7

TRIPS Meets Big Data

*Daniel J. Gervais**

'Artificial intelligence is another emerging area focusing in IPR protection, used mostly in the tech industry, producing new products and services every year. Artificial intelligence (AI) will redefine how individuals think about daily life, and start-ups will need to start leveraging AI to get ahead.'[1]

Even as the United States is playing 'hard ball' at the World Trade Organization (WTO) in the area of dispute settlement, the quote demonstrates its willingness to engage in discussions on the topic of artificial intelligence at the WTO. The United States is not alone. In this chapter, I review some of the work done at on AI and big data in the WTO and in particular under the Agreement on Trade-Related Intellectual Property Rights (TRIPS),[2] and reflect on how this work is likely to progress. I begin, however, by defining the topic.

A DEFINING BIG DATA AND AI[3]

The term 'big data' can be defined in a number of ways. A common way to define it is to enumerate its three essential features, a fourth that, though not essential, is increasingly typical, and a fifth that is derived from the other three (or four). Those

* Daniel J. Gervais, PhD, MAE, is the Milton R. Underwood Chair in Law and Director of the Intellectual Property Program,Vanderbilt Law School, as well as Professor of Information Law, University of Amsterdam. Contact: daniel.gervais@vanderbilt.edu.

[1] Summary of Statement by the United States, Council for Trade-Related Aspects of Intellectual Property Rights, Minutes of Meeting held in the Centre William Rappard, 8–9 November 2018, IP/C/M/90/Add.1, 15 January 2019, at para. 363.

[2] Agreement on Trade-Related Aspects of Intellectual Property Rights, 1869 U.N.T.S. 299; 33 I.L.M. 1197 (1994), entered into force 1 January 1995 [hereinafter: TRIPS].

[3] An earlier version of this part of the chapter appeared in the *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* in 2019.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

features are volume, veracity, velocity, variety, and value.[4] 'Volume' or size is, as the term big data suggests, the first characteristic that distinguishes big data from other ('small data') datasets. Because big data corpora are often generated automatically, the question of the quality or trustworthiness of the data ('veracity') is crucial. 'Velocity' refers to 'the speed at which corpora of data are being generated, collected and analyzed'.[5] The term 'variety' denotes the many types of data and data sources from which data can be collected, including Internet browsers, social media sites and apps, cameras, cars, and a host of other data-collection tools.[6] Finally, if all previous features are present, a big data corpus likely has significant 'value'.

The way in which 'big data' is generated and used can be separated into two phases.[7] First, the creation of a big data corpus requires processes to collect data from sources such as those mentioned in the previous paragraph. Second, the corpus is analysed, a process that may involve Text and Data Mining (TDM).[8] TDM is a process that uses an AI algorithm. It allows the machine to learn from the corpus; hence the term 'machine learning' (ML) is sometimes used as a synonym of AI in the press.[9] As it analyses a big data corpus, the machine *learns and gets better at what it does*. This process often requires human input to assist the machine in correcting errors or faulty correlations derived from, or decisions based on, the data.[10] The processing of corpora of big data is done to find correlations and generate predictions or other valuable analytical outcomes. The found correlations and insight can be used for multiple purposes, including targeted advertising and surveillance, though an almost endless array of other applications is possible. To take just one different example of a lesser known application, a law firm might process hundreds or thousands of documents in a given field, couple ML with human expertise, and produce insights about how they and other firms operate, for instance, in negotiating a certain type of transaction or settling (or not) cases.

A subset of machine learning, known as *deep learning* (DL), uses neural networks, a computer system modelled on the human brain.[11] This implies that any human

[4] J. Cano, 'The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity', *XSNet*, 11 March 2014.

[5] Ibid.

[6] The list includes 'cars' as personal vehicles are one of the main sources of (personal) data with up to 25 gigabytes per hour of driving.

[7] The two components are not necessarily sequential. They can and often do proceed in parallel.

[8] See M. Montagnani, 'Il text and data mining e il diritto d'autore', *Annali Italiani del diritto d'autore, della cultura e dello spettacolo* 26 (2017), 376–395.

[9] C. Kozyrkov, 'Are You Using the Term 'AI' Incorrectly?', *Hackernoon*, 26 May 2018.

[10] How IP will apply to the work involved in the human training function of machine learning is one of the interesting questions at the interface of big data and IP. The term 'training data' is used in this context to suggest that the machine training is supervised (by humans). See B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press, 1996), at 354.

[11] With the 'deep learning model, the algorithms can determine on their own if a prediction is accurate or not … through its own method of computing – its own "brain", if you will'.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

contribution to the output of deep learning systems is often 'second degree' and the proximate cause of the output is not the programmer. When considering the possible intellectual property (IP) protection of outputs of such systems, this separation between humans and the output challenges core notions of IP law, especially authorship in copyright law and inventorship in patent law.

ML and DL can produce high value outputs. Such outputs can take the form of analyses, insights, correlations, and may lead to automated (machine) decision-making. It can be expected that those who generate this value will try to capture and protect it, using IP law, technological measures and contracts. One can also expect competitors and the public to try to access those outputs for the same reason, namely their value. In many cases, big data corpora are protected by secrecy, a form of protection that relies on trade secret law combined with technological protection from hacking, and contracts. A publicly available corpus, in contrast, must rely on erga omnes IP protection – if it deserves protection to begin with. Copyright protects collections of data; the sui generis database right (in the European Union, EU) might apply; and data exclusivity rights in clinical trial data may be relevant.

The *outputs* of the processing of big data corpora may contain or consist of subject matter that facially could be protected by copyright or patent law. Big data technology can be – and in fact is – used to create and invent. For example, a big data corpus of all recent pop music can find correlations and identify what may be causing a song to be popular. It can use the correlations to write its own music.[12] The creation of (potentially massive amounts of) new literary and artistic material without direct human input will challenge human-created works in the marketplace. This is already happening with machine-written news reports.[13] Deciding whether machine-created material should be protected by copyright could thus have a profound impact on the market for creative works. If machine created material is copyright-free, machines will produce free goods that compete with paid ones – that is, those created by humans expecting a financial return. If the material produced by machines is protected by copyright and its use potentially subject to payment, this might level the commercial playing field between human and machine, but then who (which natural or legal person) *should* be paid for the computer's work? Then there will be border definition issues. Some works will be created by human and machine working together. Can we apply the notion of joint authorship? Or should we consider the machine-produced portion (if separable) copyright-free, thus limiting the protection to identifiably human-authored portions?

B. Grossfeld, 'A Simple Way to Understand Machine Learning vs Deep Learning', *ZenDesk*, 18 July 2017.

[12] See G. Hadjeres and F. Pachet, 'DeepBach: A Steerable Model for Bach Chorales Generation', *arXiv*:1612, 3 December 2016, 1–20, at 1.

[13] See C. Underwood, 'Automated Journalism – AI Applications at *New York Times*, *Reuters*, and Other Media Giants', *eMerj*, 17 November 2019, available at https://bit.ly/2Q84BTV.

If such major doctrinal challenges – each with embedded layers of normative inquiries – emerge in the field of copyright, big data poses existential threats in the case of patents. AI tools can be used to process thousands of published patents and patent applications and used to *expand the scope of claims in patent applications*. This poses normative challenges that parallel those enunciated earlier: Who is the inventor? Is there a justification to grant an exclusive right to a machine-made invention? To whom? There are doctrinal ones as well. For example, is the machine-generated 'invention' disclosed in such a way that would warrant the issuance of a patent?

It gets more complicated. If AI machines using patent-related big data can broaden claim scope or add claims in patent applications, then within a short horizon they could be able to *predict the next incremental steps in a given field of activity* by analysing innovation trajectories. For example, they might look at the path of development of a specific item (car brakes, toothbrushes) and 'predict' or define a broad array of what *could* come next. Doctrinally, this raises questions about inventive step: If a future development is obvious to a machine, is it obvious for purposes of patent law? Answering this question poses an epistemological as well as a doctrinal challenge for patent offices. The related normative inquiry is the one mentioned earlier, namely whether machine-made inventions (even for inventions the scope [claims] of which were merely 'stretched' using big data and AI) 'deserve' a patent despite their obviousness (to the machine).

This use of patent and technological big data could lead to a future where machines pre-disclose incremental innovations (and their use) in such a way that they constitute publicly available prior art and thus make obtaining patents impossible on a significant part of the current patentability universe. Perhaps even the best AI system using a big data corpus of all published patents and technical literature will not be able to predict the next pioneer invention, but very few patents are granted on ground-breaking advances. AI systems that soon will be able to predict *most* improvements to currently patented inventions, which tend to be only incrementally different from the prior art would wreak havoc with the patent-based incentive system.[14] Let us take an example: It is possible that deep-learning algorithms could parse thousands of new molecules based on those recently patented or disclosed in applications and even predict their medical efficacy. If such data (new molecules and predicted efficacy) were available and published, it would significantly hamper the patentability of those new molecules due to lack of novelty.

The unavailability of patents would dramatically increase the role of data exclusivity rights – the right to prevent reliance in clinical data submitted to obtain marketing approval – in the pharmaceutical field.[15] If this prediction of future

[14] See S. Y. Ravid and X. Liu, 'When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3A Era', *Cardozo Law Review* 39 (2018), 2215–2263, at 2254; T. Baker, 'Pioneers in Technology: A Proposed System for Classifying and Rewarding Extraordinary Inventions' *Arizona Law Review* 45 (2003), 445–466.

[15] See D. Gervais, 'The Patent Option', *North Carolina Journal of Law and Technology* 20 (2019), 357–403.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

inventions by AI became an established practice in fields where this separate protection by data exclusivity is unavailable, the very existence of the incentive system based on patents could be in jeopardy.

B BIG DATA IN THE WTO'S WORK

Big data has slowly made its way past the imposing iron gates of rue de Lausanne and into the WTO. Big data has made appearances in various WTO committees and at the General Council. At the committee level, it showed up in the work that the WTO is doing on 'electronic commerce', based on a Work Programme on that topic adopted by the General Council on 25 September 1998.[16] The Work Programme required the Committees on Trade in Goods and Trade in Services, the Council for TRIPS and the Committee for Trade and Development to 'examine and report' on how electronic commerce might impact each of those trade sectors.[17]

In the area of intellectual property, work began quickly after the adoption of the Work Programme. In 1998, the Secretariat published a note reflecting the thinking on IP, just a few years after the adoption of the TRIPS Agreement. The note stated that intellectual property plays an important role also in promoting the development of the infrastructure of [electronic communications networks], i.e. software, hardware and other technology that make up information highways. It provides protection to the results of investment in the development of new information and communications technology, thus giving the incentive and the means to finance research and development aimed at improving such technology. In addition, a functioning intellectual property regime facilitates transfer of information and communications technology in the form of foreign direct investment, joint ventures and licensing.[18]

Along the same lines, but in a much more recent discussion of AI and big data in the Committee on Regional Trade Agreements, in response to a question from Canada as to whether there were 'effective measures to curtail repetitive infringement of copyright and related rights on the Internet' in the China–Korea Free Trade Agreement (FTA), China and Korea stated in their joint response that China would '[p]romote the cooperation of electric [sic]-commerce Big Data between the government and the industries to ensure the efficiency of information searching and evidence obtaining'.[19] Here big data and AI were seen as adjuncts for copyright

[16] WTO, Work Programme on Electronic Commerce, WT/L/274, 30 September 1998.
[17] Ibid.
[18] WTO, General Council, WTO Agreements and Electronic Commerce: Note by the Secretariat, WT/GC/W/90, 14 July 1998.
[19] WTO, Committee on Regional Trade Agreements, Free Trade Agreement between China and the Republic of Korea (Goods and Services): Questions and Replies, WT/REG370/2, 6 November 2017, at 3.

enforcement. One might question whether what seems a high protectionist view is always warranted in the face of empirical data about open innovation models, for example.

Some WTO members have suggested a broader role. Japan, for example, mentioned the need to address issues of 'digital protectionism', noting that the digital economy has contributed to global economic growth. Furthermore, the Fourth Industrial Revolution, realised with the utilisation of the latest technology such as the Internet of Things and Big Data will permeate countless aspects of the world economy and people's lives . . . . However, a number of challenges still remain to be addressed in order to maximize the benefits from this trend. . . . Among others, it is indispensable to address emerging "digital protectionism"'.[20]

Though it is not clear exactly what Japan had in mind in this statement, digital protectionism is often shorthand for an attempt to restrain regulatory autonomy on the protection of personal data and data localization.[21]

In a so-called 'non paper', Brazil also raised the question whether 'usage of big data' would require a debate on concepts like universal jurisdiction or choice of jurisdiction applicable to electronic commerce.[22] Developing countries have also had their say. India underscored the need for developing countries 'to maintain policy space to formulate a policy on ownership, use and flow of data in sunrise sectors like cloud computing, data storage, hosting of servers as well as in big data analytics'.[23] They are, therefore, committed to reinvigorate work on the multilateral track, with its non-negotiating mandate, to understand these issues.[24] Rwanda's more sombre observation was that 'empirical evidence showed that the digital market was highly concentrated and that only a few companies worldwide were dominating the digital market, specializing in management and development of data centers and exploiting [B]ig [D]ata'.[25] It noted that only a few developing countries were able to catch up.[26] Finally, UNCTAD sought support to assist WTO members in adapting 'domestic IP frameworks to recent technological developments in big data solutions

[20] WTO, Work Programme on Electronic Commerce, Non-paper for the Discussions on Electronic Commerce/Digital Trade from Japan, JOB/GC/100, 25 July 2016, at paras. 2.1 and 2.2.

[21] See Chapter 1 in this volume and see S. Yakovleva, Privacy Protection(ism): The Latest Wave of Trade Constraints on Regulatory Autonomy, University of Miami Law Review 74 (2020), 416–519.

[22] WTO, Exploratory Work on Electronic Commerce, Non-paper from Brazil, NF/ECOM/3, 25 March 2019, at 5.

[23] WTO, General Council, Minutes of the Meeting held in the Centre William Rappard on 18 October 2018, Statement by India, WT/GC/M/174, 20 November 2018, at 41.

[24] Ibid.

[25] See WTO, *Aid for Trade Global Review 2017: Promoting Trade, Inclusiveness and Connectivity for Sustainable Development: Summary Report* (Geneva: WTO, 2017), at 203; also WTO, General Council, Minutes of the Meeting held in the Centre William Rappard on 26 July 2017, WT/GC/M/168, 22 September 2017, at 7.248.

[26] Ibid.

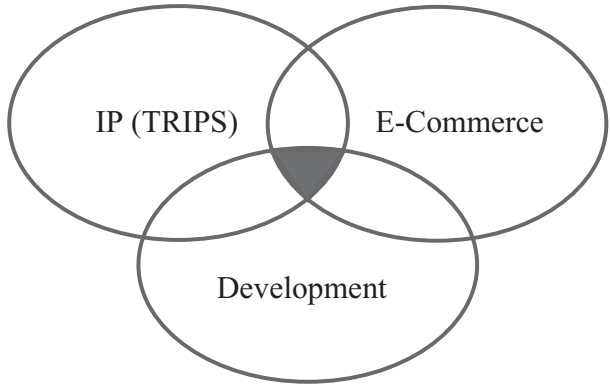

FIGURE 7.1. WTO work on AI and big data in thematic areas

and artificial intelligence'.[27] At this juncture, administratively the work on AI and big data at the WTO looks something as depicted below (Figure 7.1).

The future work of the WTO may progress in a number of different directions. It could usefully review how IP rights are actually used in the area of AI and big data, thus at least providing empirical data for future discussions. If the adoption of 'TRIPS 2.0' remains on the distant horizon, it seems clear that AI and big data issues will be on the table if and when it happens. In the intersection between IP and development, providing this type of analysis could be helpful to policymakers and development-focused international organizations outside the WTO as they develop domestic policies to facilitate the growth of AI and big data–based industries. The e-commerce and IP intersection includes how trade secret and other forms of IP apply to big data corpora. Again, more detailed work on this issue, whether comparative in nature or more theoretical, could open a useful window on various policy decisions.

In the next (and last) part of the chapter, I review a few areas in which the WTO could make analytical progress to make future discussions more productive, paying specific regard to the TRIPS Agreement.

C ADAPTING INTELLECTUAL PROPERTY TO BIG DATA AND AI

I *Intellectual Property Rights Protection of Big Data Software and Corpora*

Human-written AI software code used to collect (including search and social media apps), store and analyse big data corpora is considered a literary work eligible for copyright protection, subject to possible exclusions and limitations. That much is

[27] WTO, Council for Trade-Related Aspects of Intellectual Property Rights, Minutes of Meeting held in the Centre William Rappard on 5–6 June 2018, Statement by UNCTAD, IP/C/M/89/Add.1, 13 September 2018, at 38.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

already in TRIPS.[28] The TRIPS Agreement also protects '[c]ompilations of data or other material, whether in machine readable or other form', which might seem like mandatory protection for big data corpora.[29] This is however not necessarily so. Indeed, Article 10.2 TRIPS imposes a condition for such protection, namely that the compilations 'by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such'.[30] This condition is a way of stating that the compilation must be 'original' as the term is defined in international copyright law.

TRIPS incorporates most of the substantive provisions of the Berne Convention, to which 179 countries were party as of April 2021.[31] The convention contains important hints as to what constitutes an 'original' work. In its Article 2, when discussing the protection of 'collections', it states that '[c]ollections of literary or artistic works such as encyclopaedias and anthologies which, by reason of the *selection and arrangement of their contents*, *constitute intellectual creations* shall be protected *as such*, without prejudice to the copyright in each of the works forming part of such collections'.[32] This is the language that was reused in Article 10.2 TRIPS.

Selection and arrangement are exemplars of what copyright scholars refer to as 'creative choices'.[33] Creative choices need not be artistic or aesthetic in nature, but it seems they do have to be human.[34] Relevant choices are reflected in the particular way an author describes, explains, illustrates, or embodies their creative contribution. In contrast, choices that are merely routine (e.g., the choice to organize a directory in alphabetical order) or significantly constrained by external factors, such as the function a work is intended to serve (e.g., providing accurate driving directions), the tools used to produce it (e.g., a sculptor's marble and chisel), and the practices or conventions standard to a particular type of work (e.g. the structure of a

[28] This is recognized, for example, in Article 10(1) TRIPS, which provides that '[c]omputer programs, whether in source or object code, shall be protected as literary works under the Berne Convention (1971)'.

[29] Article 10.2 TRIPS.

[30] Ibid.

[31] Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886, last revised at Paris on 24 July 1971, and amended on 28 September 1979 [hereinafter: Berne Convention]. On membership of the Berne Union, see www.wipo.int/treaties/en/ShowResults.jsp?lang=en&treaty_id=15.

[32] Article 2.5 Berne Convention (emphasis added).

[33] See D. Gervais and E. F. Judge, 'Of Silos and Constellations: Comparing Notions of Originality in Copyright Law', *Cardozo Arts and Entertainment Law Journal* 27 (2009), 375–408.

[34] Deciding whether big data corpora are protectable in the absence of an identifiable human author is a debate well beyond the scope of this paper. See P. B. Hugenholtz, J. P. Quintais, and D. Gervais, Trends and Developments in Artificial Intelligence: Challenges to the Intellectual Property Rights Framework (Amsterdam: Institute for Information Law, 2021); D. Gervais, 'The Machine As Author', Iowa Law Review 105 (2019), 2053–2106. This statement from the United States Copyright Office is also interesting: 'Examples of situations where the Office will refuse to register a claim include: … The work lacks human authorship'. See United States Copyright Office, *Compendium of US Copyright Office Practices*, 3rd edn (Washington, DC: United States Copyright Office, 2017), at 22.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

sonnet) are not creative for the purpose of determining the existence of a sufficient degree of originality.

When the Berne Convention text was last revised on substance in 1967,[35] neither publicly available 'electronic' databases nor any mass-market database software was available. The 'collections' referred to in the convention are thus of the type mentioned by the convention drafters: (paper-based) anthologies and encyclopaedias. When 'electronic' databases started to emerge in the 1990s, data generally had to be indexed and re-indexed regularly to be useable. The TRIPS Agreement, signed in 1994 but essentially drafted in the late 1980s up to December 1990, is a reflection of this development.[36] The data in typical (relational or 'SQL') databases in existence at the time generally was 'structured' in some way, for example via an index, and that structure might qualify the database for (thin) copyright protection in the database's organizational layer. Older databases also contained more limited datasets ('small data').

Facebook, Google, and Amazon, to name just those three, found out early on that relational databases were not a good solution for the volumes and types of data that they were dealing with. This inadequacy explains the development of open source software (OSS) for big data: the Hadoop file system, the MapReduce programming language, and associated non-relational ('noSQL') databases, such as Apache's Cassandra.[37] These tools and the data corpora they helped create and use may not qualify for protection as 'databases' under the SQL-derived criteria mentioned earlier. This does not mean that no work or knowhow is required to create the corpus, but that the type of structure of the dataset may not qualify. As the CJEU explained in *Football Dataco*, 'significant labour and skill of its author … cannot as such justify the protection of it by copyright under Directive 96/9, if that labour and that skill do not express any originality in the selection or arrangement of that data'.[38] Indeed, big data is sometimes defined in *direct contrast* to the notion of SQL databases implicitly reflected in the TRIPS Agreement and the EU Database Directive discussed in the next section. Big data software is unlikely to 'select or arrange' the data in a way that would meet the originality criterion and trigger copyright protection.

[35] An Appendix for developing countries was added in Paris in 1971 but it did not modify the definition of 'work'.

[36] For a longer description of the negotiating history, see D. Gervais, *The TRIPS Agreement: Drafting History and Analysis*, 5th edn (London: Sweet and Maxwell, 2021), at Part I.

[37] See A. Reeve, 'Big Data and NoSQL: The Problem with Relational Databases', *Dell Technologies InFocus*, 7 September 2012, available at https://infocus.dellemc.com/april_reeve/big-data-and-nosql-the-problem-with-relational-databases/. It is worth noting that it is because code is protected by copyright (see TRIPS Agreement, Article 10.1) that owners of code can licence it and impose open source terms.

[38] C-604/10, *Football Dataco Ltd and others* v. *Yahoo!* [2012], ECLI:EU:C:2012:115, at 42.

Finally, it is worth noting that, in some jurisdictions, even absent copyright protection for big data, other IP-like remedies might be relevant, such as the tort of misappropriation applicable to 'hot news' in US law, or the protection against parasitic behaviour available in a number of European systems.[39] This might apply to information generated by AI-based TDM systems that have initially high but fast declining value, such as financial information relevant to stock market transactions, as data 'has a limited lifespan – old data is not nearly as valuable as new data – and the value of data lessens considerably over time'.[40]

In EU law, there is also a sui generis right in databases.[41] This right is not subject to the originality requirement,[42] but, according to Professor Bernt Hugenholtz, the way in which big data coprora are structured (or not) 'squarely rules out protection – whether by copyright or by the *sui generis* right – of (collections of) raw machine-generated data'.[43] The directive also mentions, however, that if there is an *investment* in obtaining the data, that investment may be sufficient for the corpus to qualify as a database.[44] The Court of Justice of the European Union (CJEU) defined 'investment' in obtaining the data as 'resources used to seek out existing materials and collect them in the database but does not cover the resources used for the creation of materials which make up the contents of a database'.[45] Professor Hugenholtz explains that 'the main argument for this distinction, as is transparent from the decision, is that the Database Directive's economic rationale is to promote and reward investment in database production, not in generating new data'.[46] This casts doubt on whether the notion of investment is sufficient to warrant sui generis protection of big data corpora, though Matthias Leistner suggested caution in

[39] See V. Smith Ekstrand and C. Roush, 'From "Hot News" to "Hot Data": The Rise of "FinTech", the Ownership of Big Data, and the Future of the Hot News Doctrine', *Cardozo Arts and Entertainment Law Journal* 35 (2017), 303–339.

[40] D. Sokol and R. E. Comerford, 'Antitrust and Regulating Big Data', *George Mason Law Review* 23 (2016), 1129–1161, at 1138.

[41] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, OJ L [1996] 77/20 [hereinafter: Database Directive]. See also D. J. Gervais, 'The Protection of Databases', *Chicago-Kent Law Review* 82 (2007), 1101–1168.

[42] See P. B. Hugenholtz, 'Intellectual Property and Information Law', in J. J. C. Kabel and G. J. H. M. Mom (eds), *Essays in Honour of Herman Cohen Jehoram* (The Hague/London/Boston: Kluwer Law International, 1998), 183–200.

[43] P. B. Hugenholtz, 'Data Property: Unwelcome Guest in the House of IP', in P. Drahos, G. Ghidini, and H. Ullrich (eds), *Kritika: Essays on Intellectual Property*, Vol. 3 (Cheltenham: Edward Elgar, 2018), 65–77. See also E. Derclaye, 'The Database Directive', in I. Stamatoudi and P. Torremans (eds), *EU Copyright Law* (Cheltenham: Edward Elgar, 2014), 298–354, at 302–303.

[44] Article 7(1) Database Directive.

[45] C-46/02, *Fixtures Marketing Ltd* v. *Oy Veikkaus Ab* [2004], ECLI:EU:C:2004:694; C-203/02, *British Horseracing Board* v. *William Hill Organization* [2004], ECLI:EU:C:2004:695; C-338/02, *Fixtures Marketing Ltd* v. *Svenska Spel AB* [2004], ECLI:EU:C:2004:696; C-444/02, *Fixtures Marketing Ltd* v. *Organismos prognostikon agonon podosfairou AE (OPAP)* [2004], ECLI:EU:C:2004:697.

[46] Hugenholtz, above note 43.

opining that 'the sweeping conclusion that all sensor- or other machine-generated data will typically not be covered by the sui generis right is not warranted'.[47]

II *Text and Data Mining*

The WTO could usefully consider the need for TDM exceptions, and how they mesh with the three-step test contained in Article 13 TRIPS, as many WTO members have adopted or are considering adopting exceptions for this purpose. TDM software used to process corpora of big data might infringe rights in databases that are protected either by copyright or the EU sui generis right, thus creating a barrier to TDM.[48] The rule that copyright works reproduced in a big data corpus retain independent copyright protection has not been altered. This means that images, texts, musical works, and other copyright subject-matter contained in a big data corpus are still subject to copyright protection until the expiry of the term of protection. This is clearly reflected in Article 10.2 TRIPS, second sentence: 'Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself'.

Geiger et al. opined that '[o]nly TDM tools involving minimal copying of a few words or crawling through data and processing each item separately could be operated without running into a potential liability for copyright infringement'.[49] This might explain why several jurisdictions have introduced TDM limitations and exceptions. Four examples should suffice to illustrate the point. First, the German Copyright Act contains an exception for the 'automatic analysis of large numbers of works (source material) for scientific research' for non-commercial purposes.[50] A corpus may be made available to 'a specifically limited circle of persons for their joint scientific research, as well as to individual third persons for the purpose of monitoring the quality of scientific research'.[51] The corpus must also be deleted once the research has been completed.[52] Second, France introduced an exception

[47] M. Leistner, 'Big Data and the EU Database Directive 96/9/EC: Current Law and Potential for Reform', *SSRN Publication* (2018), available at https://ssrn.com/abstract=3245937.

[48] See D. L. Rubinfeld and M. S. Gal, 'Access Barriers to Big Data', *Arizona Law Review* 59 (2017), 339–381, at 368.

[49] C. Geiger, G. Frosio, and O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Report to the European Parliament's Committee on Legal Affairs* (Brussels: European Parliament, 2018), at 6. See also C. Geiger, G. Frosio, and O. Bulayenko, 'The EU Commission's Proposal to Reform Copyright Limitations: A Good but Far Too Timid Step in the Right Direction', *European Intellectual Property Review* 40 (2018), 4–15, at 6.

[50] Copyright Act of 9 September 1965 (*Federal Law Gazette* I, 1273), as last amended by Article 1 of the Act of 28 November 2018 (*Federal Law Gazette* I, 2014), Article 60(d).

[51] Ibid.

[52] Ibid.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

in 2016 allowing reproduction, storage, and communication of 'files created in the course of TDM research activities'.[53] The reproduction must be from lawful sources.[54] Third, the UK statute provides for a right to make a copy of a work 'for computational analysis of anything recorded in the work', but prohibits dealing with the copy in other ways and makes contracts that would prevent or restrict the making of a copy for the purpose stated above unenforceable.[55] Fourth and finally, the Japanese statute contains an exception for the reproduction or adaptation of a work to the extent deemed necessary for 'the purpose of information analysis ("information analysis" means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information)'.[56]

The examples in the previous paragraph demonstrate a similar normative underpinning, namely a policy designed to allow TDM of the data contained in copyright works. They disagree on the implementation of the policy, however. Based on those examples, the questions that policymakers considering enacting an explicit TDM exception or limitation should include

1. whether the exception applies to only one (reproduction) or all rights (including adaptation/derivation);
2. whether contractual overrides are possible;
3. whether the material used should be from a lawful source;
4. what dissemination of the data, if any, is possible; and
5. whether the purpose of TDM is non-commercial.

The answers to all five questions can be grounded in a normative approach, but they should be set against the backdrop of the three-step test, which, as explained later, is likely to apply to any copyright exception or limitation.

As to the first question, if allowing TDM is seen as a normatively desirable goal, then the right holder should not be able to use one right fragment in the bundle of copyright rights to prevent it. In an analysis of the rights involved, Irini Stamatoudi came to the conclusion that right fragments beyond reproduction and adaptation were much less relevant.[57] Still, it would seem safer to formulate the exception or

[53] Geiger et al., note 49, at 830.
[54] Law No. 2016-1231§ for a Digital Republic and Article L122-5 of the Intellectual Property Code.
[55] Added by the Copyright and Rights in Performances (Research, Education, Libraries and Archives), Regulations 2014, 2014 No 1372, available at www.legislation.gov.uk/uksi/2014/1372/regulation/3/made.
[56] Copyright Law of Japan (translated by Y. Oyama et al.), at Article 47 *septies*, available at www.cric.or.jp/english/clj/doc/20161018_October,2016_Copyright_Law_of_Japan.pdf.
[57] I. A. Stamatoudi, 'Text and Data Mining', in I. A. Stamatoudi (ed), *New Developments in EU and International Copyright Law* (Deventer: Wolters Kluwer, 2016), 262–282.

limitation as a non-infringing *use*, as for example in section 107 (fair use) of the US Copyright Act.[58]

Second, for the same reason, contractual overrides should not be allowed. One can hardly see how they can be effective unless perhaps there was only one provider of TDM for a certain type of work. Even if a provision against contractual overrides was absent from the text of the statute, the restriction could be found inapplicable based on principles of contract law.[59]

Third, the lawful source element contained in French law is facially compelling. It seems difficult to oppose a requirement that the source of the data be legitimate. There are difficulties in its application, however. First, it is not always clear to a *human* user whether a source is legal or not; the situation may be even less clear for a machine. Second, and relatedly, if the source is foreign, a determination of its legality may require an analysis of the law of the country of origin, as copyright infringement is determined based on the lex loci delicti – and this presupposes a determination of its origin (and foreignness) to begin with. Perhaps a requirement targeting sources that the user *knows or would have been grossly negligent in not knowing* were illegal might be more appropriate.

The last two questions on the list are somewhat harder. Dissemination of the data, if such data includes copyright works, could be necessary among the people interested in the work. German law makes an exception for a 'limited circle of persons for their joint scientific research', and 'third persons for the purpose of monitoring the quality of scientific research'.[60] This is a reflection of a scientific basis of the exception, which includes project-based work by a limited number of scientists and monitoring by peer reviewers. This would not allow the use of TDM to scan libraries of books and make snippets available to the general public, as Google Books does, for example. An interpretation of the scope of the exception might depend on whether the use is commercial, which in turn might vary according to the definitional approach taken: is it the commercial nature of the *entity* performing the TDM that matters, or the specific use of the TDM data concerned (i.e., is that specific use monetized)?

The EU was considering a new, mandatory TDM exception as part of its digital copyright reform efforts.[61] Article 3, which contains the proposed TDM exception, has been the focus of intense debates. The September 2018 (Parliament) version of the proposed TDM exception maintained the TDM exception for scientific

[58] The US Copyright Act reads in part as follows: 'the fair use of a copyrighted work . . . is not an infringement of copyright'. US Copyright Act of 1976, 17 U.S.C. §§ 101–810 [hereinafter: US Copyright Act].

[59] See for example Lucie Guibault's detailed analysis of the possible application of the German *Sozialbindung* principle in this context. L. Guibault, *Copyright Limitations and Contracts: An Analysis of the Contractual Overridability of Limitations on Copyright* (The Hague/London/Boston: Kluwer Law International, 2002), at 224–225.

[60] See Copyright Act of 9 September 1965, note 50.

[61] Geiger et al., note 49, at 832–833. The research for this part of the chapter was completed.

research proposed by the commission but adds an optional exception applicable to the private sector, not just for the benefit of public institutions and research organizations.[62] Members of the academic community have criticized the narrow scope of the commission's proposed exception, which the Parliament's amendments ameliorated.[63] The European Copyright Society opined that 'data mining should be permitted for non-commercial research purposes, for research conducted in a commercial context, for purposes of journalism and for any other purpose'.[64] The final text of Article 3 in the now adopted directive states that EU member states must provide for an exception in their domestic laws for 'reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access',[65] as well as for 'reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining'.[66]

One should note, finally, that when a technological protection measure (TPM) or 'lock' such as those protected by Article 11 of the 1996 WIPO Copyright Treaty, is in place preventing the use of data contained in copyright works for TDM purposes, the question is whether a TDM exception provides a 'right' to perform TDM and thus potentially a right to circumvent the TPM or obtain redress against measures designed to restrict it.[67] This might apply to traffic management (e.g. throttling) measures used to slow the process down. Those questions are worth pondering, but they are difficult to answer, especially at the international level.[68]

III The Three-Step Test

The three-step test sets boundaries for exceptions and limitations to copyright rights. The original three-step test is contained in Article 9(2) of the Berne Convention. Instead of enumerating acceptable exceptions and limitations, Berne negotiators

[62] The Parliamentary version and the commission's proposal are compared in amendments 64 and 65 of European Parliament, Amendments Adopted by the European Parliament on 12 September 2018 on the Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market (COM(2016)0593 – C8–0383/2016 – 2016/0280 (COD)), OJ C [2019] 433/248.

[63] See, e.g., M. Senftleben, 'EU Copyright Reform and Startups – Shedding Light on Potential Threats in the Political Black Box', March 2017, at 9, available at https://bit.ly/2kiJgFq.

[64] European Copyright Society, 'General Opinion on the EU Copyright Reform Package', 24 January 2017, available at https://bit.ly/2k2k3jD.

[65] Article 3 of Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, OJ L (2019) 130/92.

[66] Ibid., at Article 4.

[67] WIPO Copyright Treaty, adopted in Geneva on 20 December 1996, entered into force 6 March 2002.

[68] For a brief discussion, see Geiger et al., note 49.

https://doi.org/10.1017/9781108919234.010 Published online by Cambridge University Press

decided to introduce this test which allows countries party to the convention to make exceptions to the right of reproduction (i) 'in certain special cases'; (ii) 'provided that such reproduction does not conflict with a normal exploitation of the work'; and (iii) 'does not unreasonably prejudice the legitimate interests of the author'. The test was extended to all copyright rights by the TRIPS Agreement, with the difference that the term 'author' at the end was replaced with the term 'right holder'.[69]

The test was interpreted in two panel reports adopted by the WTO Dispute Settlement Body. The first step ('certain special cases') was interpreted to mean that 'an exception or limitation must be limited in its field of application or exceptional in its scope'. In other words, 'an exception or limitation should be narrow in quantitative as well as a qualitative sense'.[70] The normative grounding to justify a TDM exception is fairly clear. Indeed, exceptions and limitations have already been introduced in major jurisdictions. A well-justified exception or limitation with reasonable limits and a clear purpose is likely to pass the first step.

The second step (interference with normal exploitation) was defined as follows: First, exploitation was defined as any use of the work by which the copyright holder tries to extract/maximize the value of their right. 'Normal' is more troublesome. Does it refer to what is simply 'common', or does it refer to a normative standard? The question is particularly relevant for new forms and emerging business models that have not, thus far, been common or 'normal' in an empirical sense. If the exception is used to limit a commercially significant market or, a fortiori, to enter into competition with the copyright holder, the exception is prohibited.[71]

Could a TDM exception be used to justify scanning and making available entire libraries of books still under active commercial exploitation? The answer as regards the full text of books is negative, as this would interfere with commercial exploitation. For books still protected by copyright *but no longer easily available on a commercial basis*, the absence of active commercial exploitation would likely limit the impact of the second step, however, subject to a caveat. Some forms of exploitation are typically done by a third party under licence and do not need any active exploitation *by the right holder*. For example, a film studio might want the right to make a film out of a novel no longer commercially exploited. That may in turn

[69] Article 13 TRIPS. The test is now used as the model for exceptions to *all copyright rights* in TRIPS; Article 10(1) and (2) WIPO Copyright Treaty; Article 16(2) WIPO Performances and Phonograms Treaty, adopted on 20 December 1996; Article 13(2) Beijing Treaty on Audiovisual Performances, adopted 24 June 2012; and Article 11 of the Marrakesh Treaty to Facilitate Access to Published Works for Persons Who Are Blind, Visually Impaired or Otherwise Print Disabled, adopted 27 June 2013. Interestingly, in TRIPS, it is also the test for exceptions to industrial design protection (Article 26(2)) and patent rights (Article 30).

[70] Panel Report, United States – Section 110(5) of the US Copyright Act (US – Section 110(5) Copyright Act), WT/DS160/R, adopted 15 June 2000, at 6.109 (emphasis added and citations omitted). The second case was decided in Panel Report, Canada – Patent Protection of Pharmaceutical Products, WT/DS114/R, adopted 17 March 2000.

[71] P. Goldstein, *International Copyright: Principles, Law, and Practice* (Oxford: Oxford University Press, 1998), at 295.

generate new demand for the book. This is still normal exploitation. One must be careful in extending this reasoning too far, for example, by assuming that every novel will be turned into a movie.

One way to pass the second step is for a TDM exception to allow limited uses that do not demonstrably interfere with commercial exploitation, such as those allowed under the German statute. Another example is the use of 'snippets' from books scanned by Google for its Google Books project, which was found to be a fair use by the US Court of Appeals for the Second Circuit. This is important not just as a matter of US (state) practice but because at least the fourth US fair use factor ('the effect of the use upon the potential market for or value of the copyrighted work') is a market-based assessment of the impact of the use resembling the three-step test's second step.[72] The Second Circuit noted that this did not mean that the Google Books project would have *no* impact, but rather that the impact would not be meaningful or significant.[73] It also noted that the type of loss of sale created by TDM 'will generally occur in relation to interests that are not protected by the copyright. A snippet's capacity to satisfy a searcher's need for access to a copyrighted book will at times be because the snippet conveys a historical fact that the searcher needs to ascertain'.[74] In the same vein, one could argue that the level of interference required to violate the second step of the test must be significant and should be a use that is relevant from the point of view of commercial exploitation.

The third step (no unreasonable prejudice to legitimate interests) is perhaps the most difficult to interpret. What is an 'unreasonable prejudice', and what are 'legitimate interests'? Let us start with the latter. 'Legitimate' can mean sanctioned or authorized by law or principle. Alternatively, it can just as well be used to denote something that is 'normal' or 'regular'. The WTO Panel Report concluded that the combination of the notion of 'prejudice' with that of 'interests' pointed clearly towards a legal-normative approach. In other words, 'legitimate interests' are those that are protected by law.[75] Then, what is an 'unreasonable' prejudice? The presence of the word 'unreasonable' indicates that *some level or degree* of prejudice is justifiable. Hence, while a country might exempt the making of a small number of private copies entirely, it may be required to impose a compensation scheme, such as a levy, when the prejudice level becomes unjustified.[76] The WTO panel concluded that 'prejudice

[72] The fourth fair use factor contained in the US Copyright Act reads as follows: 'the effect of the use upon the potential market for or value of the copyrighted work . . .'.

[73] *The Authors Guild* v. *Google, Inc.*, 804 F.3d 202 (2d Cir, 2015), cert. denied 136 S.Ct. 1658.

[74] Ibid.

[75] Panel Report, note 70, at paras. 6.223–6.229. In para. 6.224, the Panel tried to reconcile the two approaches: '[T]he term relates to lawfulness from a legal positivist perspective, but it has also the connotation of legitimacy from a more normative perspective, in the context of calling for the protection of interests that are justifiable in the light of the objectives that underlie the protection of exclusive rights'.

[76] WIPO, *Records of the Intellectual Property Conference of Stockholm: June 11 to July 14, 1967*, Vol. 1 (Geneva: WIPO, 1971), at 1145–1146.

to the legitimate interests of right holders reaches an unreasonable level if an exception or limitation causes or has the potential to cause an unreasonable loss of income to the copyright holder'.[77] Whether a TDM exception is liable to cause an unreasonable loss of income to copyright holders is analytically similar to the second step of the test as interpreted by the WTO panels. It is not, however, identical: The owner of rights in a work no longer commercially exploited may have a harder case on the second step. It is not unreasonable, however, for a copyright holder, to expect some compensation for some uses of a protected work even if it is not commercially exploited. For example, the owner of rights in a novel may expect compensation for the republication by a third party or translation of the book. The major difference between the second and third step as interpreted by the two WTO dispute-settlement panels in this regard is that the third step condition may be met by compensating right holders. This could allow the imposition of a compulsory licence for specific TDM uses that overstep the boundary of free use – for example, to make available significant portions of, or even entire, protected works that are no longer commercially exploited subject to a series of conditions such as the existence of any plan or preparation by the right holder to exploit the work.

D CONCLUSION

Multilateral trade rules, such as the General Agreement on Tariffs and Trade (GATT) 1947 began as an effort to facilitate trade in goods by removing tariff and non-tariff barriers. In 1995, with the establishment of the WTO, this was extended to services and IP protection. IP is perhaps the odd man out, as GATT Article XX considers IP as not much more than an acceptable barrier to trade. Moreover, IP is often not traded per se but rather embedded in a good or service. Data is arguably a new area of trade, as data, especially big data corpora and the inferences that can be derived from their analysis by AI machines, have become a commodity in themselves, but with special features, including the fact that many corpora are based on personal data.[78] Given its trajectory as a multilateral organization that addresses all main areas of trade, it would be normal for the WTO to extend its normative reach in trade in data. As it does so, it will need to see whether the rules contained in the TRIPS Agreement are up to the task of supporting the data economy, which must begin by a massive data gathering and analysis phase, as the GATT did when preparing the TRIPS Agreement. In this chapter, I offered a few suggestions on areas in which it could shine its analytical spotlight to illuminate a path for future negotiations.

[77] Panel Report, US – Section 110(5) Copyright Act, note 70, at para. 6.229.

[78] See S. Yakovleva, 'Should Fundamental Rights to Privacy and Data Protection Be a Part of the EU's International Trade "Deals"?', *World Trade Review* 17 (2018), 477–508, at 478; also Chapter 1 in this volume.