International Journal of Law and Information Technology, 2017, 25, 26–37
doi: 10.1093/ijlit/eaw014
Advance Access Publication Date: 8 January 2017
Article

OXFORD

Big data and the legal framework for data quality

Thomas Hoeren*

ABSTRACT

Power has a lot to do with knowledge, access to, and utilization of data. But in the context of the debate about power, the question of data quality is hardly ever raised. This is because legal standards for data quality are lacking. The first attempts to regulate this question can be found hidden in Article 6 of the EU Data Protection Directive and in the regulation on scoring in section 28b of the German Federal Data Protection Act (BDSG). From this, with the help of initial research attempts by computer science and sociology, we can develop a provisional, fragmentary framework for legal standards in data quality, as I will demonstrate in the following 10 theses.

KEYWORDS: big data, data quality, data protection, EU Regulation, tort law

THE SILENCE OF THE LAMBS: WHY DOES RESEARCH HAVE NOTHING TO SAY ABOUT DATA QUALITY?

Data are the backbone of power. Only someone who knows something and has access to data can control, plan and effect changes. Data are often interpreted as the currency of the digital economy, not without reason. So, it is all the more astonishing that until now there has been hardly any debate about the protection of data quality within the discussion of power and powerlessness.[1] What remains of an organisation's power such as Google when spectacular big data cases such as their Google Flu Trends turned out to be ex post false?[2]

This ignorance is still promoted by articles in the daily press that extol the sloppiness of data research as an actual asset in big data, for example, as here in the *Süddeutsche Zeitung*.

Large amounts of data, dirty data, indicate a trend but do not provide an exact result—in just about all of this, big data methods contradicts the way in which

* Professor, ITM, Germany. Email: hoeren@uni-muenster.de

1 For example, X Meng and X Ci, 'Big Data Management: Concepts, Techniques and Challenges' (2013) 50 J Comp Res Devel 146. But the debate is different, however, in certain areas such as aeronautical data where the data quality is regulated and standardized extensively. See Annex IV ('Data quality requirements') of EU Commission Reg No 73/2010 of 26 January 2010 for the qualitative requirements in aeronautical data and aeronautical information for the entirety of European airspace [2010] OJ L23/6 <http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32010R0073&from=DE> accessed 17 August 2016.

2 D Butler, 'When Google Got Flu Wrong' (2013) 494 Nature 155. Equally shocking in this respect is Sharona Hoffman's empirical study on the deficiencies in data quality in the medical world. S Hoffman, 'Medical Big Data and Big Data Quality Problems' (2014) 21 Conn Ins L J 289.

statisticians have worked up to now. But if large amounts of data are processed, it is this sheer mass whose analysis ultimately brings one very close to one's goal.[3] Especially 'big data' changes the research structure in science- and evidence-based decision-making from causation models to correlation paradigms. Data-based research traditionally proceeded from a hypothesis, which was used to understand relations between data and causation. Big data changes this concept by obviating the need for a hypothesis. Instead understanding is gained and knowledge is derived from data patterns. This new data mining technique leads to unknown epistemological consequences for data quality criteria including legal requirements for data quality in civil law or data protection law.

Data quality requirements arise and must arise in at least three situations. The first is the question of buying raw data: the buyer's issue essentially is the protection of the contractually stipulated quality of such data. The second concerns the protection of those who acquire the results of big data research. And finally, it is always about the rights of those who are affected by the assessment results in whatever way outside of contractual relationships, but within the legal protection regime against unfair discrimination.

THE LAW ON WARRANTY IN MODERN CODIFIED CIVIL LAW IS OUTDATED

The contractual rules on the protection of data quality are obsolete.[4] They derive from 19th-century commodity-oriented economic structures and safeguard at best exceptionally a liability in contractual or quasi-contractual relationships. Accordingly, the few published opinions on data quality in big data essentially only discuss the liability for transmission errors.[5]

The real test on the subject of information liability in the information society is from now on data themselves are being made the subject of contracts. Traditionally, in what was then the only conceivable case of selling information in book form, the law proceeded on the basis that the contractually agreed use was hard to determine.[6] In law, the buyer/reader of a book entertained no expectations of a book's content that were worthy of protection;[7] such expectations were as a rule only irrelevant desires for information.[8] Boundaries were only overstepped if a larger-than-average number of printing errors were present, pages were missing or a statute book was

3 H Martin-Jung, 'Warum wir Big Data verstehen müssen' *SZ* (10 October 2015).

4 The following ideas are based on the premises of German civil law. However, the legal position in other EU Member States is no better.

5 See C Peschel and S Rockstroh, 'Big Data in der Industrie - Chancen und Risiken neuer datenbasierter Dienste' (Vol 9 2014) MMR 571.

6 German Federal Supreme Court (Vol 41 1988) NJW 2597ff; also J Wertenbruch, 'Gewährleistung beim Kauf von Kunstgegenständen nach neuem Schuldrecht' (Vol 28 2004) NJW 1977, 1979ff; H Haberstumpf, 'Verkauf immaterieller Güter' (Vol 22 2015) NJOZ 793, 796 maintains that only tangible property can be the starting point for a product purchase.

7 HP Westermann, *Säcker/Rixecker/Oetker/Limperg Münchener Kommentar zum Bürgerlichen Gesetzbuch* (7th edn, Beck 2015) para 73.

8 German Federal Supreme Court (1958) NJW 138ff.

completely obsolete.[9] Alternatively, one worked with assurances and guarantees[10] or an independent consultancy agreement.[11] Otherwise, there was the danger that the usually expected reliability of the facts would lead to a warranty irrespective of which party was at fault.[12]

The background to this restrictive attitude can be found in Article 5 paragraph 3 clause 1 of the *Grundgesetz* (GG) [German Constitution]. It privileges both the author of the book and its publisher.[13] From this, the German Federal Court of Justice concludes printing errors 'can indeed be largely avoided by a customary and commercially generally acceptable method of production, although not with certainty. In individual cases, therefore, it may be that trade and communication does not and may not rely on the absence of a single such error.'[14] Even if we recognize such privileging, this is not applied to data providers in the age of big data. At least, since the Law of Obligations reform, data are 'other objects', according to section 453 paragraph 1 (second alternative) Bürgerliches Gesetzbuch BGB [German Civil Code], with the result that the normal rules on the sale of goods (sections 433ff BGB) are correspondingly applicable to data.[15] Under traditional German law, the sales law provisions in sections 433ff BGB only included the purchase of movable and immovable physical objects according to the conception of the legislature. The purchase of rights and other immaterial objects was not covered by these provisions. However, with the last law reform, the sales law is applicable *mutatis mutandis* to these objects based on the referencing norm of section 453 paragraph 1 BGB. In the same way, German law leaves the path to apply the law of service contracts to data contracts. Someone who is selling data is more likely to be sued under German sales of goods law than according to the regulations on the law of services.

THE LAW OF TORTS IN THE BGB AND OTHERS IS ALSO WORTHLESS

As we can clearly see in the example of the German Civil Code, the rules of tort liability too are obsolete. German tort law is based on the assumption that mere pecuniary loss is not enough to justify claims, but that a violation of absolute rights (property, health or another similar right) is needed. Thus the only provision is

9 Local Court Stuttgart (Vol 9 1995) NJW-RR 565ff; see also F Faust, *Bamberger/Roth Kommentar zum bürgerlichen Gesetzbuch* (39th edn, Beck 2016) para 70; U Foerste, 'Die Produkthaftung für Druckwerke' (Vol 23 1991) NJW 1433, 1436; U Huber, *Soergel Bürgerliches Gesetzbuch mit Einführungsgesetz und Nebengesetzen: BGB* (12th edn, Kolhammer 1991) s 459, para 344.

10 German Federal Supreme Court (1973) NJW 843–46; criticized in Huber, ibid, para 19.

11 German Federal Supreme Court (1978) BGHZ 70, 356–65, Vol 20 NJW 997ff; cf for more details J Köndgen, 'Die Haftung von Börseninformationsdiensten' (Vol 24 1978) JZ 389; C von Hertzberg, 'Die Haftung von Börseninformationsdiensten' (1978) *Fachmedien Recht und Wirtschaft in Deutscher Fachverlag GmbH.*

12 Westermann (n 7) para 3.

13 See criticism in Foerste (n 9) 1433ff.

14 German Federal Supreme Court (1970) NJW 1963ff.

15 RegE, BT-Drs 14/6040 24; J Jickeli and M Stieper, *Staudinger BGB* (rev edn, Sellier/de Gruyter 2012) s 90, para 17; RM Beckmann, *Staudinger BGB* (rev edn, Sellier/de Gruyter 2014) s 453, para 37. Also Case No 17 U 167/09, 2010 BeckRS 09514, Düsseldorf Higher Regional Court, Judgement of 17 February 2010; Case No 16 HK O 10382/08, 2009 BeckRS 88429, Munich Regional Court I, Judgement of 10 December 2008.

protection against a worst-case scenario in information law, the complete loss of data through the construct of a loss of property (section 823 paragraph 1 BGB).[16] This construction is made possible because the loss of property under section 823 paragraph 1 BGB does not require damage to an object itself, but any negative influence on the owners' wish to use its property as he sees fit. Deleting data from a data carrier derives the owner of the respective carrier of this right. Yet, data as such cannot be seen as 'another right' according to section 823 paragraph 1 BGB due to the circumstances that it is not characterized as an absolute right, a right that applies to everyone, and provides the owner with the authority to use it as he sees fit. Only those are protected.

For the same reason, tort liability for negligent misstatements exists only if one of the legally, through section 823 paragraph 1 BGB protected rights (life, body, health, freedom, property in the understanding of German law (i.e. physical objects) or an absolute right in the above-mentioned sense) is infringed. Under section 823 paragraph1 BGB, there is no claim for the loss of assets beyond this, especially financial losses due to trusting negligently made false statements.

Section 824 paragraph 1 BGB on the other side protects against endangerments to the credit of a person or a company and entitles the claimant to damages in the amount of the incurred financial losses. Claims solely for financial losses can result out of section 824 paragraph 1 BGB on the condition that these endangerments result out of factual claims rather than value judgements. However, claims for financial losses as a result of reliance on advice can usually not arise in the big data sector. These are the dire consequences of the fact that the assessment of raw data, for example, in the case of scoring, is seen as the creation and communication of value judgments, even in the opinion of the Federal Court of Justice. Thus section 824 paragraph 1 BGB requires that untruthful statements are being disseminated, not just value judgments, according to the Court.[17] By contrast, the Court said, section 824 paragraph 1 BGB offers no protection against pejorative expressions of opinion and value judgments. An exception would only come into force, according to the Court, 'if during the expression, in the recipient's view the elements of the opinion fade into the background in the face of the underlying facts' (paragraph 11). For instance, section 824 offers a protection against a scoring result denying a financial credit to a company, but not against the wrong factual basis of this decision if it is derived from big data.

Equally, the Court said, the law concerning the right to carry on an established business cannot be of further help. For in the necessary weighing of interests in the context of section 823 paragraph 1 BGB, it should be noted that Article 5 paragraph 1 GG 'does not prohibit the dissemination of true and objective information on the market, which can be important for the competitive behavior of market participants, even if the content has an adverse affect on individual competitive positions'.[18]

16 Case No 3 U 15/95, 1996 CR 32, Karlsruhe Higher Regional Court, Judgment of 7 November 1995.

17 Case No VI ZR 120/10, Vol. 30 2011 NJW 2204, German Federal Supreme Court, Judgment of 22 February 2011. See criticism in T Weichert, 'Scoring in Zeiten von Big Data' (Vol 6 2014) ZRP 168, 170ff.

18 Judgment of 22 February 2011, ibid.

These antiquated guidelines appear not only in Germany, but also, for example, in US law, as a reading of the famous *Winter v GP Putnam's Sons* of 1991 demonstrates.[19] In this case, two mushroom enthusiasts sued the publisher, who had published the British book *The Encyclopedia of Mushrooms* in the USA, for damages, for damages. The book was a work of reference on the subject of collecting and preparing mushrooms. It contained erroneous and misleading information concerning the identification of highly toxic mushrooms. The plaintiffs trusted the book's descriptions, ate the mushrooms they had collected accordingly—mushrooms that turned out to be highly toxic—and became seriously ill. The plaintiffs based their claim on, among other things, 'products liability', on 'breach of warranty' and on 'negligent misrepresentation', but without success.

The Court rejected the plaintiffs' view that this work of reference was a 'product' in the meaning of the term 'products liability',[20] as only 'items of a tangible nature'[21] are included in this term. In the case of the contents of books, the court said, it was a question of non-tangible ideas, which were not comparable to 'products' in the abovementioned sense. In addition, the Court said that no other judgment arose in consideration of the type of publication. Moreover, any differentiation between the contents of guidebooks, encyclopedias, and abstract ideas was illusory.[22]

The Court also rejected liability on the grounds of negligent misrepresentation. It said although publishers had a fundamental duty to investigate the contents of their publications, insofar as there were no grounds, however, a further examination of the contents for its accuracy was not required.[23] In addition the Court rejected liability based on the law of warranty; for the above-mentioned reasons, the Court regarded it as unlikely that a book publisher would offer a warranty for the accuracy of the information.[24]

Another example worth mentioning on the question of tort liability of organs of the press is the case of *Alm v Van Nostrand Reinhold Co Inc*[25] Here the case involved instructions in the book *The Making of Tools*, published by the defendant. The plaintiff had incurred injuries while making a woodcarving tool as explained in the book, and brought a claim against the publisher. The plaintiff alleged 'negligent misrepresentation' on the grounds that the defendant failed to verify the accuracy of the book's contents themselves and independently of the author. The Court dismissed the claim, on the grounds of disproportionate scope of verification, as otherwise the publisher would be obliged to check all publications in detail and test them for their accuracy in order to prevent liability towards an unspecified number of people.[26]

Examining other US jurisprudence[27] as well shows that the liability of publishers for published content from the point of view of data quality/data accuracy is

19 [1991] 938 F 2d 1033 (9th Cir); see also *Jones v JB Lippincott Co* [1988] 694 F Supp 1216 (D Md) 15 Media L Rep 2155.

20 cf s 402A of the Restatement (Second) of Tort.

21 *Winter v GP Putnam's Sons* (n 19) 1034.

22 ibid 1036.

23 ibid 1037.

24 ibid 1038.

25 [1985] 480 NE 2d 1263 (Ill App 1 Dist.).

26 ibid 1264.

27 *Winter v GP Putnam's Sons* (n 19); ibid 1263; *Jones v JB Lippincott Co* (n 19).

interpreted globally in a restricted way. Extending the journalistic duty of care to book publishers from this point of view would create a crass discrepancy with regard to the mass of published works and could entail effects that might threaten their very existence.

To impose the suggested broad legal duty upon publishers of nationally circulated magazines, newspapers and other publications, would not only be impractical and unrealistic, but would have a staggering adverse effect on the commercial world and our economic system. For the law to permit such exposure to those in the publishing business who in good faith accept paid advertisements for a myriad of products would open the doors 'to a liability in an indeterminate amount for an indeterminate time to an indeterminate class'.[28]

According to US law, there is no fundamental distinction between liability for erroneous information in print and in digital form and the applicability of each law of liability depends on the individual case. In general, especially the English Courts have avoided creating an indeterminate liability towards an indeterminate class of the public by limiting sales liability to the defectiveness of physical goods and by limiting liability under tort law to clearly defined class to whom the duty of care is owed. [29]

THE RULES OF THE EU DATA PROTECTION DIRECTIVE ON DATA QUALITY

One initial fragmentary approach to a juristic validation of data quality is offered by Article 6 (1) (d) of the EU Data Protection Directive,[30] with its assertion that data, insofar as they relate to a person, have to be up to date and accurate ('accurate and, where necessary, kept up to date'). Astonishingly, this regulation has never been implemented in Germany and in this Germany remains almost alone in Europe. For example, in Austria the provisions concerning quality have been implemented in section 6 of the Austrian Data Protection Act. Switzerland has even extended the regulations. According to Article 5 of the Swiss Data Protection Act, the processor of personal data has to make sure of their accuracy. He must take all reasonable steps to correct or destroy data that are incorrect or incomplete in light of the purpose of its collection or processing.

In the UK, the EU Data Protection Directive was implemented as the Data Protection Act 1998. While the latter regulates the fundamentals of British data protection law, a concretization of these rules takes place through statutory instruments and Codes of Practices.[31] The 1998 Data Protection Act sets up a total of eight data protection principles. The quality regulation in Article 6 (1) (d) of the EU Data

28 *Yuhas v Mudge* [1974] 129 NJSuper 207, 209–10, 322 A 2d 824, 825; accord *Suarez v Underwood* [1980] 103 Misc 2d 445, 426 NYS 2d 208.

29 See, for example, *Smeaton v Equifax* [2013] EWCA Civ 108, paras 73–76.

30 This regulation can be found in almost identical form in art 5 of the draft of the EU General Data Protection Regulation.

31 V Bange and others, 'An Overview of UK Data Protection Law' (2012) <https://united-kingdom.taylorwessing.com/uploads/tx_siruplawyermanagement/NB_000168_Overview_UK_data_protection_law_WEB.pdf> accessed 17 August 2016.

Protection Directive was implemented through the fourth data protection principle, which prescribes that personal data must be up to date and accurate.[32]

For reasons of practicability, the Act provides special regulations for cases in which individuals provide information about themselves or personal data is acquired from third parties. In these cases, even if personal data is factually inaccurate, this is not considered a breach of the fourth data protection principle if, in the case of the data subject or a third party false information was entered correctly, the data controller has taken reasonable steps to ensure the quality of the data and the data indicate that the data subject has alerted the data controller to the inaccuracies.[33] What precisely is to be understood by 'reasonable steps' depends on the kind of personal data and on the importance of accuracy in each individual case.[34]

In the case of *Smeaton v Equifax Plc*, the UK Court of Appeal pointed out that the 1998 Data Protection Act justified no absolute obligation to maintain the accuracy of personal data, but merely demanded the taking of reasonable steps to maintain data quality. This reasonableness is to be judged according to each individual case. The Court also said that the fourth data protection principle did not lead to a parallel obligation in the area of legal torts.[35]

The silence of civil law is particularly astonishing if we look at the current significance of Article 6 of the EU Data Protection Directive in the discussion about legal policy. In its Google ruling,[36] the Court of Justice of the EU emphasized the principles of data quality and not without cause. It said that any processing of personal data must comply with the principles established in Article 6 of the Directive in relation to the quality of the data (paragraph 71).[37] On the principle of data accuracy the Court also said 'even initially lawful processing of accurate data may, in the course of time, become incompatible with the Directive where those data are no longer necessary in the light of the purposes for which they were collected or processed' (paragraph 93).

In the USA, the Data Quality Act (DQA), also known as the Information Quality Act (IQA), was adopted in 2001 as a component of the Consolidated Appropriations Act. It empowers the Office of Management and Budget to issue guidelines, which should guarantee and improve the quality and integrity of the information that is published by state institutions ('Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility and Integrity of Information Disseminated by Federal Agencies'[38]).[39] In addition, mechanisms should be created

32 Data Protection Act 1998, sch 1 pt 1, para 4; detailed information on the fourth principle of data protection: Information Commissioner's Office, 'Guide to Data Protection' (2016) <https://ico.org.uk/for-organisations/guide-to-data-protection/principle-4-accuracy/> accessed 17 August 2016.

33 Data Protection Act 1998, sch 1 pt 2 para 7.

34 Information Commissioner's Office (n 32).

35 [2013] EWCA BPIR 231.

36 Case C–131/12 *Google Spain Official Journal* C 212(7 July 2014) 4

37 cf *Rechnungshof v Österreichischer Rundfunk and others and Neukomm and Lauermann v Österreichischer Rundfunk* [2003] C–294 (ECJ) para 65; *ASNEF and FECEMD v Administración del Estado* [2011] C–777 (ECJ) para 26; *Worten v ACT* [2013] C–355 (ECJ) para 33.

38 JD Graham, 'Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility and Integrity of Information Disseminated by Federal Agencies' (2001) <https://www.whitehouse.gov/omb/fedreg_final_information_quality_guidelines/> accessed 17 August 2016.

39 ibid.

that enable data subjects affected by the dissemination of false information to have this flagged up and corrected.[40]

A distinction between personal and non-personal data is not made in this case, however. In addition, the scope of the DQA is limited only to the dissemination of information by state authorities to the public.[41]

Furthermore, there is no federal law that establishes guidelines for the data quality of personal data in the non-state domain. Since US data protection law is regulated by numerous laws and guidelines at both federal and state level, there are individual sector-specific laws that contain regulations touching on data quality (eg the Fair Credit Reporting Act or the 1996 Health Insurance Portability and Accountability Act). For example, the Fair Credit Reporting Act requires users of consumer reports to inform consumers of their rights to contest the accuracy of reports concerning them. Another example is the Health Insurance Portability and Accoutabbility Act HIPAA Security Rule according to which institutions concerned (eg health programs, settlement facilities in healthcare or healthcare companies) must guarantee the integrity of electronically protected health data.[42]

The example of the USA and the EU Data Protection Directive demonstrate that the growing relevance of data quality as an issue has at least been recognized. On the other hand, veracity[43] of data can only be attained if effective tools are created that can ensure quality standards for data. Both the EU Directive and the DQA are giving a lead in the right direction.

However, the fact that until now Germany has not implemented Article 6 of the EU Data Protection Directive at a national level, and that the DQA in the USA recognizes solely the dissemination of information by state institutions, nevertheless indicates a need for correction and reform exists.

SCORING AND BIG DATA

A further element of a legal validation of data quality is provided by section 28b of the German Federal Data Protection Act (BDSG) and its regulations on scoring. This offers a new criterion for the assessment of the way in which information is collected, namely the establishment of a scientifically recognized mathematical–statistical procedure for the calculation of probability value (no 1).[44]

The scope of this regulation is unclear. It may be interpreted the way that it can be applied beyond the narrow discipline of financial scoring to profiling and other big data assessments as well. This is backed up, for example, by the Federal Government's justification of the draft legislation: 'Scoring is a mathematical-statistical procedure that makes it possible to calculate the probability of a certain

40 DQA 2001, sub-s (2) (B).

41 AD Wait and JP Maney, 'Regulatory Science and the Data Quality Act' (2006) 18 ECJ 145, 148.

42 RP Jay, *Data Protection and Privacy 2015* (3rd edn, Law Business Research Ltd, 2014) 210ff.

43 Referring to T Douglas's 'Four Vs of Big Data' (volume, variety, velocity and veracity) Big Data And Beyond: How Companies Can Find Insight In Big Data (2015).

44 See also Niko Härting's early ideas on this topic: N Härting, 'Vier Thesen zur neu entbrannten Scoring-Debatte' (2015) <http://www.cr-online.de/blog/2015/05/20/vier-thesen-zur-neu-entbrannten-scoring-debatte/> accessed 17 August 2016.

person demonstrating certain behavior.'[45] There is absolutely no reference that says that scoring must be related and limited to credit checks. The only limitation contained in the regulation is the reference that scoring may be used 'for the purpose of deciding on the creation, execution or termination of a contractual relationship with the data subject'. Bringing the concept of probability values into play goes far beyond the usual procedure of credit scoring. For example, in all business transactions, prognoses inevitably have an influence on the decision concerning a business deal. In a similar way, many big data processes are based on scoring that has an influence on the creation of differentiated business models (in the area of health insurance, for example).

This categorization has far-reaching consequences for the world of big data. According to section 28b German Federal Data Protection Act BDSG, the mathematical standards must be 'demonstrably essential' for calculating the probability of the action. The reference to 'demonstrability' shifts the burden of explanation and proof onto the big data analysts and gives the data protection supervisory authority, under section 38 (3) sentence 1 BDSG, the opportunity of being informed about the parameters of demonstrability in the case of the use of personal data.[46]

THE QUESTION OF DATA QUALITY IS NOT A PROBLEM OF DATA PROTECTION LEGISLATION

Notably, both specifications (Article 6 of the EU Data Protection Directive and section 28b BDSG) are incorrectly qualified as data protection legislation. The background to this is the deliberate legal confusion of consumer protection and data protection in such a way that data protection law becomes an extension of consumer protection law. However, questions as to the accuracy of data or the basis of scoring affect not only consumers but business people as well. To that extent here too it is not a question of consumer protection, but of a general legal promotion of the accuracy of data analysis in light of big data.[47]

In this respect, scoring is by no means a question relating to the admissibility of the use of personal data, but rather to the accuracy of the relevant procedures and their results. However, data protection legislation has nothing to do with the question of the accuracy of data. Equally, consumer protection law does not address the central question of data accuracy: in the world of business in particular, there is also a very pressing need for protection against the irresponsible use of big data tools. The guarantee of data quality is an issue of civil law in general.

In this respect, the BDSG legislates for a situation that dogmatically bypasses the objective of data protection law. Accordingly, the objective of section 38 BDSG is also incorrect, which in combination with section 28b (1) is intended to enable the

45 BT-Drs 16/10529 1.

46 To this extent, it is regrettable that precisely this component of s 28b BDSG is not to be incorporated into the EU's General Data Protection Regulation. According to suggestions made by the Commission, Parliament and Council on art 20 DCGVO, first of all an 'automatic decision' (Council) or 'measure' (Commission) should be provided based on profiling that is 'profiling which leads to measures producing legal effects concerning the data subject or does similarly significantly affect the interests, rights or freedoms of the concerned data subject' (European Parliament).

47 N Härting (n 44); BT-Drs 16/10529 1ff.

supervisory authority to understand the established context scientifically. The data protection supervisory authority is in no position to judge the mathematical–statistical validity of scoring procedures. It has never been their job, or their core area of competence. In that case the data protection supervisory authority would obviously have had to employ mathematicians to check the validity, which would lead to additional administration costs, which, however, the government's draft of the then existing BDSG definitively excluded in its justification of section 28b BDSG.[48]

THE QUESTION IS HOW DATA SHOULD BE COLLECTED IN THE INTEREST OF EVERYONE

The stereotypically defensive attitude against scoring/profiling makes the error of thinking the issue of data accuracy must be of interest to all participants in the flow of data. In this respect, it cannot be a question of fighting against scoring/profiling, but of promoting data accuracy within the scoring system. We achieve nothing by polemicizing against the overpowering hunger for data; rather we will have to regulate in a focused way on the 'how' of data assessment in the context of today's data society.

IT IS A NOT A QUESTION OF 'RIGHT' OR 'WRONG'

One idea might be to protect data quality through a connection to the relevant data quality standards and to demand it through tough instruments of civil law. Categorizing data as 'right' or 'wrong' is not appropriate. Big data is concerned with correlations and probabilities, and is not suited to dualistic assertions of truth. But it is precisely in equating probabilities and facts that we find one of the biggest cases of liability in the debate about big data. As early as 2010, Danah Boyd, a renowned US sociologist, issued a warning about the tragic misunderstandings in this field: 'Bigger data are not always better data.'[49] And she warned justifiably.

Interpretation is the hardest part of doing data analysis. And no matter how big your data is, if you do not understand the limits of it, if you do not understand your own biases, you will misinterpret it.

MODERN MODELS OF DATA QUALITY

It turns out to be disastrous that, after an initial period of activity, the discussion in IT about standards in data quality has subsided once again. The 1990s gave birth to current data quality standards such as accuracy, consistency, timeliness, completeness and uniqueness.[50] This debate continued until 2008 and led to the foundation of a

48 BT-Drs 16/12011, 18 below.

49 D Boyd, 'Privacy and Publicity in the Context of Big Data' (2010) <http://www.danah.org/papers/talks/2010/WWW2010.html> accessed 17 August 2016.

50 B Heinrich and M Klier, 'Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement' in K Hildebrand and others (eds), *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (2nd edn, Vieweg+Teubner 2011) 49–66; B Heinrich and others, 'How to Measure Data Quality? - A Metric-Based Approach' (2007) 28th International Conference of Information Systems (ICIS) 2007 Proceedings Paper 108 <http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1265&context=icis2007> accessed 30 November 2016.

Deutsche Gesellschaft für Datenqualität (DGDQ) [German Society for Data Quality] which was then *de facto* disbanded. Currently, an ISO standard (ISO 8000)[51] is under consideration whose form cannot yet be determined.

Meanwhile the differentiation between five levels of quality has become standard: availability, usability, reliability, relevance and presentation quality.[52] These five levels have formed the basis for Chinese researchers, for example, to summarize the current debate in the context of a highly differentiated model of the postulated standard.[53]

The standard for the evaluation of data quality, as established by Li Cai and Yangyong Zhu uses five levels of quality standard.

The availability of data consists of its accessibility and timeliness: accessibility meaning that the data can be accessed through an interface and that it can be made public or purchased easily; and timeliness meaning that data arrive on time, are regularly updated, and that the time interval between data collection and processing meets these requirements.

The test for the usability of data is its credibility. It has to be asked whether data come from a specialized organization of a country, field or industry, whether experts or specialists regularly audit and check the correctness of its content and whether the data exist in the range of known or acceptable values.

The most comprehensive test is that of data reliability which is divided into data accuracy, consistency, integrity and completeness. In addition to the data provided being accurate, their representation and value reflect the true state of the source information and will not cause ambiguity. Data consistency implies that data concepts, values and formats still match after it has been processed, that data remain consistent and verifiable during a certain period of time and that the data are consistent and verifiable in relation to data from other sources. Data integrity can be summarized as the data format being clear and within certain criteria, and data being consistent with structural as well as content integrity. Finally, for data completeness it has to be tested whether the deficiency of a component will impact data usage for data with multi-components or data accuracy and integrity.

Data relevance can be translated as data fitness. This means the collected data may not completely match a certain theme, but expound one aspect, most of the retrieved data sets are within the retrieval theme of the data user, and that the information theme provides matches with users' retrieval theme.

Ultimately, data presentation quality is a test for their readability. Data (content, format etc) have to be clear and understandable, particularly as to description, classification and coding content. It must be easy to judge that the data provided meet certain needs.

This model highlights the complexity of quality assurance in the case of big data. What is required here is not only the accuracy of the input data (mentioned here under point 3.1 'Accuracy'), but rather the entire procedure from the inputting of

51 JL Wang and others, 'Research on ISO 8000 Series Standards for Data Quality' (2010) 12 Stan Sci 44.

52 C Cappiello and others, 'Data Quality Assessment from the User's Perspective' (2004) ACM 68.

53 Cited from: L Cai and Y Zhu, 'The Challenges of Data Quality and Data Quality Assessment in the Big Data Era' (2015) <http://datascience.codata.org/article/10.5334/dsj-2015-002/> accessed 17 August 2016.

data to the presentation of the final data correlations must be structured appropriately.

Under 'availability' they distinguish between 'accessibility' of data and 'timeliness'. The authors measure accessibility using indicators such as being able to access data through an interface and the possibility of receiving data free of charge or at a reasonable price. They would like to ensure timeliness using procedures that guarantee the regular updating of input data and an appropriate projection of the time periods from input through processing to output. Usability too must be ensured, for example, through regular auditing by experts or by examination of the source of input data. Moreover, we should note that they demand not only the relevance of the output (its 'fitness for purpose'), but that emphasis is placed on 'readability' as well—the intelligibility of the output and its presentation in a way that avoids misunderstandings.

AND NOW: INPUT, PROCESSING, OUTPUT—AND LIABILITY

It is particularly important to revisit the old debates on data quality in light of big data, as the output of a big data assessment can be disastrous if the data entered are assessed incorrectly, twice over or inconsistently. This is what gives rise to factually incorrect results on the basis of mathematically correct and apparently clean methods.

An efficient legal system would be able to distinguish between input, processing and output. The input would have to meet classic contractually binding data quality standards that have an influence on the contractual relationship between data sellers and buyers in the form of the usual stipulated conditions. This includes above all the aforementioned criteria of availability. In the contractual relationship between big data analyst and big data customer, the classic data quality criteria equally apply, in particular the 'fitness test; as is usually required. In relation to interested third parties, the infringement of the data quality standards as per sections 823 (1) and 824 BGB in the context of the test for negligence would apply. This would presume that during big data processing a record of the tools used would be required analogous to section 28b BDSG and that big data companies must reveal the basis of their assessment of individual data to their customers and interested parties.[54]

The requirement of data quality has thus to become one of the crucial points in the discussion on big data. Especially, the law has to establish a regime of requirements that guarantee a high level of data quality—in the interest of the buyers of data material and the person concerned by decisions based on big data tools. Such a regime cannot be constructed simply within law itself. It is based on a high level of interdisciplinary discussions between lawyers and computer scientists. These discussions may result in technical standards on data quality, for instance, established with the ISO system. These standards can then be used in the legal setting to define the normative level of expectations within contract and tort law.

54 cf the US perspective in O Tene and J Polonetsky, 'Big Data for All: Privacy and User Control in the Age of Analytics' (2013) 11 NW J Tech Intell Prop 239, 270ff. On the 'Big Data Disclosure Problem', see also M Mattioli, *Disclosing Big Data* (2014) 99 Minn L Rev 535.