, and radio stations, or non-profit organisations also generate them (von Lucke and Geiger 2010: 3). Licences serve to identify and regulate the free use of these data. (ibid.).

Linked open data

Linked open data are all data that are interlinked with other data via the World Wide Web and made available in the public's interest without restrictions for free use, dissemination, and re-publishing. (von Lucke and Geiger 2010: 4). Added value is created if databases that were not previously linked are merged, resulting in new findings. Especially the easy addressability of data assets on the internet lowers the bars for data retrieval. By means of Uniform Resource Identifiers (URI) and the Resource Description Framework (RDF), subsets of data, information, and knowledge can be prepared, shared, exported, or linked (ibid.: 3).

The German Freedom of Information Act (IFG)

Freedom of Information is a general and basic right to view public documents and files stored by public administrations. It can be exercised by individual citizens but also within the framework of journalistic or academic research activities or projects. To this effect, public offices and authorities can, for example, be obligated to make their files and processes public (principle of public access) or accessible (transparency of administration) and to define binding quality standards for access.

The Freedom of Information Act (IFG), which regulates access to information from the federal government, is how freedom of information is regulated in Germany. Until now, thirteen of the German Länder (states) have passed similar laws for their own jurisdictions. Bavaria, Lower Saxony, and Saxony, however, do not have any Freedom of Information Acts at the state (Land) level.

The IFG contains many exceptions that restrict or fully deny the right of access to information:

- Freedom of information exclusively refers to completed and documented processes and therefore does not grant access to ongoing projects (sec. 3 Protection of public interests, sec. 4 Protection of administrative decision-making processes).
- Freedom of information excludes personal data (sec. 5) and company data (sec. 6). Access to personal data can therefore only be granted if the applicant's interest in information outweighs the data subject's protectable interest or if the data subject has consented. There is no entitlement to access personnel files and personnel management systems. Information about the names or work addresses of employees, however, should generally be made available. The same applies to information about experts and appraisers.

Every individual or legal entity under private law is entitled to file applications (e.g., registered associations). Civic action groups and associations that are not acting as legal entities under private law are not entitled to apply.
Generally, applications need not be substantiated. Exceptions are made only in the event of infringements of third-party rights or to protect intellectual property or trade secrets. In that case, providing reasons is necessary to allow any third parties that must be notified by the authorities to decide whether or not they wish to consent.

Notes and recommendations:

In view of their potential use and depending on the type of data, open data and data that are requested under the IFG should be made available in a structured and machine-readable form.

It should be possible to find, search, and filter open data, and to process them using other applications. Currently, data provided by government offices are often only available as PDF files and therefore cannot be easily processed for research. Moreover, it should become easier to find such data, e.g. through suitable search engine optimisation or central listings (e.g., in the form of metadatabases).

3 Web scraping: data access and research data management

A common feature of individual and institutional access to big data is that they are based on contracts or binding agreements. The legal framework for open data and freedom of information is based on special laws at the federal and state (Länder) level.

With respect to web scraping, there are major uncertainties as to what is permitted or required under the law. This concerns the use of interfaces, but also the use of screen scraping methods that involve the retrieval of information from websites made for humans. Therefore, the German Data Forum (RatSWD) obtained a legal opinion from the Robotics Law Research Centre at Julius Maximilian University of Würzburg on the legal framework for these access paths. The following pages contain a brief overview of the potential of web scraping for research and a summary of the key points from the expert opinion. The full expert opinion can be found in the Appendix.[17]

Expert opinion p. 31

3.1 Case study: reputation and collaboration in anonymous online marketplaces

Many studies in the social, behavioural, and economic sciences use web scraping in order to, e.g., enrich their own data with data from the internet or to use these internet data as stand-alone data sources for descriptive or analytical, old or new research problems. An example is a study by Diekmann et al. (2014), published in the American Sociological Review, which explores the relationship between reputation and collaboration in anonymous online marketplaces. This study is especially interesting because it uses new data sources and measurement techniques to examine long-existing problems. Moreover, there is a wide range of other studies that use these data for new research questions (for an overview, see Gosling and Mason 2015).

Trust, reciprocity, and reputation are considered essential requirements for collaboration in social relationships, whether private or business in nature. Every exchange between anonymous stakeholders bears risks because every stakeholder may act more or less cooperatively or fraudulently. Repeated interactions between the same stakeholders reduce such problems because former experiences influence mutual expectations with respect to future interactions.

Users of anonymous internet auctions do not have any interaction history because sellers and buyers usually only interact once. Furthermore, fraudulent actions by one or the other party often cannot be prosecuted. Online marketplaces counter this lack of institutional safeguards for economic transactions by establishing reputation systems. The effectiveness of these reputation systems is based on the willingness of buyers and sellers to mutually rate each other after transactions have been completed. This is not a matter of course as reputation systems – representing a collective good – can be used by all stakeholders (free of charge) even if some do not participate themselves, i.e. do not provide any ratings. In their study, Diekmann et al. examined the mechanisms that are beneficial for collaborations of this kind. Firstly, they were specifically interested in whether the reputation itself has a market value that would make it rational for participants on online marketplaces to build a good reputation.

17 Note: In contrast to this paper, the expert opinion at one point describes screen scraping as being more "generic" than web scraping.
In the legal opinion and this article, the umbrella term "web scraping" refers to both the use of interfaces and the retrieval of data from websites. The statements in the legal opinion refer to web scraping.

Secondly, they studied the rating behaviour of participants because these (realistic) ratings are the foundation for effective reputation systems and thus the basis for functioning anonymous online marketplaces. They used auctions on the German-language version of the eBay platform as the data source. Relevant data were collected with screen scraping, and the researchers observed auctions for two selected products (mobile phones and DVDs). Over a period of one month, they collected data on the selected products from more than a million auctions. After the auctions were completed, the following information was collected by means of screen scraping:

1) Item page of the auction including HTML code;
2) List of bidders;
3) Seller and buyer profile pages;
4) Seller and buyer product pages;
5) Former ratings of the participants (rating history).

This one-month full survey was used to make a selection based on content so that ultimately 350,000 auctions could be analysed. The details of this data collection were published in an online supplement.[18] Interested researchers can find the data and also the analysis scripts online[19] so that it should be possible to replicate the analyses.

Whether auctions were successful was measured by two indicators: (1) sale (yes/no) and (2) achieved sales price. The reputations of sellers and buyers were measured by the number of positive and negative ratings. The rating behaviour after successful auctions was measured using different indicators, e.g., days until sellers and/or buyers rated transactions, the types of ratings (positive, negative, neutral), and former ratings submitted by the two parties for each other. It can further be stated that although the subject of the study was the behaviour of the individuals, the products (or auctions) were monitored and analysed, not the individuals.

The main findings of this study support the theory that reputations have a market value. In line with other studies, the authors found that a seller's positive (negative) reputation can have a positive (negative) effect on the probability of a sale and on the sales price. Furthermore, it showed that rating behaviour is strongly influenced by the principle of reciprocity: As soon as one of the parties rated the other, the probability of the other party also submitting a rating increased significantly. Alongside the principle of reciprocity, the authors noted another affect, characterised as altruistic. Buyers tended to rate sellers with very few positive ratings rather positively[20] while sellers with many positive ratings were rated more negatively. If stakeholders had rated each other for previous transactions, the probability of a new rating was lower.

18 https://journals.sagepub.com/doi/suppl/10.1177/0003122413512316 (last accessed: 15.04.2019).
19 https://www.ethz.ch/content/specialinterest/gess/chair-of-sociology/en/publikationen/data.html (last accessed: 15.04.2019).
20 According to Diekmann et al. (2014), this could indicate the willingness of (happy) buyers to help sellers build their reputation.

3.2 Collection, archiving, and secondary use of data

3.2.1 Data access

The expert opinion "Web scraping in independent academic research" by the Robotics Law Research Centre (FoRoRe) discusses data access and data archiving via web scraping under the aspect of competition law, copyright law, and general civil law. In this regard, the copyright act (UrhG) is particularly relevant for restrictions on use (Vogel and Hilgendorf 2020: 33). Especially the limitations rule under sec. 60d UrhG addresses the privileges of non-commercial academic research. While web scraping for commercial use requires consent and can be subject to claims for injunctive relief or damages in the event of violations, consent is not required for independent academic research if certain conditions are fulfilled.

The legal opinion determined that the following criteria must be met to qualify as non-commercial academic research:

1) The information to be analysed must be publicly accessible. This also applies to information that can only be retrieved after paying a fee (ibid.: 46).
2) Technical protective measures intended to prevent web scraping must not be circumvented. The expert opinion lists robots.txt files as an example of protective measures (ibid.).
3) Academic research must be for non-commercial use only (ibid.).
4) "The use of scraping technologies must not cause any technical damage to the operator's website [...]" (ibid.). Damage would occur, for example, "[...] where the mass retrieval of data severely strains or even overloads the server infrastructure and the proper operation of the website or database cannot be maintained – even if this is only temporarily the case" (ibid.: 44).
5) The rights holder is entitled to payment of an appropriate remuneration. This claim to remuneration can only be asserted by a collection society and not by the rightholders themselves. The institution where the researcher is employed owes this remuneration (ibid.: 45f.).

The last point especially implies that the rightholders of websites should be contacted before web scraping activities are carried out. After all, "[...] the parties negotiating the compensation will usually agree to the amount and type of remuneration" (ibid.: 45).

Contacting the other party is also recommended because it has not been conclusively clarified whether the place of a potential violation of copyrights in the event of a cross-border data transfer would be deemed to be the location of the server or the location of the scraper - in which case the copyright law of the server location would apply (ibid.: 48).

The retrieval of data via API usually results in the conclusion of a contract between the rightholder and the researcher. The latter consents to the terms of use. The legitimacy of such web scraping activities will then be subject to this contract (ibid.: 47f.).

3.2.2 Archiving and secondary use

If material obtained via web scraping contains personal data, the European General Data Protection Regulation (GDPR) must of course be observed (Vogel and Hilgendorf 2020: 47). Personal data is "[...] any information relating to an identified or identifiable natural person [...]" (Art. 4 no. 1 GDPR). Irrespective of the data's origin, researchers in Germany are subject to the German legal system (Vogel and Hilgendorf 2020: 48). The expert opinion points out that in the event of pseudonymisation, care must be taken that such data cannot be attributable to any identified or identifiable natural persons (ibid.: 49). For example, this would be the case if aliases contain the real names of the people using them.

Regarding the archiving and re-use of data collected through web scraping activities, the expert opinion further states that after a research project is completed, the data stored by the researcher must be deleted (= irretrievable deletion). For this purpose, researchers are recommended to draft a deletion plan in advance and then document their compliance with the deletion obligation. In this regard, completion of the research project also includes quality control steps, e.g., peer reviews (ibid.: 46).

However, a transfer to libraries, archives, museums, or educational institutions (privileged institutions within the meaning of sec. 60e and sec. 60f UrhG) is allowed to ensure long-term compliance reviews with respect to academic standards and to allow the use of the data in citations or as references (ibid.: 46f.). There are no rulings yet as to whether research data centres (RDCs) are deemed privileged institutions or not (ibid.: 41). The expert opinion concludes that "RDCs can – insofar as they do not serve any direct or indirect commercial purpose – qualify as privileged institutions on their merits" (ibid.: 41f.). Due to the heterogeneity of the RDCs accredited by the German Data Forum (RatSWD), however, no general statements can be made (ibid.: 42). Archiving institutions can provide the material they receive to other researchers for non-commercial research (ibid.: 46f.). It is not clear whether such provision must be in a form that excludes printing or storage (ibid.: 41). Corpora obtained via web scraping generally may not be transferred to academic journals (ibid.: 42).

4 Institutionalised access via third parties (fiduciary)

As explained in sections 2 and 3, individual access to private big data sources poses many challenges and uncertainties. Existing institutionalised data access paths and data collection via web scraping – even if the above recommendations are followed – cannot fulfil all the requirements that are necessary for sustainable and user-friendly data access for research. The German Data Forum (RatSWD) therefore suggests establishing independent research institutions or commissioning existing institutions with the task of acting as a fiduciary to provide standardised access to anonymised (micro)data from public or private big data sources.

A data trust would have to receive (micro)data from big data sources of certain companies on a contractual basis. This agreement would have to include provisions regulating which data can be transferred under which conditions, in which formats, and at which times. The data trust would have to act as a broker and represent the interests of both the researchers and the companies. For example, the data trust could demonstrate the added value that would be created for the companies when researchers analyse their data. An improved reputation would not be the only value added for companies; analyses of company data could also offer important insights into their activities, especially in the case of smaller enterprises. The data trust could also consult companies on the quality of their data sources or methods used.

International perspective

In the Netherlands, the collaboration between Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS), and individual enterprises, e.g., from the mobile phone or energy sectors, is based on cooperation agreements that include provisions regarding added values for both parties.

Another model is the Social Science One initiative that originates from and focuses on the academic system in the US. Under this model, a board of renowned academics negotiates the provision of data by large private enterprises within the framework of common research perspectives and then issues calls for tenders for research projects to use these data. As yet, this kind of collaboration exists solely with Facebook. Government agencies are not directly involved. This project was initiated by Harvard University. For more information, see: https://socialscience.one/overview.

The task of a data trust would be to develop and implement standardised access paths to anonymised (micro)data originating from big data sources for approved entities. The researchers would assist the data trust with respect to the related legal questions. It would be important that the data trust is an independent non-commercial and not-for-profit institution. The necessary infrastructure should therefore be financed by public research funding agencies, which could be a potential project for the National Research Data Infrastructure (NFDI)[21].

21 National Research Data Infrastructure (NFDI) is a project of the Joint Science Conference (Gemeinsame Wissenschaftskonferenz – GWK) established by the federal and state governments. It is intended to connect and expand the existing research data infrastructures in Germany and provide user-friendly services for the research community. Further information can be found at: https://www.dfg.de/nfdi (last accessed: 07.05.2020).

In addition to the actual provision of the data, the responsibilities of the data trust would include:

- Point of contact and interface between the research community and the producers of big data:
  The data trust can be contacted by researchers to assist them with gaining access to specific big data sources. The data trust can potentially make requests to producers of data to make more big data material available to researchers.
- Collection of tried and tested contracts and agreements for the regular provision of relevant data from big data sources
- Clarification of the legal requirements as to which entities have been approved for the use of data from big data sources, and under which conditions
- Clarification of the technical options for storing, processing, and linking data from heterogeneous big data sources
- Development of plans to anonymise and transfer data in accordance with the EU General Data Protection Regulation (GDPR) and the German Federal Data Protection Regulation
- Creation of suitable and standardised metadata descriptions of the managed data
- Provision of data in suitable and standardised formats to approved entities for user-friendly processing
- Consultation for interested entities about specific data requests and data uses

If data are transferred through a data trust, it would be generally helpful to have a legal framework that stipulated under which conditions data must be submitted to the data trust, and the requirements for providing data from big data sources that were processed by the data trust.

This legal framework should limit the transfer of data from the data trust to certain user groups – with special emphasis in this regard on the research community and official statistics. Data access should be regulated and granted to each group accordingly.

Research user group:

Independent researchers are interested in analysing data originating from big data sources, and use these analyses to answer their research questions or participate in discussions within national or international research communities. Access to data from big data sources would clearly improve the conditions for empirical science in Germany, and thus the international competitiveness of this field. In this context, it is essential for researchers that the findings of these analyses can be published and that analyses can be replicated by researchers working with the same datasets. It would be the data trust's responsibility to share the data while adhering to these basic conditions.

Official statistics user group:

Official statistics are interested in the use of data from big data sources, among others, to supplement their official data compilations with new content, to improve quality, and to keep data up to date. Potentially, the use of big data could help reduce obligations to collect and supply information. For such purposes, it would have to be ensured that the data trust provides microdata to the extent necessary.

This could lead to the data being integrated into the official statistics and thus becoming part of the data collection, processing, and publishing programme. If this were the case, it should be possible to make statistics enriched with big data available via the research data centres (RDC) of the federal and state statistical offices for academic purposes in accordance with the applicable regulations. To this effect, the German Statistical Advisory Committee (Statistischer Beirat) stated in its recommendations on the development of official statistics: Today, the data in the official statistics are mainly provided to the research community via their RDCs in compliance with the statutory requirements. In the future, this should also be possible for the new digital data processed in official statistics. Furthermore, findings from the official statistics must remain accessible for academic review (Statistischer Beirat 2018: 11).

A concept of how data could be provided by the data trust to the approved user groups (research community and official statistics) needs specification. It would be reasonable to use the current infrastructure and the current regulations for the RDCs of the federal and state statistical offices as templates. Among others, the following conditions could apply:

- Access to data from big data sources is granted to all users from the approved user groups (e.g., academic institutions, statistical offices) based on applications that must include information about, among others, the purposes for which the data are to be used.
- The use of the data is subject to a time limit that can differ depending on the user group (for research, this might be the duration of a project; for official statistics, the duration of a specific survey).
- The data can be used repeatedly by the same user as long as the purpose of use is admissible.
- The same (standardised) data and metadata can also be used by other users from approved user groups (e.g., other academic institutions, other statistical offices).
- In accordance with the legal basis, access to data depends on their degree of anonymisation. The bar for access to aggregated or de facto anonymised data could be lower than for access to less anonymised microdata.
- In accordance with the legal basis, access to data is based on the user group. Official statistics offices could receive some of the data through other channels and in other formats than the research community.
- Publications of findings based on the use of the data must name the sources used.
- Institutions could cooperate with other organisations within the approved user groups when using the data.

Notes and recommendations

To strengthen the German research infrastructure in social, behavioural and economic sciences, the German Data Forum (RatSWD) suggests establishing independent research institutions or commissioning existing institutions with the task of providing standardised access to anonymised (micro)data from public or private big data sources on a trust basis. The infrastructure required for this could be a key issue for the dynamically developing National Research Data Infrastructure (NFDI).

5 Glossary

AJAX	Asynchronous JavaScript and XML (web programming technique)
API	Application Programming Interface
BDSG	German Federal Data Protection Act
BGB	German Civil Code
BGH	German Federal Supreme Court
Crawling	Retrieval of documents from the internet that can be obtained by following hyperlinks. The hyperlinks to these documents are also starting points for further queries.
CSS	Cascading Style Sheets (style language used to design digital, mostly web-based documents)
Destatis	Federal Statistical Office (Germany)
GDPR	General Data Protection Regulation
	Statistisches Bundesamt (Deutschland)
CJEU	European Court of Justice
RDC	Research data centre
GWK	Joint Science Conference (coordinates joint funding for the science community from the federal and state governments in Germany)
HTML	Hypertext Markup Language (text-based markup language)
	Hypertext Markup Language (textbasierte Auszeichnungssprache)
IFG	Freedom of Information Act
IP	Internet protocol
LG	Regional court
LOD	Linked open data (standardised, retrievable, clearly identifiable data that are freely available on the internet)
NFDI	National Research Data Infrastructure (Germany)
NSF	National Science Foundation (Research funding agencies USA)
OLG	Higher regional court in Germany
PDF	Portable Document Format (file format)
RatSWD	Rat für Sozial- und Wirtschaftsdaten (German Data Forum)
RDF	Resource Description Framework (formal language for the provision of metadata on the internet)
	Ressource Description Framework (formale Sprache zur Bereitstellung von Metadaten im Internet)
RL	Directive
RWI	Leibniz Institute for Economic Research
Screen Scraping	Collection of information by targeted extraction of data that are used to (typo)graphically present contents on users' end devices (sub-concept of web scraping)
Smart Meter	Digital meter and consumption meter, for example, for electrical energy, natural gas, district heating, or water)

SMS	Short message service (telecommunication service for text-based short messages)
Spidering	See “Crawling”
UrhG	Act on Copyright and Related Rights
UrhWissG	Act to adjust the copyright law to the current requirements of a knowledge society
URI	Uniform Resource Identifier (identifier for resources on the internet)
URL	Uniform Resource Locator (naming standard for network resources)
UWG	Act against Unfair Competition
Web Scraping	Collection of information by targeted extraction of data that are available on the World Wide Web (see also “Screen Scraping” for further differentiation)
XML	Extensible Markup Language

6 References

Adjerid, Idris and Ken Kelley (2018): Big data in psychology: A framework for research advancement. American Psychologist 73(7), 899–917.

An de Meulen, Philipp; Martin Micheli and Sandra Schaffner (2014): Documentation of German Real Estate Market Data – Sample of Real Estate Advertisements on the Internet Platform ImmobilienScout24. RWI Materialien 80. https://www.rwi-essen.de/publikationen/rwi-materialien/327 (last accessed: 24.05.2019).

Boeing, Geoff and Paul Waddell (2016): New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. https://journals.sagepub.com/doi/full/10.1177/0739456X16664789 (last accessed: 07.05.2020).

Boelmann, Barbara and Sandra Schaffner (2019): FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED) – Advertisements on the Internet Platform ImmobilienScout24. RWI Projektberichte. Essen. http://www.rwi-essen.de/media/content/pages/publikationen/rwi-projektberichte/pb_fdz_-_rwi-geo-red_data_description.pdf (last accessed: 24.05.2019).

Bug, Mathias (2015): Ansätze und Datenquellen in der Kriminalitätsmessung: ein Überblick zu den offen zugänglichen WISIND-Daten. DIW Vierteljahrshefte zur Wirtschaftsforschung 84(2), 69–101. https://doi.org/10.3790/vjh.84.2.5 (last accessed: 07.05.2020).

Carrière-Swallow, Yan and Felipe Labbé (2013): Nowcasting with Google Trends in an Emerging Market. Journal of Forecasting 32(4), 289–298.

Chen, Le; Alan Mislove and Christo Wilson (2016): An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. https://cbw.sh/static/pdf/amazon-www16.pdf (last accessed: 07.05.2020).

Davis, Donald R.; Jonathan I. Dingel; Joan Monras and Eduardo Morales (2019): How Segregated is Urban Consumption? Journal of Political Economy 127(4), 1684-1738. http://faculty.chicagobooth.edu/jonathan.dingel/research/davisdingelmonrasmorales.pdf (last accessed: 07.05.2020).

Di Bella, Enrico; Lucia Leporatti and Filomena Maggino (2018): Big Data and Social Indicators: Actual Trends and New Perspectives. Social Indicators Research 135(3), 869–878. https://doi.org/10.1007/s11205-016-1495-y (last accessed: 07.05.2020).

Diekmann, Andreas; Ben Jann; Wojtek Przepiorka and Stefan Wehrli (2014): Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets. American Sociological Review 79(1), 65–85.

Edelman, Benjamin (2012): Using Internet Data for Economic Research. The Journal of Economic Perspectives 26(2), 189–206.

Einav, Liran and Jonathan Levin (2014): The Data Revolution and Economic Analysis. in: Lerner, Josh und Scott Stern (Hrsg.): Innovation Policy and the Economy. National Bureau of Economic Research Innovation Policy and the Economy, 1–24.

Fishman, Elliot (2016): Bikeshare: A Review of Recent Literature. Transport Reviews 36(1), 92–113.

Frees, Beate and Wolfgang Koch (2018): ARD/ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. Media Perspektiven 2018(9): 398-413. https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2018/0918_Frees_Koch_2019-01-29.pdf (last accessed: 07.05.2020).

Gosling, Samuel D. and Winter Mason (2015): Internet research in psychology. Annual Review of Psychology 66, 877–902.

Gyódi, Kristóf (2017): Airbnb and Booking.com: Sharing Economy Competing Against Traditional Firms? Working Paper DELab UW 2017(3). http://www.delab.uw.edu.pl/wp-content/uploads/2017/09/WP_3_2017_K.Gyodi_.pdf (last accessed: 07.05.2020).

Hadam, Sandra (2018): Nutzung von Mobilfunkdaten für amtliche Statistiken. Methoden - Verfahren - Entwicklungen. Nachrichten aus dem Statistischen Bundesamt 2018(2), 6–9.

Hannak, Aniko; Claudia Wagner; David Garcia; Alan Mislove; Markus Strohmaier and Christo Wilson (2017): Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr. CSCW '17 Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM, 1914–1933. https://doi.org/10.1145/2998181.2998327 (last accessed: 07.05.2020).

Hyunyoung, Choi and Hal Varian (2012): Predicting the Present with Google Trends. Economic Record 88(S1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x (last accessed: 07.05.2020).

Kinne, Jan and Janna Axenbeck (2018): Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany, ZEW Discussion Paper No. 18-033. http://ftp.zew.de/pub/zew-docs/dp/dp18033.pdf (last accessed: 07.05.2020).

Lazer, David; Ryan Kennedy; Gary King and Allessandro Vespignani (2014): The Parable of Google Flu: Traps in Big Data Analysis. Science 343(6176), 1203–1205.

McCormick, Tyler H.; Hedwig Lee; Nina Cesare; Ali Shojaie and Emma S. Spiro (2017): Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods & Research 46(3), 390–421. https://doi.org/10.1177/0049124115605339 (last accessed: 07.05.2020).

McLaren, Nick and Rachana Shanbhogue (2013): Using Internet Search Data as Economic Indicators. Bank of England Quarterly Bulletin 51(2), 134–140.

Moat, Helen Susannah; Chester Curme; Adam Avakian; Dror Y. Kennett; Eugene Stanley and Tobias Preis (2013): Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Scientific Reports 3(1801). https://www.nature.com/articles/srep01801; https://doi.org/10.1038/srep01801 (last accessed: 07.05.2020).

Morstatter, Fred; Jürgen Pfeffer; Huan Liu and Kathleen M. Carley (2013): Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. Seventh International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPaper/6071 (last accessed: 07.05.2020).

Ørmen, Jacob (2019): From Consumer Demand to User Engagement: Comparing the Popularity and Virality of Election Coverage on the Internet. The International Journal of Press/Politics 24(1), 49–68. https://doi.org/10.1177/1940161218809160 (last accessed: 07.05.2020).

Pfeffer, Jürgen; Katja Mayer and Fred Morstatter (2018): Tampering with Twitter's Sample API. EPJ Data Science 7(1), 1–21.

Powell, Ben; Guy Nason; Duncan Elliott; Matthew Mayhew; Jennifer Davies and Joe Winton (2017): Tracking and modelling prices using web‐scraped price microdata: towards automated daily consumer price index forecasting. Journal of the Royal Statistical Society, Series A. Statistics in Society 181(3), 737–756. https://doi.org/10.1111/rssa.12314 (last accessed: 07.05.2020).

Reinsel, David; John Gantz and John Rydning (2018): Data Age 2025. The Digitization of the World From Edge To Core. IDC White Paper, November 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (last accessed: 07.05.2020).

Rieckmann, Johannes and Jan-Lucas Schanze (2015): Sicherheitsempfinden in sozialen Medien und Suchmaschinen – ein realistisches Abbild der Kriminalitätsbelastung? DIW Wochenbericht 2015(12), 271–279. https://www.diw.de/documents/publikationen/73/diw_01.c.498951.de/15-12.pdf (last accessed: 07.05.2020).

Schmidt, Torsten and Simeon Vosen (2011): Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Journal of Forecasting 30(6), 565–578.

Schmidt, Torsten and Simeon Vosen (2012): A Monthly Consumption Indicator for Germany Based on Internet Search Query Data. Applied Economics Letters 19(7), 683–687.

Schmidt, Torsten and Simeon Vosen (2013): Forecasting Consumer Purchases Using Google Trends. Foresight: The International Journal of Applied Forecasting (30), 38–41.

Sen, Indira; Fabian Flöck; Katrin Weller; Bernd Weiß and Claudia Wagner (2019): A Total Error Framework for Digital Traces of Humans. Computing Research Repository. Working Paper. July 22, 2019. arXiv:1907.08228.

Slivko, Olga (2018): 'Brain Gain' on Wikipedia: Immigrants Return Knowledge Home. ZEW - Centre for European Economic Research Discussion Paper 18(008). https://doi.org/10.2139/ssrn.3124193 (last accessed: 07.05.2020).

Statistischer Beirat (2018): Fortentwicklung der amtlichen Statistik. Empfehlungen des Statistischen Beirats für die Jahre 2018 bis 2022. https://www.destatis.de/DE/Ueber-uns/Leitung-Organisation/Statistischer-Beirat/FortentwicklungNov2018_2022_Teil3.pdf (last accessed: 07.05.2020).

Verma, Amit; Kirill M. Yurov; Peggy L. Lane and Yuliya V. Yurova (2019): An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements, Journal of Education for Business, 94:4, 243–250, https://doi.org/10.1080/08832323.2018.1520685 (last accessed: 07.05.2020).

Vogel, Paul and Eric Hilgendorf (2019): Web scraping in independent academic research. Expert opinion on behalf of the Berlin Social Science Centre (WZB) – German Data Forum (RatSWD). In: RatSWD: Big data in social, behavioural, and economic sciences: data access and research data management. RatSWD Output 4(6). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.52. Part of this publication, p. 31.

Von Lucke, Jörn and Christian Geiger (2010): Open Government Data. Frei verfügbare Daten des öffentlichen Sektors. Gutachten für die Deutsche Telekom AG zur T-City Friedrichshafen. https://www.zu.de/institute/togi/assets/pdf/TICC-101203-OpenGovernmentData-V1.pdf (last accessed: 07.05.2020).

Von Schönfeld, Max (2018): Screen Scraping und Informationsfreiheit. Baden-Baden, Nomos.

Yongping, Zhang and Mi Zhifu (2018): Environmental benefits of bike sharing: A big data-based analysis. Applied Energy 220, 296–301. https://doi.org/10.1016/j.apenergy.2018.03.101 (last accessed: 07.05.2020).

Contributors to this report

Working group members

Prof. Dr. Thomas K. Bauer *(Co-Chair Working group)*
RWI – Leibniz Institute for Economic Research, Ruhr-University Bochum, German Data Forum (RatSWD)

Prof. Dr. Michael Eid
Free University Berlin, German Data Forum (RatSWD)

Hans-Josef Fischer *(Co-Chair Working group)*
Landesbetrieb Information und Technik Nordrhein-Westfalen (IT.NRW), German Data Forum (RatSWD)

Dr. Fabian Flöck
GESIS – Leibniz Institute for the Social Sciences

Prof. Dr. Anja Göritz
Albert-Ludwigs-University of Freiburg, German Data Forum (RatSWD)

Heike Habla
German Pension Insurance, German Data Forum (RatSWD)

Prof. Dr. Kai Maaz
DIPF | Leibniz Institute for Research and Information in Education, Goethe University Frankfurt am Main, German Data Forum (RatSWD)

Sabine Ohsmann
German Pension Insurance, German Data Forum (RatSWD)

Prof. Dr. Mark Trappmann
Institute for Employment Research (IAB) of the German Federal Employment Agency, Otto-Friedrich-University of Bamberg, German Data Forum (RatSWD)

Dr. Heike Wirth, GESIS
GESIS – Leibniz-Institute for the Social Sciences, German Data Forum (RatSWD)

Consultation

Prof. Dr. Wolfgang Nagel
ScaDS Dresden/Leipzig – Competence Center for Scalable Data Services and Solutions, Technische Universität Dresden

German Data Forum (RatSWD) business office

Dr. Mathias Bug

Dr. Tim Deeken

Dr. Nora Dörrenbächer

Thomas Runge

Appendix

robot forschungs | recht stelle

Robotics Law Research Centre
Julius Maximilian University of Würzburg

Web scraping in independent academic research

Expert opinion on behalf of the Berlin Social Science Centre (WZB) – German Data Forum (RatSWD)

Paul Vogel,
Prof. Dr. Dr. Eric Hilgendorf

Würzburg, August 30th, 2018

Recommended citation: Paul Vogel and Eric Hilgendorf (2020): Web scraping in independent academic research. Expert opinion on behalf of the Berlin Social Science Centre (WZB) – German Data Forum (RatSWD). In: RatSWD: Big data in social, behavioural, and economic sciences: data access and research data management. RatSWD Output 4 (6). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.52.

Contents

Executive Summary

In social, behavioural and economic research, we witness an increasing use of web scraping technology to automatically retrieve and compile data that are publicly available on the internet. This makes it easier to discover anomalies and correlations that would be hard or even impossible to find manually due to the large amounts of data.

However, from a legal point of view, there are concerns regarding the legitimacy of this method. Questions particularly arise under competition, copyright and general civil law. This expert opinion finds that web scraping for academic research does not have any implications under competition law because this research does not usually involve any "commercial activity" that could give reason to such implications. In order to evaluate the permissibility of web scraping for academic research from a copyright perspective, more profound justification is needed. The queried websites and databases are mainly protected works or database works within the meaning of the copyright act (UrhG). Consequently, every use of the information – and this also applies to web scraping – would be subject to consent and compensation and could result in serious claims for injunctive relief or damages in the event of violations. The German legislators responded to some – in part contradictory – decisions of the highest courts, the German Federal Court of Justice (BGH) and the European Court of Justice (CJEU), and included a limitation rule in the copyright act (UrhG) that applies to academic research and, under certain circumstances, exempts text and data mining, as is practised in web scraping, from the requirement of consent. Researchers can generally assume that their activities are permissible under copyright law, whereas the legitimacy of commercial scrapers is highly questionable. Often, contractual provisions do not exist for scraping processes, and, if applied wisely, the virtual property rights (Hausrecht) of website operators do not contradict scraping operations.

Finally, this expert opinion determines some criteria for the use of web scraping in academic research based on the discussions of the legal position and for the matters of archiving and provision following the scraping process.

As a result, it can be noted that, from a legal perspective, scientists can certainly use the method of web scraping to support their research activities; however, they must comply with several requirements to avoid any conflicts with the law.

A. Introduction and problem statement

In a data-driven society with a global data volume that is expected to multiply tenfold between 2016 and 2025, from 16.1 to 163 zettabyte (Seagate n. d.), the value of data increases in proportion to the growing number of options to link and analyse it. Big data is no longer just a fashionable buzzword but has become an established business practise that will generate sales of €6.4 billion in 2018 in Germany alone (Bitkom 2018). This upswing confirms the frequently quoted comparison of data being the "new oil"[1]; consequently, establishing new data sources has become extremely important.

In recent years, a wide range of methods have been developed for this purpose. Web scraping is a technique for automatically retrieving contents from third-party databases. It is important to differentiate web scraping from the more generic term "screen scraping". Web scraping involves the retrieval of data from websites or databases provided by websites (e.g. programming interfaces [API]) (von Schönfeld 2018: 25). The following explanations refer to this use case. Search engines are common use cases that extract masses of data from websites – for example from hotels – and display them in the form of a list of available hotel rooms, including rates, for the customers. The resulting "follow-up market" and its comparison platforms have gained enormous economic and social relevance (Schapiro and Zdanowiecki 2015: 497(498)). More and more consumers now prefer these comparison platforms to the proprietary websites of retailers or providers of goods and services (Bitkom 2013: 32f.). Conflicts between the relevant providers and web scrapers are therefore almost inevitable (Schapiro and Zdanowiecki 2015: 497(498)).

Nonetheless, web scraping has gained popularity not only for business purposes. The various disciplines of academic research that depend on analysing large volumes of data increasingly use this technology for their own benefit to automatically retrieve data from publicly accessible databases. This saves time and allows researchers to reserve their manpower for the subsequent analysis of the large volumes of data. Moreover, growing computer capacities enable researchers to find correlations and anomalies that would have been overlooked in manual analyses or when using traditional methods, simply due to the huge volumes of data (cf. Mörike 2018: 2).

From a technical perspective, web scraping usually involves two basic steps: first, a web bot (also referred to as a crawler) accesses a website. Then, the information there is analysed and extracted, if necessary (von Schönfeld 2018: 51). The problem is that websites are designed to be accessed by people and not by machines. Often there are no machine-readable versions of the accessed websites are available, which, as a result, considerably impairs the retrieval of data (Kukulenz 2008: 22). Nevertheless, as the technological developments advance, the options for differentiating between relevant and irrelevant information also increase.

1 Former EU commissioner *Meglena Kuneva* with respect to personal data in a speech in Brussels on 31 March 2009 – SPEECH/09/156 ("Personal data is the new oil of the Internet and the new currency of the digital world.").

In summary, web scraping can be described as a "generic term for an algorithm-based method to retrieve publicly-accessible information and data from the World Wide Web by technological means. For this purpose, a screen scraping program simulates human behaviour to gain access to websites, retrieve, and then analyse the available information and data" (von Schönfeld 2018: 58).

The legal limits of web scraping have yet to be clarified. The same is true for the requirements that must be fulfilled – especially within the context of academic research. The interests of academic researchers differ greatly from those of commercial services such as comparison platforms; after all, researchers do not practice scraping for purely pecuniary reasons. Instead, the interests of database or website providers must be aligned with the (general or public) interest of unhindered academic research as much as possible.

B. Legal requirements

For the most part, the legitimacy of web scraping must be evaluated by considering three fields of law. Firstly, problems could arise under competition law. Especially when web scraping is used for flight agency and price comparison platforms, conflicts with the law against unfair competition (UWG) are almost inevitable. Whether such conflicts also arise within the context of academic research must first be explored.

Secondly, the implications under copyright law are considerably more relevant and must therefore be analysed in more depth. Databases and their authors (or makers) benefit from a wide range of protections under the copyright act (UrhG). However, the legislators have recently privileged the access to third-party database works for the purpose of academic research. Following a series of divergent decisions of the highest courts, the issue of the legitimacy of web scraping under copyright law may now be conclusively settled.

Lastly, considerations under contract law must not be neglected: depending on the design of the "scraped" database, a user licence agreement could limit retrieval options. The unauthorised use of web crawlers, in particular, could interfere with a website operator's "virtual property rights" *(Hausrecht)*.

I. Competition law

Implications under competition law are often discussed when it comes to a legal evaluation of web scraping. The German Federal Court of Justice (BGH) has considered these issues from the perspective of competition law several times. The legal matter that is mainly regulated in the law against unfair competition (UWG) involves prohibitions of certain actions and penalties for violations of its provisions, sometimes even sentences under criminal law.

Within the context of web scraping, discussions mostly involve violations of the provisions under the Act against Unfair Competition pertaining to imitations (sec. 4 no. 3 UWG), the deliberate obstruction of competitors (Sec. 4 no. 4 UWG), and misleading competitive information (sec. 5 para. 1 sentence 2 no. 1 UWG). All of these practises are illegal, giving reason to the legal consequences listed under sec. 8 et. seq. UWG such as claims for injunctive relief or damages. A common requirement for all these provisions, however, is the existence of a commercial practise as defined under sec. 2 para. 1 no. 1 UWG.

Sec. 2 UWG [Act against Unfair Competition] – definitions[2]

(1) Within the meaning of this Act the following definitions shall apply:

1. "Commercial practise" means any conduct by a person for the benefit of that person's or a third party's business before, during or after the conclusion of a business transaction, which conduct is objectively connected with promoting the sale or the procurement of goods or services, or with the conclusion or the performance of a contract concerning goods or services; "goods" shall be deemed to include immovable property as well, and "services" to also include rights and obligations;

According to this definition, web scraping would have to be objectively related to promoting the sale or purchase of goods and services or to concluding or executing a contract pertaining to goods or services. Academic research is usually not connected to such purposes, which means that there will not generally be any conflicts under competition law when web scraping is carried out by researchers (cf. also Mörike 2018: 4).

2 https://www.gesetze-im-internet.de/englisch_uwg/englisch_uwg.html (last accessed: 07.05.2020). This translation is not binding and has no legal effect for compliance or enforcement purposes (c.f. https://www.gesetze-im-internet.de/Teilliste_translations.html (last accessed: 07.05.2020)).

II. Copyright

Copyright is one of the key rights regarding intellectual property and is largely codified within the German legal system in the German Act on Copyright and Related Rights (UrhG). The rules of the act grant primary and subjective exclusive rights (exploitation rights) if the requirements for protection are fulfilled (particularly if personal intellectual creations are involved) (Rehbinder and Peukert 2019: Sec. 1 marginal nos. 7, 9). It is designed to protect owners of copyrights and related proprietary rights and their personal and intellectual relationships to their works and to ensure that an exploitation by third parties is appropriately compensated (Nordemann 2014: Einleitung UrhG marginal no. 8).

Web scraping could violate the relevant proprietary rights of the initiators of the queried website or database.

1. Protection of database works, sec. 4 para. 2 UrhG

Operators of a queried website or database could benefit from database work protection under sec. 4 para. 2 UrhG. Within the meaning of this provision, database works are compilations whose elements are systematically or methodically arranged and individually accessible. In order for a database or website to qualify as a compilation, the selection or arrangement of its content must be the author's own intellectual creation (Sec. 4 para. 1 UrhG).[3] The decisive factor is whether it is possible to make discretionary decisions; this is not the case if a selection or arrangement must be sorted according to strict criteria (for example, alphabetically or chronologically) (Marquardt 2014: Sec. 4 marginal no. 9; Kotthoff 2013: Sec. 4 marginal no. 8). Simply collecting and updating data in a database, including a website, therefore does not achieve the required degree of creativity and, as a consequence, does not meet the criteria to qualify as a work (Leupold and Demisch 2000: Sec. 4 marginal no. 10; Schapiro and Zdanowiecki 2015: 497 (499)). Most of the scraped contents do not qualify as database works within the meaning of the provision so that the protection of their authors according to sec. 4 para. 2 UrhG does not conflict with web scraping processes (cf. also von Schönfeld 2018: 191).

2. Protection of database makers, sec. 87a, 87b UrhG

Sec. 87a et. seq. UrhG, which are based on Directive 96/9/EC of the European Parliament and of the Council on the legal protection of databases[4], grant protection to the makers of databases, referred to as sui generis database rights, in recognition of the investment that is made in compiling a database (Dreier 2018: before Sec. 87a et. seq. marginal no. 1). According to sec. 87b para. 1 sentence 1 UrhG, it is the exclusive right of a database maker to copy, disseminate or publish a database in full or essential parts of it. Nonetheless, these provisions do not protect the contents of a database, for example in the form of individual data records or information. Such protection is instead provided for the compilation and systematising of a database (cf. von Schönfeld 2018: 205). Creativity is not a requirement – as opposed to the protection of a database work as set out under sec. 4 para. 2 UrhG (Kotthoff 2013: Sec. 87a marginal no. 1). In order for someone to qualify as maker, it is not relevant whether or not they did in fact create or maintain the database; the relevant factor is instead who bears the economic risk – that is the costs – of its operation (Shapiro and Zdanowiecki 2015: 497 (499)).

a) Suitable object of protection

To qualify as a suitable object of protection in the meaning as set out under sec. 87a para. 1 UrhG, a database must consist of independent elements that are systematically or methodically arranged and individually accessible by electronic or other means; in addition, the obtaining, verification, or presentation must require substantial investment.[5]

In the area of e-commerce, platform databases and databases such as, for example, review sites, online marketplaces for vehicles, or flight schedule databases are generally considered to be databases

3 Cf. also *CJEU*, judgment of 01 March 2012 – C-604/10 – *Football Dataco*.
4 Directive 96/9/EC of the European Parliament and of the Council of 11.03.1996 on the legal protection of databases, law gazette No. L 77, 20 et. seq.
5 Insofar, the definitions of the terms are identical with those in Art. 1 para. 2 of the database directive 96/9/EC.

within the meaning of sec. 87a para. 1 sentence 1 UrhG.[6] Assessments as to whether static HTML code can qualify as a database within this meaning differ in the jurisprudential literature. Despite the broad definition of the term database[7] as intended by the EU legislators, it would be difficult to claim that plain HTML code meets the requirements set out in sec. 87a para. 1 UrhG. After all, an aggregation of digital data in the form of HTML code does not serve to make individual elements of a database available to users but solely to display the website (von Schönfeld 2018: 213: likewise, Schack 2001: 9 (11f.)). Nonetheless, websites that provide search functions and thus allow users to directly access individual elements of a website meet the criteria of a database (von Schönfeld 2018: 214). As a rule, however, websites that define themselves as collections of independent elements (e.g. online lexicons or online encyclopaedia) do meet these requirements (Thum and Hermes 2014: Sec. 87a marginal no. 94; Vogel 2017: Sec. 87a marginal no. 28). Social media platforms, such as Facebook and Twitter that are based on content management systems (CMS) like most other dynamic Web 2.0 websites must be classified as databases within the meaning of sec. 87a para. 1 UrhG because the CMS stores the contents in databases and makes them available for quick retrieval through indices (Thum and Hermes 2014: Sec. 87a marginal no. 95).

According to rulings of the German Federal Court of Justice (BGH), a significant investment can be assumed to exist if under objective considerations, the effort required to create the database were considerable and could not be easily provided by anyone.[8] The required effort to obtain, verify, or present the contents of the queried database or website determines whether a suitable object of protection exists or not.

b) Web scraping as an act of exploitation

After a compilation of third-party content has been positively classified as a database within the meaning of sec. 87a para. 1 sentence 1 UrhG, the question arises whether the technological process of web scraping represents an act of exploitation that potentially impairs the database maker's rights or not. According to sec. 87b para. 1 sentence 1 UrhG, it is the exclusive right of a database maker to copy, disseminate, or publish the database in full or essential parts of it. The decisive question therefore is whether essential elements of a database are copied during the automated mass retrieval of data through web scraping.

aa) Precedents in Germany

Thus far, the precedents in the context of web scraping have not followed a uniform guideline or allowed for the establishment of certain principles. In the "Online market place for automobiles" case regarding the permissibility of scraping services, the Hamburg regional court ruled in favour of the maker of the database and found the scraping to represent a copying of substantial parts of the database.[9] Copying must be assumed if the use of a part of the database posed the risk of causing "considerable damage to the amortisation of the investment for the maker of the database".[10] The Hanseatic higher regional court reversed this decision[11] and their decision was finally confirmed by the BGH as the court of appeal. The BGH based its ruling on the conviction that the provider of the scraping software was not the main violator of the rights of the maker of the database but that the violation was committed by the user alone.[12]

In a more recent case regarding this issue, in which the airline *Ryanair* defended itself against an air travel comparison platform, the Hanseatic higher regional court concurred with this argumentation

6 Cf. *BGH*, judgment of 01.12.2010 – I ZR 196/08, marginal no. 15 – *Second dentist's opinion II* = GRUR 2011, 724; *BGH*, judgment of 22.06.2011 – I ZR 159/10, marginal no. 27f. – *On-line market place for automobiles* = NJW 2011, 3443.
7 Cf. *CJEU*, judgment of 09.11.2004 – C-444/02, marginal no. 22 – *Fixtures Marketing II* = GRUR 2005, 254.
8 *BGH*, judgment of 01.12.2010 – I ZR 196/08, marginal no. 23 – *Second dentist's opinion II* = GRUR 2011, 724 with further references.
9 *LG Hamburg*, judgment of 09.04.2009 – 310 O 39/08 = BeckRS 2009, 20109.
10 *LG Hamburg*, judgment 09.04.2009 – 310 O 39/08, marginal no. 60 (juris) = BeckRS 2009, 20109.
11 *OLG Hamburg*, judgment of 18.08.2010 – 5 U 62/09 = GRUR 2011, 728.
12 *BGH*, judgment of 22.06.2011 – I ZR 159/10, Rn. 20ff. – *On-line market platform for automobiles* = NJW 2011, 3443.

by deciding in favour of the latter.[13] Although the BGH, as the court of last resort, had to try this case, it did not have to make a decision on whether the rights of the maker of the database had been infringed on as the decision in this respect had already become final and conclusive.[14] Therefore, the question of the permissibility of web scraping under the database-related laws was considered closed for Germany (Schapiro and Zdanowiecki 2015: 497(499); approved by von Schönfeld 2018: 244).

bb) Precedents of the European Court of Justice

Two judgments that were later passed by the European Court of Justice (CJEU) have materially challenged the German rulings.

(1) "Innoweb/Wegener" decision

This precedent was based on the "Innoweb/Wegener" decision in which the CJEU banned web scraping via what is referred to as a metasearch engine.[15] Similar to the earlier BGH ruling in the "online market place for automobiles" case, the issue involved a website that allowed users to search for specific vehicles offered for sale via a search engine and different search criteria. This website browsed third-party websites, matching them against the entered parameters, searched their databases, and finally presented the relevant hits as search results on their own website (cf. Schapiro and Zdanowiecki 2015: 497(499)). As opposed to the BGH, the CJEU set a lower bar for the protection of database maker rights: even the provision of a platform qualified as an infringement of database maker rights so that the actions of the users were no longer relevant factors in the decision.[16] The differences of opinion are mainly due to the fact that the European Database Directive 96/9/EC considered the range of uses reserved for the makers of databases to be much broader ("extraction" and "re-utilization", Art. 7 para. 2 of the directive) than the German legislators did ("reproduction", "distribution", or "communication to the public", sec. 87b, para. 1 UrhG) (Schapiro and Zdanowiecki 2015: 497(499)).

Article 7 Directive 96/9/EC – object of protection

(2) For this chapter, the following definitions apply:

a) "Extraction" means the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form.

b) "Re-utilization" means any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. The first sale of a copy of a database within the Community by the rightholder or with his consent shall exhaust the right to control resale of that copy within the Community.

Public lending is not an act of extraction or re-utilization.

Especially the omnibus clause "or any other form of transmission" within the meaning of Art. 7 para. 2 letter b of the Directive is traditionally interpreted in a broad sense and this was also applied in the CJEU decision (cf. Schapiro and Zdanowiecki 2015: 497(499) with further references). As opposed to the BGH, the CJEU puts less emphasis on the technical functionality and instead deemed it highly relevant to protect the investment of the database maker (Schapiro and Zdanowiecki 2015:497(500)).[17] Consequently, in this specific case, the CJEU considered the web scraping operations of the defendant platform operator to be an infringement of the claimant's database maker rights and, as a result, the claimant to be entitled to assert claims for injunctive relief or damages (cf. to German law, sec. 97 UrhG).

Nonetheless, the decision in the "Innoweb/Wegener" case is not suitable as a precedent of general importance for all scraping situations: in its decision, the CJEU emphasised the very special circumstances of the case and generally described the platform as a "specialised metasearch engine"

13 *OLG Hamburg*, judgment of 24.10.2012 – 5 U 38/10, marginal no. 183f. (juris) = GRUR-RS 2012, 22946.
14 *BGH*, judgment of 30.04.2014 – I ZR 224/12, marginal no. 20 – *On-line flight agency* = MMR 2014, 740.
15 *CJEU*, judgment of 19.12.2013 – C-202/12 – *Innoweb/Wegener* = MMR 2014, 185.
16 *CJEU*, judgment of 19.12.2013 – C-202/12, marginal no. 37ff. – *Innoweb/Wegener* = MMR 2014, 185.
17 *CJEU*, judgment of 19.12.2013 – C-202/12, marginal no. 36 – *Innoweb/Wegener* = MMR 2014, 185.

(von Schönfeld 2018: 246). Moreover, it established criteria for the applicability of its decision: its statements apply solely to metasearch engines "that provide a search form for end users that contains most of the same options as the database's search form", "that submits the end user's queries to the search engine in real-time", and "that displays the search results in the style of their website".[18]

(2) "Ryanair" decision

About a year later, another decision of the CJEU sparked further discussions about the permissibility of web scraping. After Ryanair took legal action against a flight comparison platform in the Netherlands, the CJEU decided in favour of the claimant - as opposed to the BGH in the earlier case.[19] The reasoning for this, however, differed significantly from that in the "Innoweb/Wegener" case. The appeal courts in the Netherlands had ruled that the flight booking database on Ryanair's website did not represent a database work or database in the meaning of the copyright act because it lacked significant investments in the collection and recording of contents in a database[20] – the CJEU was bound by these findings in its decision (Schapiro and Zdanowiecki 2015: 497(500)).

The court then had to deal with the issue of whether a database that is not protected under the database directive, such as the Ryanair database, can nevertheless be lawfully queried in accordance with the limitation provisions as set out under Art. 6 and 8 of the directive. This was of importance due to the fact that Ryanair's general terms and conditions, which users are required to accept prior to submitting a search request, expressly forbid screen scraping. Such a prohibition, however, impaired the users' rights pursuant to Art. 6 para. 1 of the directive and, as a result, was invalid according to Art. 15 of the directive. The CJEU decided in favour of the claimant, Ryanair Ltd., that their database generally did not fall under the scope of this directive within this context (and therefore is not subject to the prohibition under Art. 15) and that the claimant was therefore allowed to forbid its users by contract to carry out scraping operations.[21] To this extent, an infringement of the copyright or ancillary copyright on the part of the comparison platform did not exist; nevertheless, Ryanair could validly assert claims for injunctive relief based on the contractual provisions forbidding it. (Schapiro and Zdanowiecki 2015: 497(500)).

c) Interim Results

After the decisions of the BGH generally affirmed the permissibility of web scraping in Germany, the two decisions passed by the CJEU severely challenged this legal assessment. The highest court of the European Union gave the investment protection of the maker of the database much more weight than the German courts had. With regard to commercial scraping services, it will be interesting to see how the BGH responds to the latest CJEU decisions.

3. Limitation rule under sec. 60d UrhG

Within the context of academic research, which is relevant in this case, the German legislators were quicker to respond than the BGH. Under sec. 44a et seq. UrhG, the German copyright act declares that many use cases are permissible that would otherwise be protected under copyright law (Rehbinder and Peukert 2018: Sec. 1 marginal no. 17). In the event that the limitation rule applies, an infringement of copyrights including the consequences set out under sec. 97 et. seq. UrhG (right holder's entitlement to injunctive relief, damages, etc.) does not take place. With the Act to Align Copyright Law with the Current Demands of the Knowledge-based Society (UrhWissG)[22], which came into force on 01.03.2018, the legislators established new privileges for the use of copyrighted works for teaching, research, and institutions. This includes the newly stipulated limitation rule under sec. 60d UrhG that privileges text and data mining.

According to sec. 60d para. 1 UrhG, a large number of works is the basis and subject of an automated analysis. The term that is legally defined as "source material" in the act refers to a large and initially

18 *EuGH*, judgment of 19.12.2013 – C-202/12, Ls. – *Innoweb/Wegener* = MMR 2014, 185.
19 *EuGH*, judgment of 15.01.2015 – C-30/14 – *Ryanair* = MMR 2015, 189.
20 *Gerechtshof Amsterdam*, judgment of 13.03.2012 – 200.078.395, 4.13 = ECLI:NL:GHAMS:2012:BW0096.
21 *EuGH*, judgment of 15.01.2015 – C-30/14, marginal no. 39 – *Ryanair* = MMR 2015, 189.
22 Act to adjust the copyright law to the current requirements of a knowledge society of 01.09.2017, federal gazette. I 2017, p. 3346.

unstructured text and data volume comprising any type of protected contents (Hagemeier 2018: Sec. 60d marginal no. 6). This source material is then rendered machine-readable, for example, by normalising, structuring, and, if necessary, converting it (for example, PDF documents into XML data records).[23] The resulting corpus is the compilation to be analysed, for example, through an automated search for statistical frequencies or correlations.[24]

A basic requirement for the applicability of the limitation rule is that academic research is carried out exclusively for non-commercial purposes (sec. 60d para. 1 sentence 2 UrhG). This is the case if research is "not profit-driven" or "carried out in a public interest acknowledged by the state" (Hagemeier 2018: Sec. 60d marginal no. 8). On the contrary, if companies conduct research to develop and then market goods or services, this is deemed a commercial purpose.[25] The created corpus (but not the source material (ibid.: Sec. 60d marginal no. 16)) can continue to be provided to a limited group of people for common academic research and to third parties to review the quality of the academic research according to sec. 60d para. 1 sentence 1 no. 2 UrhG (for example, via a common intranet (Mörike 2018:3)).

The law does not specify whether any other forms of reproduction are allowed beyond this. Construed literally, the wording would apply solely to the provision for viewing so that the corpus could only be made available in data formats that excluded printing or saving (Hagemeier 2018: Sec. 60d marginal no. 15). The saving, printing, or sending by e-mail (Leupold and Demisch 2000: 379(385)) would, in fact, represent copying within the meaning of sec. 16 UrhG, which is not covered by the wording under sec. 60d para. 1 sentence 1 no. 2 UrhG (Hagemeier 2018: Sec. 60d marginal no. 15). This problem is reminiscent of a lawsuit about the provision of works via electronic reading stations in libraries in which the CJEU found that a reproduction within the meaning of the underlying directive did not allow the printing or storing on a data medium but could allow subsequent utilisation under the limitation rule.[26] Therefore, one author of jurisprudential literature considers the limitation rule under sec. 60c para. 1 and para. 3 UrhG to be proof that the corpus can be made available in a storable and printable data format (Hagemeier 2018: Sec. 60d marginal no. 15). However, because the wording of the law is imprecise in this regard, such an interpretation should be regarded with suspicion until the issue has been confirmed by the legislators or the courts.

Finally, the corpus and any reproductions of the collected material must be deleted after completion of the research activities pursuant to sec. 60d para. 3 UrhG. Submitting the corpus to a library or a comparable privileged institution within the meaning of sec. 60e and 60f UrhG is allowed, however. Besides libraries, the norms only list archives, museums, and educational institutions as privileged institutions. Whether research data centres (RDC) that are accredited by the RatSWD can qualify as privileged institutions within the meaning of the law must be determined by means of interpretation. Within the meaning of sec. 60e para. 1 UrhG, "libraries" are "publicly accessible libraries that have no direct or indirect commercial purpose".

Public accessibility must be understood within the meaning of the provisions of sec. 15 para. 3 UrhG (Dreier 2018: Sec. 61 marginal no. 16). Accordingly, the institution must act primarily in the service of society; this service is not ruled out by the fact that the use of a database is subject to certain conditions (e.g. membership) (Hagemeier 2018: Sec. 60e marginal no. 19). By contrast, the user group must not be restricted to a certain group of people – it is, however, allowed to limit access to a defined group of people (e.g. students of a university) (ibid.). RDCs are generally accessible for independent academic research. Therefore, there is no restriction to a certain group of people but rather to an intended use. Insofar, RDCs qualify as publicly accessible. Whether they are considered libraries within the meaning of sec. 60e UrhG or archives in accordance with sec. 60f UrhG is ultimately irrelevant for the result because sec. 60d para. 3 sentence 2 UrhG does not differentiate in this regard. Therefore, RDCs can – insofar as they do not serve any direct or indirect commercial purpose –

23 Official reasoning, Bundestag paper 18/12329, p. 40.
24 Bundestag paper 18/12329, p. 40.
25 Bundestag paper 18/12329, p. 39.
26 Cf. *CJEU*, judgment of 11.09.2014 – C-117/13 = MMR 2014, 822; preceding *BGH*, decision of 20.09.2012 – I ZR 69/11 – Electronic reading places I = MMR 2013, 529; following *BGH*, judgment of 16.04.2015 – I ZR 69/11 – Electronic reading places II = MMR 2015, 820.

qualify as privileged institutions on their merits, although individual assessments must be based on the concrete designs; general binding statements cannot be made due to the heterogeneity of the entities accredited by RatSWD. Moreover, it must be noted that due to the recentness of the provisions (effective date: 01.03.2018), there are no substantiating precedents and the interpretation herein represents an expert evaluation only.

The transfer of a corpus to a journal – if created using text and data mining techniques – would have to be verified according to the provisions of sec. 60d para. 3 sentence 2 UrhG and would likely not be permissible under the above-mentioned criteria because journals or their publishers do not represent publicly accessible institutions within the meaning as explained above.

By permitting transfers to privileged institutions, the provision intends to balance the interests of researchers and the interests of authors and publishers. After all, researchers must ensure the availability of the contents used for their research, for example, in order to verify whether they adhere to scientific standards. By contrast, it is mainly science publishers that are interested in preventing parallel databases of articles from emerging.[27] However, according to the reasoning given by the legislators, researchers are no longer allowed to store corpus or source material themselves.[28] In this respect, the law addresses the deletion obligation to the privileged parties under the limitation rule according to sec. 60d para. 1 UrhG. If there is more than one researcher (e.g. a consortium), it can be assumed that the head of the consortium is responsible for overseeing that all of the co-researchers delete the data accordingly. However, there are no concrete suggestions in this respect, neither in the law nor in its reasoning.

The ban on avoiding technical protection measures within the meaning of sec. 95a UrhG remains unaffected by the new limitation rule. If a website operator protects its contents against web scraping, for example through a robots.txt file[29], the scraper is not allowed to circumvent such protection measures (Mörike 2018: 3). If users were in fact entitled to assert claims against rightholders in accordance with sec. 95b para. 1 sentence 1 no. 11 UrhG that allowed them to exercise their rights under sec. 60d UrhG, this would apply only to works that are accessible offline and would not be relevant for the purpose of web scraping (Raue 2017: 656(658)).

In summary, it can be concluded that web scraping can be assumed to be permissible within the context of academic research from a copyright perspective, despite the CJEU decisions, insofar as the requirements under sec. 60d UrhG are met and the conditions thereunder are fulfilled (cf. also Mörike 2018: 2f.). Furthermore, it must be noted that sec. 60a through 60h UrhG only apply provisionally until 28.02.2023 and are not applicable after this date in accordance with sec. 142 para. 2 UrhG. This is due to the fact that the changes to this act are to be evaluated in 2022 after which a decision will be made as to whether they will continue to be valid or require modification.[30]

27 Bundestag paper 18/12329, p. 41.
28 Bundestag paper 18/12329, p. 41 et. seq.
29 Cf. regarding technical protective measures against web scraping, von Schönfeld 2018: 60 et. seq.
30 Bundestag papers 18/12329, p. 49.

III. General civil law

Along with implications under competition or copyright law, issues of general civil law are also frequently discussed, especially with respect to "virtual property rights" (*Hausrecht*).

1. Contract law

Website operators can set conditions for the use of their web presence through contracts and thus prohibit, for example, the automated retrieval of data by means of web scraping.[31] To obligate users to refrain from scraping activities, however, a valid contract must be concluded. A valid contract is concluded, for example, if a website can only be accessed after entering a login and user identification, following prior registration during which the user must consent to the terms of use. Simply visiting a website is not sufficient for the valid conclusion of a contract if such website only contains a simple note referencing the operator's terms of use (Deutsch 2009, 1027 (1028)).[32]

If the first is the case and a valid contract for use was concluded between the website operator and the user, the legitimacy of the provision that forbids scraping is subject to the provisions of sec. 305 et. seq. BGB [German Civil Code) regarding general terms and conditions (GTCs). According to this, the first criterion for the provision to be valid is that it qualifies as a GTC term or condition within the meaning as set out under sec. 305 para. 1 BGB, i.e., the contract terms were drafted for an indefinite number of contracts and unilaterally stipulated by one of the parties to the contract without any individual negotiations. This is usually the case with the terms of use for websites.

Furthermore, the clause must be validly incorporated into the contract. That means, the user of a website must be able to acknowledge the clause in an appropriate manner (sec. 305 para. 2 BGB). Links to terms of use that are hidden and cannot be easily found by website visitors do not meet this requirement so that such terms do not become part of the contract.

In a third step, the effectiveness of the substance of the clause is verified according to the provisions under sec. 307 et. seq. BGB. According to these, a contract term is invalid if it violates statutory provisions or unfairly disadvantages the other party to the contract. The concrete wording of a clause is decisive for evaluating the effectiveness of the substance so that general statements in this respect cannot be made (also Mörike 2018: 4).

Especially within the context of web scraping in academic research, another requirement must be taken into consideration: in addition to sec. 60d UrhG (see above), sec. 60g UrhG was introduced in conjunction with UrhWissG (the Act to Align Copyright Law with the Current Demands of the Knowledge-based Society); and para. 1 of sec. 60g UrhG provides that contract terms must not limit the uses allowed under sec. 60d UrhG. As a consequence, rightholders cannot rely on contract provisions that generally forbid web scraping if such web scraping falls under sec. 60d UrhG (Mörike 2018:4). However, this does not prevent them from taking protective technical measures that hinder or prevent web scraping. According to sec. 95b para. 3 UrhG, the obligation set out under sec. 95b para. 1 sentence 1 no. 11 UrhG to allow a text or data miner the use for privileged purposes as set out under sec. 60d UrhG does not apply to works made available online.

31 For example, *Ryanair's*, terms of use, available at https://www.ryanair.com/de/de/CorporateLinks/nutzungsbedingungen, No. 3 (last accessed: 20.08.2018).

32 *OLG Frankfurt/M.*, judgment of 05.03.2009 – 6 U 221/08.

2. Virtual property rights (*Hausrecht*)

Another obstacle under civil law could be virtual property rights (*Hausrecht*), the general existence and basic outline of which have been under discussion since the Bonn regional court[33] and the Cologne higher regional court[34] issued their rulings in 1999 and 2000, and the Munich higher regional court[35] in 2007 (for general information on the discussion, cf. Maume 2007: 620 (623 et. seq.)). Traditional "physical" property rights (*Hausrecht*) are generally derived from the ownership rules under Sec. 903, 1004 BGB [German Civil Code], and sec. 858 et. seq. BGB.[36] Virtual property rights (*Hausrecht*), for example, allow operators of a website to exclude users from accessing or using their website by means of technical measures.[37] A requirement for this is that a factual reason exists; an arbitrary ban on certain IP addresses is not allowed (Mörike 2018: 4).[38]

It could be considered an offence under the prohibition of arbitrary action if an excluded scraper behaved like a typical human user, for example, by submitting a single query to a specific database (von Schönfeld 2018: 341). Irrespective of whether virtual property rights (*Hausrecht*) are recognised in a direct or analogous application of the above provisions, the question arises as to what extent they are able to prevent potential web scraping. The virtual property rights (*Hausrecht*) was developed to exclude people who disrupt ordinary operations (for example, through repeated insults in online forums) from further use. The purpose of the virtual property rights (*Hausrecht*) therefore is the prevention of further disruptions. That means, the relevant question is whether the web scraping method represents "normal" behaviour or a disruption of operations (a valid argument ibid.: 341f.).

If an application programming interface (API) is used to retrieve data, disruptions of operations can be ruled out if the relevant conditions are fulfilled; after all, service providers usually offer APIs for this specific purpose. And even if there is no access to an API, automated data collection does not generally qualify as dishonest or abnormal user behaviour; such a treatment would unfairly disadvantage start-up companies operating in the area of innovative information services and thus hinder innovation (ibid.: 342). The boundary will have to be drawn where the mass retrieval of data severely strains or even overloads the server infrastructure and the proper operation of the website or database cannot be maintained – even if this is only temporarily the case (ibid.: 343).

In line with the opinions found in jurisprudential literature, it can be assumed that the concept of the virtual property rights (*Hausrecht*) within the context of web scraping is not considered to be that important and that web scraping does not conflict with these rights if used rationally (also Mörike 2018: 4; von Schönfeld 2018: 343).

33 *LG Bonn*, judgment of 16.11.1999 – 10 O 457/99 = MMR 2000, 109.
34 *OLG Köln*, decision of 25.08.2000 – 19 U 2/00 = MMR 2001, 52.
35 *OLG München*, judgment of 26.06.2007 – 18 U 2067/07 = MMR 2007, 659.
36 cf. instead of many *BGH*, judgment of 08.11.2005 – KZR 37/03, marginal no. 23 et. seq. – *radio broadcasting rights* = NJW 2006, 377.
37 *OLG Köln*, decision of 25.08.2000 – 19 U 2/00 = MMR 2001, 52; against the need for a virtual property rights (*Hausrecht*) Redeker 2007: 265 (266).
38 cf. *OLG Köln*, decision of 25.08.2000 – 19 U 2/00 = MMR 2001, 52.

C. Determining criteria for web scraping in research

Based on the above presentation of the legal position, we intend to develop some criteria for the use of web scraping technologies within the framework of independent academic research.

I. General requirements for the use of web scraping

From a legal perspective, it can be stated that a central criterion for the use of scraping procedures is that the information to be analysed must be accessible to the general public. The circumvention of technical protection measures that expressly serve to prevent scraping violates the discretionary rights of website or database operators to determine the recipients of their contents (von Schönfeld 2018: 356f.).[39] In this respect, "publicly- accessible" data not only means freely and directly accessible information but also information that can only be viewed after paying a fee or consideration (Schulze-Fielitz 2013: Art. 5 marginal no. 80; von Schönfeld 2018: 357). This criterion was determined by the highest courts in their judgments regarding the implications of web scraping under the copyright act and act against unfair competition, because circumventing protective technical measures represents an integrity-related offence[40] on the one hand and is prohibited by sec. 95a UrhG on the other.[41] The underlying concept is that anyone who benefits from the public accessibility of the internet must also generally accept such access to their own contents by others that use precisely the same public accessibility for their own profit (von Schönfeld 2018: 357).

As a result of the CJEU's newly opened discussion about the permissibility of scraping procedures under the copyright act, the legal position cannot be deemed conclusively settled until the German Federal Court of Justice (BGH) passes a judgement creating a legal precedent or the legislators establish general regulations regarding web scraping. The question of whether scrapers illegally infringe on database makers' rights as set out under sec. 87b UrhG was answered in the above court decisions, in each case differently depending on the specific circumstances. With regard to academic research, the newly created sec. 60d UrhG provides a certain degree of legal clarity and security by introducing a new limitation rule for text and data mining ensuring that – even if an infringement of database maker rights exists – such access to the data must be tolerated. However, researchers who use web scraping must ensure that all requirements set out under sec. 60d UrhG are met, including but not limited to the deletion of the corpus and source material after the research project is completed. Furthermore, the rightholder (e.g. the database maker) is entitled to payment of an appropriate remuneration in accordance with sec. 60h para. 1 UrhG (cf. Raue's critique 2017: 565(661f.)). Such a claim to remuneration in accordance with sec. 60h para. 4 UrhG can only be asserted by a collection society; rightholders are not authorised to collect claims on their own (Hagemeier 2018: Sec. 60h marginal no. 11). According to sec. 60h para. 5 sentence 1 UrhG, researchers conducting research for an institution do not owe the compensation themselves; compensation is owed solely by the institution. Such compensation can be paid in the form of a flat rate or based on use (representative samples), according to sec. 60h para. 3 sentence 1 UrhG, which will likely be left to the discretion of the party owing the compensation, that is the researcher or the institution. As a rule, however, the parties negotiating the compensation will usually agree to the amount and type of remuneration (Pflüger und Hinte 2018: 153(157). Based on the fact that the compensation is owed by law, the researcher must actively contact the rightholder and negotiate the modalities of such compensation, even if an express information obligation is not provided in the law.

39 *BVerfG*, judgment of 24.01.2001 – 1 BvR 2623/95 and 622/99 – Filming for television in the court room II = BVerfGE 103, 44 = ZUM 2001, 220 (224f.).

40 cf. BGH only, judgment of 30.04.2014 – I ZR 224/12 – On-line flight agency = MMR 2014, 740; BGH, judgment of 22.06.2011 – I ZR 159/10 – Online market place for automobiles = NJW 2011, 3443.

41 For this, cf. above B. II. 3.

From a legal perspective, it is necessary to ensure that the use of scraping technology does not cause any technical damage to the website or database operator (cf. also von Schönfeld 2018: 358). A concrete limit or threshold for the retrieval of data that would be considered damaging cannot however be determined. A decisive criterion in this regard would ultimately be the computing capacity of the host server and the technical design of the scraping software. If, however, the functionality of the host server is impaired (even if only temporarily) due to excessive query activities, the use can be considered damaging and operators can respond by excluding the accessing IP address/es, thereby relying on their virtual property rights (*Hausrecht*).[42]

In summary, the following criteria must be taken into consideration in this regard:

- The information to be analysed must be publicly accessible. "Publicly accessible" can also mean that data can only be retrieved after payment of a consideration.
- The circumvention of protective technical measures aimed at preventing web scraping infringes on the right of beneficiaries to choose the addressees of their contents.
- Sec. 60d UrhG provides clarity with respect to the permissibility of web scraping for the purpose of academic research. However. the requirements made therein must be fulfilled, for example:
  - The academic research must be for non-commercial purposes only.
  - After completion of the research activities, the created corpus must be deleted; transfer to privileged institutions (e.g. library) is allowed, however.
  - The rightholder is entitled to payment of an appropriate remuneration.
- The use of scraping technologies must not cause any technical damage to the operator's website or database.

II. Conditions for archiving and long-term provision of the collected data

1. Copyright requirements

With respect to copyright, the corpus and any copies of the source material must be deleted upon completion of the research project as described above (sec. 60d para. 3 sentence 1 UrhG). The law does not explicitly specify the type and scope of erasure, but means irretrievable deletion, for example by destroying the storage media or deleting the digital datasets (Hagemeier 2018: Sec. 60d marginal no. 19). The development of a deletion plan is recommended. This plan should, for example, include a description of how to comply with the statutory requirements and the technical and organisational requirements to protect researchers from any inadmissible re-utilisation of the data (by third parties or themselves) and to document compliance with the deletion obligation (ibid.). It is difficult to determine the point of time at which the research project is completed within the meaning of the law. For this purpose, the completion of the research report alone cannot be deemed to mark the end of the project because according to sec. 60d para. 1 sentence 1 no. 2 UrhG, quality control measures (e.g. peer reviews) are also expressly allowed (Dreier 2018: Sec. 60d marginal no. 12).

According to sec. 60d para. 3 sentence 2 UrhG, the corpus and any copies of the source material can be transferred to libraries, archives, museums, and educational institutions (privileged institutions within the meaning of sec. 60e and 60f UrhG) to ensure verification of compliance with academic standards and whether the research can be used for citations or as references in the long term.[43] The purpose and intent of the provision is to ensure that the archiving institutions can make the received material accessible to other researchers for text and data mining as well; storage of the data would otherwise be meaningless (Raue 2017: 656(661)).

Any electronic transfer from an archiving library to other researchers qualifies as a communication to the public within the meaning of sec. 15 para. 2 UrhG that does not generally fall under the limitation rule as provided under sec. 60a et. seq. UrhG (ibid.). A relevant authorisation must result from the provision set out under sec. 60d para. 3 sentence 2 UrhG to properly comply with the

42 See B. III. 2.
43 Bundestag papers 18/12329, p. 41.

legislator's intent (ibid.). This also does not contradict the copyright directive 2001/29/EC[44] (the "InfoSoc" directive) on which sec. 60d UrhG is based and which allows communication to the public for the purpose of academic research in Art. 5 para. 3 letter a (Raue 2017: 656 (661)). As a precaution, the archiving institutions should restrict access to such persons that fulfil the requirements set out under sec. 60d para. 1 UrhG (recommended by Raue 2017: 656(661)).

2. Data protection requirements

Problems under the data protection act arise where the scraped material contains personal data. According to Art. 4 no. 1 GDPR, personal data "means any information relating to an identified or identifiable natural person ("data subject")". This is particularly relevant if the data for analysis originate from social media. Due to the principle of "prohibited unless authorised" in data privacy law, every processing of personal data is initially forbidden, unless a legal norm allows their processing, or the data subject consents to the processing of the data (Art. 6 para. 1 GDPR (Ingold 2017: Art. 7 marginal no. 8f). If the latter is not the case, Art. 6 para. 1 letter f GDPR can be considered a norm that is suitable for allowing web scraping because it permits data processing where – after considering the conflicting interests – the data processing researcher's interests override the data protection interests of the data subject. This must be individually decided in each case because there are no general criteria that can be applied in every case.

Especially within the context of academic research, sec. 27 Federal Data Protection Act (BDSG), as amended on 30.06.2017, establishes a legal basis for the processing of data for academic research with respect to special categories of personal data within the meaning of Art. 9 para. 1 GDPR. If the requirements are met (weighing of interests and taking measures to protect the interests of data subjects), health and genetic data, in particular, can be processed without consent contrary to the express prohibition as set out under Art. 9 para. 1 GDPR.

In addition, sec. 28 BDSG, as amended on 30.06.2017, provides a special legal basis for processing data for archiving in the public interest according to which the rights of data subjects are restricted with the intent of preventing archives from becoming meaningless (Pauly 2018: Sec. 28 BDSG marginal no. 2). Similar to sec. 27 BDSG, as amended on 30.06.2017, this provision applies exclusively to special categories of personal data within the meaning of Art. 9 para. 1 GDPR, for example health data. According to recital no. 158 GDPR, the object of archiving includes but is not limited to the acquisition, preservation, and provision of access to "records of enduring value for general public interest". In any case, however, it must be considered whether the objects of archiving cannot be attained using anonymised or at least pseudonymised data as well (Art. 89 para. 1 sentence 4 GDPR) (Pauly 2018: Sec. 28 BDSG marginal no. 6).

III. Excursus: Practical example

As a practical example, we would like to discuss several legal problems regarding web scraping and social media websites (particularly Twitter).

1. Binding effect of the Twitter API terms of use

If researchers intend to use the Twitter API to automatically retrieve and analyse tweets, they must first consent to the Twitter terms of use[45]. As explained above (B. III. 1.), a licence agreement is not concluded simply by accessing a website.

Due to the requirement of express consent, a licence agreement is concluded between the user and Twitter that binds the former by contract to comply with the Developer Agreement and Policy. The Developer Policy, inter alia, contains an obligation to respect the control and privacy of users with respect to their contents, including the obligation to destroy copies of deleted tweets on short

44 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, Official Journal 2001 No. L 167, p. 10 et. seq.

45 https://developer.twitter.com/en/developer-terms/agreement-and-policy (as of 24.08.2018).

notice.[46] This is unfavourable for researchers: conditions for grants often include terms that obligate researchers to archive their research data for several years for replication purposes. These policies represent general terms and conditions (GTCs) the validity of which is subject to sec. 305 et. seq. BGB (see above B. III. 1.). Since users must expressly agree to the terms, they are validly incorporated into the licence agreement (sec. 305 para. 2 BGB). A summary examination, however, failed to identify an unfair disadvantage for users because of the obligation to erase copies of deleted tweets. Therefore, one can assume the terms to be effective.

However, the terms of Twitter's licence agreement are binding for the contractual party if the API is used. If researchers prefer not to be bound by these terms, their only option is to terminate the previously concluded licence agreement – as a precaution – if one exists, and to scrape the data without using the programming interface. Limitations of use that are stipulated in the Twitter API also only bind the parties to the contract. Users who have not subjected themselves to these terms by entering into a contract by private act will not be bound by them. However, claims in tort, for injunctive relief or damages, may arise if data scraping causes an impairment of the server functionality or if operations are otherwise disrupted.

2. Governing law in the event that foreign countries are involved

If researchers working in Germany intend to use scraping methods for US websites in order to examine social groups in third countries, the question arises as to which legal system is relevant for them. As a rule, the researchers in the first instance are subject to German jurisdiction because they reside within German territory. Copyright law poses some difficulties: according to the principle of *lex loci protectionis*, the laws of the country for whose territory protection is sought are applicable to matters of intellectual property.[47] However, considerable uncertainties arise with respect to the question of the applicable law if actions take place on the internet because cross-border data transfers, in particular, affect different legal systems and allocating an action to the territory of a certain state is therefore complicated (as demonstrated by Bollacher 2005: 101 et. seq.). According to the principle of *lex loci protectionis*, the above-mentioned matter could be subject to US copyright law if the servers, on which the queried databases are stored, are located within the territory of the US. As a consequence, copyright holders under US law could assert claims against scrapers within that legal system if scrapers infringed on copyrights or proprietary rights in the course of their activities. A prerequisite for this would be, however, that the place where such infringement of copyright takes place is deemed to be in the US. Whether in this specific use case of web scraping the place of action is considered to be the location of the server (here: USA) or the location of the scraper (here: Germany), has, to the best of our knowledge, yet to be examined.

Moreover, other legal systems can become relevant if the researcher becomes a party to a contract that includes a clause pertaining to governing law. Such a provision would then only apply to such disputes or circumstances that arise directly out of the respective contract. Otherwise, German law remains relevant and applicable (as described in detail under B. II. 3. And B. III.).

46 Twitter Developer Policy, I. Guiding Principles, C. Respect Users' Control and Privacy 3., available at https://developer.twitter.com/en/developer-terms/agreement-and-policy (as of 24.08.2018).

47 As provided in Art. 5 para. 2 p. 2 of the Berne Convention for the Protection of Literary and Artistic Works seized in Art. 9 of the TRIPS Agreement (Agreement on Trade-Related Aspects of Intellectual Property Rights) and in Art. 8 para. 1 of the Rome II Regulation; cf. Rehbinder and Peukert 2018: Sec. 49 marginal no. 1206.

3. Data protection implications

a) Public interest in the meaning of Art. 89 GDPR

Besides academic and historic research, Art. 89 para. 1 GDPR also privileges archiving that is "in the public interest". According to recital no. 158 sentence 2, GDPR understands archives to mean "public authorities or public or private bodies that hold records of public interest". However, this only includes archiving purposes that are in the public interest. As examples for such, recital no. 158 sentence 4 lists "the provision of specific information related to the political behaviour under former totalitarian state regimes, genocide, crimes against humanity, in particular the Holocaust, or war crimes". A decisive factor is that the archiving is in the interest of society as a whole and not only in the interest of the processor (Schantz 2017: marginal no. 1347).

At the same time, Art. 89 para. 2 and 3 GDPR allow the introduction of statutory exemptions from the rights of data subjects as set out under Art. 12 et. seq. GDPR. Inter alia, even the right to erasure ("right to be forgotten") according to Art. 17 GDPR is overridden if it is necessary to store data for a longer period to protect the subject of the research project (Hense 2017: Art. 89 marginal no. 12). This does not result from Art. 89 GDPR but instead directly from the exceptions to the right to erasure under Art. 17 para. 3 letter d and Art. 5 para. 1 letter e GDPR (cf. Schantz 2017: marginal no. 1361).

b) Personal attribution and pseudonymisation

Ultimately, the question arises as to when data are considered attributable to individuals – despite mentioning aliases only (pseudonymisation) – and thus fall within the scope of the data privacy laws according to Art. 2 para. 1 GDPR. Art. 4 no. 5 GDPR defines pseudonymisation as the "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person." In contrast to fully anonymised data that cannot (or can no longer) be attributed to individuals and therefore do not fall within the scope of the GDPR, pseudonymised data potentially continue to be personal data and, therefore, are in fact entirely subject to the GDPR (cf. recital no. 26 sentence 2) (Schantz 2017: marginal no. 303).

In jurisprudential literature, some authors would limit this absolute applicability, in the event that the probability of identifying the person behind the alias is so low that this risk is practically negligible (Roßnagel 2018: 243 (244)). Persons can only be considered identifiable if they can in fact be identified.[48] The probability of an attribution is determined by weighing effort against benefit. For this, recital 26 sentence 4 stipulates that "the costs of and the amount of time required for identification" and "the available technology at the time of the processing and technological developments" must be taken into consideration (cf. Roßnagel 2018: 243(244)). After all, this is a question that must be answered on a case-by-case basis and general criteria cannot be determined. If researchers are by all means unable to attribute pseudonyms to specific natural persons, they are not deemed to be processing personal data and as a result need not comply with the requirements under the GDPR and BDSG for their processing.

48 *EuGH*, judgment of 19.10.2016 – C-582/14, marginal no. 46 – *Breyer* = NJW 2016, 3579.

D. Conclusion and prospects

In conclusion, it can be stated that a legal evaluation of web scraping requires significant analysis and discussion, especially with respect to copyright law. After the highest courts, the German Federal Court of Justice and the European Court of Justice, could not find a common ground, and after the subsequent European courts were not willing to issue generally accepted statements with regard to the permissibility of web scraping, the legal position was far from clear. With the UrhWissG and the newly established provision under sec. 60d UrhG, the legislators took countermeasures and provided at least a certain degree of legal clarity in the area of academic research that is the relevant area for this expert opinion.

It was a conscious decision of the legislators to act ahead of the European Union. The EU is currently working on a draft of a directive on copyright in the digital single market.[49] The new provisions under sec. 60d UrhG are based on Art. 3 of the draft of the directive. Due to the directive, the German legislators felt obligated to limit the use of text and data mining to non-commercial academic research; extending this to include researching businesses is not expected due to the European requirements (Hagemeier 2018: Sec. 60d marginal no. 4). In their reasoning for the UrhWissG, the legislators have already announced they would amend the provisions in line with the directive, if necessary, as soon as it comes into effect.[50]

49 Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016) 593 final, available at https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52016PC0593 (24.08.2018).

50 Bundestag papers 18/12329, p. 4

References

Bitkom (2013): Trends im E-Commerce. Konsumverhalten beim Online-Shopping. Berlin, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM). https://www.bitkom.org/Publikationen/2013/Studien/Trends-im-E-Commerce/BITKOM_E-Commerce_Studienbericht.pdf (last accessed: 16.08.2018).

Bitkom (2018): Markt für Big Data wächst in Deutschland zweistellig. Presseinformation vom 13.03.2018. Berlin, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. (BITKOM). https://www.bitkom.org/Presse/Presseinformation/Markt-fuer-Big-Data-waechst-in-Deutschland-zweistellig.html (last accessed: 16.08.2018).

Bollacher, Philipp D. (2005): Internationales Privatrecht, Urheberrecht und Internet. Frankfurt am Main, Verlag Peter Lang.

Deutsch, Askan (2009): Die Zulässigkeit des so genannten „Screen-Scraping" im Bereich der Online-Flugvermittler. GRUR (Gewerblicher Rechtsschutz und Urheberrecht) 2009(11), 1027–1032.

Dreier, Thomas (2018): Kommentierung des § 60d UrhG sowie der Vorbemerkung zu §§ 87a ff. UrhG. In: Thomas Dreier and Gernot Schulze (eds.): Urheberrechtsgesetz. Kommentar. 6th ed., Munich, C.H.BECK.

Hagemeier, Stefanie (2018): Kommentierung der §§ 60d, 60f, 60h UrhG. In: Hartwig Ahlberg and Horst-Peter Götting (eds.): Beck'scher Online-Kommentar zum Urheberrecht. 21th ed., (as of 04.06.2018), Munich, C.H.BECK.

Hense, Ansgar (2017): Kommentierung des Art. 89 DSGVO. In: Gernot Sydow (ed.): Europäische Datenschutzgrundverordnung. Handkommentar. Baden-Baden, Nomos.

Ingold, Albert (2017): Kommentierung des Art. 7 DSGVO. In: Gernot Sydow (ed.): Europäische Datenschutzgrundverordnung. Handkommentar. Baden-Baden, Nomos.

Kinne, Jan and Janna Axenbeck (2018), Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany, ZEW Discussion Paper No. 18-033, Mannheim.

Kotthoff, Jost (2013): Kommentierung der §§ 4 und 87a UrhG. In: Gunda Dreyer, Jost Kotthoff and Astrid Meckel: Urheberrecht. Kommentar. 3rd ed, Heidelberg and others, C.F. Müller.

Kukulenz, Dirk (2008): Die Dynamik des World Wide Web und Konsequenzen für die Informationssuche. Habilitationsschrift. https://www.ifis.uni-luebeck.de/fileadmin/user_files/veroeffentlichungen/Habil-Kukulenz08.pdf (last accessed: 16.08.2018).

Leupold, Andreas and Dominik Demisch (2000): Bereithalten von Musikwerken zum Abruf in digitalen Netzen. ZUM (Zeitschrift für Urheber- und Medienrecht) 2000(5), 379–390.

Loewenheim, Ulrich (2017): Kommentierung des § 4 UrhG. In: Gerhard Schricker and Ulrich Loewenheim (eds.): Urheberrecht. Kommentar. 5th ed, Munich, C.H.BECK.

Marquardt, Malte (2014): Kommentierung des § 4 UrhG. In: Artur-Axel Wandtke and Winfried Bullinger (eds.): Praxiskommentar zum Urheberrecht. 4th ed, Munich, C.H.BECK.

Maume, Philipp (2007): Bestehen und Grenzen des virtuellen Hausrechts. MMR (MultiMedia und Recht) 2007(10), 620–625.

Mörike, Matthias (2018): Im Namen der Wissenschaft! Zur Zulässigkeit von Screen Scraping im Forschungsbetrieb vor dem Hintergrund des neuen Urheberrechts. DFN-Infobrief 6/2018. https://www.dfn.de/fileadmin/3Beratung/Recht/1infobriefearchiv/2018/Infobrief_Recht_06-2018.pdf (last accessed: 16.08.2018).

Nordemann, Axel (2014): Einleitung zum Urheberrechtsgesetz. In: Friedrich Karl Fromm und Wilhelm Nordemann (eds.): Urheberrecht. Kommentar zum Urheberrechtsgesetz, Verlagsgesetz, Urheberrechtswahrnehmungsgesetz. 11th ed., Stuttgart, Kohlhammer.

Pauly, Daniel A. (2018): Kommentierung des § 28 BDSG. In: Boris P. Paal and Daniel A. Pauly (eds.): Datenschutz-Grundverordnung und Bundesdatenschutzgesetz. Kommentar. 2nd ed., Munich, C.H. BECK.

Pflüger, Thomas and Oliver Hinte (2018): Das Urheberrechts-Wissensgesellschafts-Gesetz aus Sicht von Hochschulen und Bibliotheken. ZUM (Zeitschrift für Urheber- und Medienrecht) 2018(3), 153–161.

Raue, Benjamin (2017): Text und Data Mining. Die neue Urheberrechtsschranke des § 60d UrhG. CR (Computer und Recht) 2017(10), 656–662.

Redeker, Helmut (2007): Anmerkung zu LG München I, Urteil v. 25.10.2006 – 30 O 11973/05. CR (Computer und Recht) 2007(4), 265–267.

Rehbinder, Manfred and Alexander Peukert (2018): Urheberrecht und verwandte Schutzrechte. 18th ed., Munich, C.H. BECK.

Roßnagel, Alexander (2018): Pseudonymisierung personenbezogener Daten. Ein zentrales Instrument im Datenschutz nach der DS-GVO. ZD (Zeitschrift für Datenschutz) 2018(6), 243–247.

Schack, Haimo (2001): Urheberrechtliche Gestaltung von Webseiten unter Einsatz von Links und Frames. MMR (MultiMedia und Recht) 2001(1), 9–17.

Schantz, Peter and Heinrich Amadeus Wolff (2017): Das neue Datenschutzrecht. Datenschutz-Grundverordnung und Bundesdatenschutzgesetz in der Praxis. Munich, C.H.BECK.

Schapiro, Leo and Konrad Żdanowiecki (2015): Screen Scraping. Rechtlicher Status quo in Zeiten von Big Data. MMR (MultiMedia und Recht) 2015(8), 497–501.

Schulze-Fielitz, Helmuth (2013): Kommentierung des Art. 5 GG. In: Horst Dreier (ed.): Grundgesetz Kommentar, Band I: Präambel, Artikel 1-19. 3rd ed., Tübingen, Mohr Siebeck.

Seagate (n.d.): Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2016 und 2025 (in Zettabyte). Statista - Das Statistik-Portal. https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/ (last accessed: 16.08.2018).

Thum, Dorothee and Kai Hermes (2014): Kommentierung des § 87a UrhG. In: Artur-Axel Wandtke and Winfried Bullinger (eds.): Praxiskommentar zum Urheberrecht. 4th ed., Munich, C.H. BECK.

Vogel, Martin (2017): Kommentierung des § 87a UrhG. In: Gerhard Schricker and Ulrich Loewenheim (eds.): Urheberrecht. Kommentar. 5th ed., Munich, C.H. BECK.

Von Schönfeld, Max (2018): Screen Scraping und Informationsfreiheit. Baden-Baden, Nomos.

Imprint

Publisher:
German Data Forum (Rat für Sozial- und Wirtschaftsdaten, RatSWD)
Rungestr. 9
10179 Berlin
office@ratswd.de
https://www.ratswd.de

Editors:
Thomas Runge, Dr. Nora Dörrenbächer, Dr. Mathias Bug, Lea Salathé

Layout and design:
Claudia Kreutz

Translation:
Regina Seelos, seelos-sprachendienste.de

Icons:
made by Freepik from www.flaticon.com
Font Awesome, fontawesome.com (modified)

Berlin, July 2020

RatSWD Output:
The RatSWD Output Series documents the German Data Forum's (RatSWD) activities during its 6th appointment period (2017–2020). It serves to publish its statements and recommendations and to make them available to a broad audience.

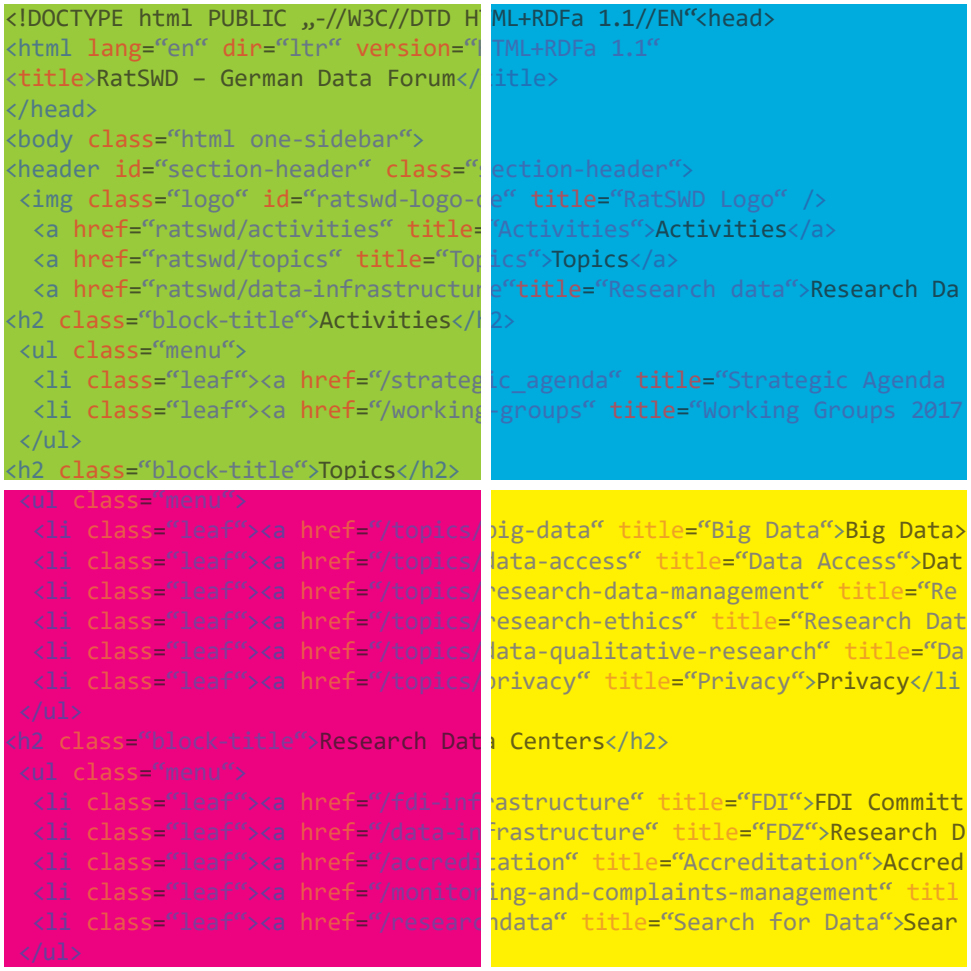
This report is the result of a project that is funded by the Federal Ministry for Education and Research (reference number: 01UW1802). Unless otherwise stated, the responsibility for this publication lies with the German Data Forum (RatSWD).

doi: 10.17620/02671.52

Suggested citation:
RatSWD [German Data Forum] (2020): Big data in social, behavioural, and economic sciences: Data access and research data management. RatSWD Output 4 (6). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.52.

Established in 2004, the **German Data Forum** (Rat für Sozial- und Wirtschaftsdaten, RatSWD) is an independent council. It advises the German federal government and the federal states (Länder) in matters concerning the research data infrastructure for the empirical social, behavioural, and economic sciences. The German Data Forum (RatSWD) has 16 members. Membership consists of eight elected representatives of the social, behavioural, and economic sciences and eight appointed representatives of Germany's most important data producers.

The German Data Forum (RatSWD) offers a forum for dialogue between researchers and data producers, who jointly issue recommendations and position papers. The council furthers the development of a research infrastructure that provides researchers with flexible and secure access to a broad range of data. The German Data Forum (RatSWD) has accredited 38 research data centres (as of May 2020) and fosters their interaction and collaboration.



www.ratswd.de