Direitos autorais e mineração de dados e textos no combate à Covid-19 no Brasil

Copyright and data and text mining in the fight against Covid-19 in Brazil

Allan Rocha de Souza[a, *]

Luca Schirru[b]

Miguel Bastos Alvarenga[c]

RESUMO: A explosão da pandemia de COVID-19 intensificou a importância das técnicas e ferramentas de mineração de dados e textos (TDM), as quais estão por trás de diversas aplicações essenciais ao combate ao SARS-CoV-2, desde o monitoramento médico e da expansão da doença ao desenvolvimento de vacinas. Nesse cenário, indagamos como a pandemia evidencia a importância dos instrumentos de TDM na inovação científica e tecnológica, assim como os efeitos do atual sistema de proteção por direitos autorais sobre as bases de dados e o desenvolvimento dessas tecnologias, que dependem fortemente do acesso e circulação aberta de informação. Para tanto, fazemos uso de pesquisa bibliográfica e documental, centrada nos casos do observatório de COVID-19 da Johns Hopkins University, do projeto NextStrain e sistemas de inteligência artificial. Primeiramente, apresentamos suscintamente as tecnologias de mineração de dados e textos, bancos de dados e aprendizado de máquina, suas aplicações e importância para a inovação científica e tecnológica. Em seguida, discutimos o papel do direito autoral sobre bases de dados e os obstáculos para o desenvolvimento de pesquisas e tecnologias intensivas em dados. Concluímos que a atual proteção sobre bancos de dados por direito autoral cria empecilhos ao acesso e uso de dados e para a pesquisa, e que a promoção das limitações e exceções, especialmente para mineração de textos e dados e desenvolvimento de pesquisas, é crucial para o desenvolvimento científico e inovação tecnológica, e ainda mais especificamente para o sucesso do combate a esta e outras pandemias.

Palavras-chave: Direito Autoral; Banco de Dados; Mineração de Dados e Textos; Limitações e Exceções; COVID-19.

ABSTRACT: The explosion of the COVID-19 pandemic has intensified the importance of text and data mining techniques and tools (TDM), which serve as basis for several applications involved in the combat against the SARS-CoV-2 virus, from the monitoring of medical cases and disease expansion to vaccine development. Under this scenario, we ask how the pandemic highlights the importance of TDM tools and the effects of the current copyright protection system on databases for the development of such technologies, which depends heavily on the access and open circulation of information. To this end, we make use of bibliographic and documental research, centered on the cases of the COVID-19 observatory at Johns Hopkins University, the NextStrain project, and artificial intelligence systems. Firstly, we conceptualize data and text mining technologies, databases and machine learning, their applications and importance for scientific and technological innovation. Next, we discuss the role of copyright on databases and the barriers it imposes for the development of research and data-intensive technologies. We conclude that the current protection of databases by copyright creates obstacles to data access and use for research purposes, and that the promotion of limitations and exceptions, especially for data and

[a] Programa de Pós-Graduação em Políticas Públicas, Estratégias e Desenvolvimento, Instituto de Economia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. Curso de Direito, Instituto Três Rios, Universidade Federal Rural do Rio de Janeiro, Três Rios, RJ, Brasil.

[b] Escola de Direito e Ciências Sociais, Universidade Positivo, Curitiba, PR, Brasil.

[c] Núcleo de Pesquisa em Direitos Fundamentais, Relações Privadas e Políticas Públicas, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.

[*] Correspondência para/*Correspondence to*: Allan Rocha de Souza. E-mail: allan@rochadesouza.com.

Recebido em/*Received*: 15/08/2020; Aprovado em/*Approved*: 29/12/2020.

text mining and for research purposes, is crucial for scientific development and technological innovation and, more specifically, for the success of the fight against this and other pandemics.

Keywords: Copyright; Database; Limitations and Exceptions; Data and Text Mining; COVID-19.

INTRODUÇÃO

A abundância contemporânea de dados é reflexo direto da penetrabilidade social (CASTELLS, 1999) das tecnologias de informação e comunicação (TICs), na medida em que cidadãos e instituições, uma vez conectados às redes digitais, passam também a ser produtores constantes e regulares de grandes volumes de dados. Ao mesmo tempo, estas tecnologias, ao trazerem um contínuo crescimento da capacidade de coleta, processamento e armazenamento (DREX, HILTY *et al*, 2019), um aumento da velocidade de conexão à rede mundial de computadores e de transmissão, uma ampliação do volume e diversidade de fontes e tipos de dados, em conjunto passaram a permitir a análise dessa vastidão de dados em tempo real, de forma cada vez mais completa e detalhada (DEAN, 2014, p. 8-9, 11-14; MARTENS, 2018, p. 6-7; PINHEIRO e TIGRE, 2019). Embora carentes de significados, valor e utilidade intrínsecos quando vistos isoladamente, se devidamente processados por ferramentas adequadas, podem ser convertidos em informações úteis e efetivo conhecimento que tornam possível encontrar caminhos novos para compreender – ou mesmo descobrir – determinados fenômenos.

Este conjunto de processos relacionados ao potencial e utilidade dos dados foi abruptamente descortinado à sociedade com a evolução da pandemia causada pelo novo coronavírus. O acompanhamento da evolução da epidemia e o desenvolvimento de respostas eficazes nos seus aspectos clínicos e sociais carece de dados precisos, dinâmicos, variados e volumosos, que necessitam ser obtidos de múltiplas fontes, processados, agregados e disponibilizados para serem utilizados. Sua própria utilização resulta em novos dados e informações que são realimentados no sistema, assegurando sua atualidade. No entanto, o próprio volume e variedade de fontes e tipos impõem a aplicação de tecnologias de mineração de dados - as quais são operadas, principalmente, por meio de processos automatizados.

Este é, por exemplo, o caso do sistema desenvolvido, aprimorado e disponibilizado pela John Hopkins University (DONG, DU e GARDNER, 2020), que se tornou a principal referência mundial de informações em tempo real sobre a pandemia e permite a reunião, organização e disponibilização pública dos dados epidêmicos sobre a COVID-19 (JOHN HOPKINS UNIVERSITY, 2020), sendo notável a diversidade e amplitude das fontes de dados utilizados para compor o resultado e permitir a visualização do todo e as consequentes inferências.

Igualmente relevante é a mineração de textos, por meio da qual se busca extrair os dados e informações relevantes de um determinado conjunto de textos, permitindo com isso uma visão panorâmica do estágio dos debates e conhecimento sobre determinado tema. No Brasil, pesquisadores da Universidade de São Paulo desenvolveram uma ferramenta de mineração de dados e textos, batizada de "*Websensors*", que "extrai dados de notícias para obter informação sobre o que aconteceu, quando e onde" (WEBSENSORS, 2020). O principal objetivo do projeto é saber se os modelos preditivos podem ser ajustados com base em dados suplementares extraídos de notícias e, para tal, os pesquisadores aplicam a técnica de mineração de dados para "identificar os eventos que estão ocorrendo em cada país e ajustar as projeções para o Brasil." (ARANTES, 2020).

Indubitável, nos parece, que a mineração de dados e textos já está sendo largamente empregada e que seu potencial para a inovação, essencialidade para a pesquisa e o avanço da ciência e conhecimento restam comprovados, especialmente diante das circunstâncias excepcionais do momento sendo vivido. Esta prática, entretanto, esbarra no sistema de apropriação e atribuição de exclusividade promovido pelos direitos autorais. Isso porque, embora os dados *per se* não sejam abarcados pela proteção - a não ser que sejam em si expressões criativas da literatura, das artes ou das ciências -, os bancos de dados, os *datasets*, enquanto conjunto reunido, original, o são, como expresso no artigo 7º, XIII da Lei de Direitos Autorais (BRASIL, 1998), e, como tal, estão reservados aos titulares dos bancos de dados os direitos de autorizar ou proibir sua reprodução, modificação, extração e distribuição de sua base e dos resultados destas operações (art. 87 da LDA), bancos estes que são hoje especialmente digitais e seus titulares são, em grande medida, os investidores e organizadores. Esta restrição é ainda aprofundada diante da inexistência de uma limitação aos direitos autorais que equacione o problema.

Dita tensão no seio dos direitos autorais não é novidade, pois as mesmas questões ressurgem a todo instante em que se faça necessária a compatibilização entre a exclusividade inerente à proteção de direitos de propriedade intelectual e, de modo geral, o direito de acesso, que consubstancia uma série de direitos e interesses, tais quais informação, conhecimento, educação, pesquisa, cultura, inovação, todos constitucionais e vários de índole fundamental. A discussão sobre o papel, fundamentos e extensão das limitações e exceções (L&E) aos direitos autorais é internacional, tendo se intensificado desde o início deste século. No Brasil, o ponto crucial foi quanto à adequada interpretação das L&E, cuja pacificação se deu por meio da consolidação pelo Superior Tribunal de Justiça (STJ), em uma série de decisões inauguradas pelo Recurso Especial 964.404/11 e que resultaram no Enunciado 115 da III Jornada de Direito Comercial do Conselho Federal de Justiça, a partir do entendimento sistemático e constitucionalizado da legislação especial, de que os casos expressos na LDA são apenas exemplos da ponderação feita pelo legislador ordinário, e não a totalidade de situações em que a utilização de obras permitidas sem necessidade de autorização prévia ou remuneração do autor é juridicamente possível (BRASIL, 2011, 2019).

Assim, o objetivo restrito deste artigo é, no contexto de pandemia e uso intensivo de dados como essencial para o seu enfrentamento, avaliar se, sob a estrutura atual do direito autoral, é possível a satisfação de sua função social no que se refere ao direito à pesquisa, a partir da perspectiva do interesse público em se ter acesso a informações e da liberdade de investigação científica e produção de conhecimentos capazes de impactar positivamente no combate ao coronavírus, que pode ser viabilizado por meio da mineração de dados e textos.

Recorremos, metodologicamente, à análise documental e bibliográfica, ilustrando as questões a partir de casos concretos. No desenvolvimento da questão, são inicialmente tratadas as características centrais da mineração de dados e textos e da proteção das bases de dados. Considerando seu destaque no processamento de grandes volumes de dados, serão trazidas algumas considerações técnicas a respeito do funcionamento dos sistemas de inteligência artificial ("IA"), notadamente aqueles baseados em *machine learning* e suas diferentes formas de treinamento. Finalmente, será analisada a proteção autoral sobre as bases de dados sob a perspectiva do direito à pesquisa por meio das práticas de mineração de dados e textos no contexto da pandemia.

MINERAÇÃO DE DADOS E A PESQUISA CIENTÍFICA

A posse e uso intensivo de grandes quantidades de dados é, hoje, vista por governos, negócios e entidades não-comerciais como fonte de informação valiosa para a introdução de melhorias e inovações nas mais diversas áreas. A análise de dados possui inúmeras aplicações, sendo utilizada para apontar caminhos ainda inexplorados dentro da pesquisa científica e prever futuras descobertas, além de auxiliar na gestão de governos e empresas, no desenvolvimento de novas tecnologias, no aprimoramento de sistemas de segurança e, no campo da saúde, introduzir melhores diagnósticos médicos e fazer uso de técnicas de sequenciamento genético para melhor compreender diversas doenças (CHEN, CHIANG e STOREY, 2012, p. 1168-1172; DEAN, p. 3-5). Estas técnicas, inclusive, têm operado de forma vital nas pesquisas sobre o comportamento do novo coronavírus e o desenvolvimento de vacinas e medicamentos para tratamentos contra a COVID-19 (BRASIL, 2020).

Um exemplo de projeto que atua com estes processos de sequenciamento é o NextStrain, uma iniciativa internacional dedicada a monitorar e divulgar a evolução e disseminação de diversos agentes patológicos, tais como os vírus do Ebola e da Zika. O grupo utiliza programas desenvolvidos com o propósito de cruzar e interpretar dados genéticos obtidos de diversas fontes públicas, a fim de gerar, em tempo real, um mapa interativo sobre a evolução genética e da propagação de populações de vírus, bactérias e outros patógenos. O objetivo do site é fornecer informações para virologistas, epidemiologistas, funcionários de saúde pública e cientistas que auxiliem no entendimento epidemiológico e na melhoria das medidas de resposta a epidemias (NEXTSTRAIN, 2020).

A viabilidade desses projetos, entretanto, depende não só da posse de grandes quantidades de dados, mas também do desenvolvimento de ferramentas de coleta e análise cada vez mais sofisticadas (DEAN, 2014, p. 4-5). Isto ocorre porque dados brutos, quando considerados de forma isolada, não possuem sentido ou utilidade intrínsecos, fora do contexto em que foram concebidos e dos propósitos para os quais foram gerados, pois a informação que se busca a partir desses dados só se revela quando estes são devidamente contextualizados e interpretados (ROWLEY, 2007, p. 170-171).

Isto se torna possível através da *mineração de dados* ou *data mining*, também chamada de *mineração de dados e textos* (MDT), ou *text and data mining* (TDM), definida como sendo o conjunto de técnicas dedicadas a encontrar padrões de interesse a partir de grandes quantidades de dados, em um complexo processo de coleta e análise de informações. Este processo começa pela coleta dos dados, normalmente oriundos de diversas fontes, e por sua limpeza (retirada de ruídos e dados inconsistentes). O conteúdo remanescente é, então, extraído e reunido em um único local, para que os dados relevantes sejam selecionados e transformados em formato inteligível para análise. A partir daí, buscam-se correlações e padrões entre os dados, de onde são extraídas diversas informações. O analista ou pesquisador avalia, então, quais informações são relevantes para o fim pretendido e, por fim, as apresenta ao usuário para que, então, possam ser utilizadas (HAN, PEI e KAMBER, 2011, p. 7; KELLEHER e TIERNEY, 2018, p. 241-242; KROENKE *et al*, 2016, p. 493).

Todas estas ações implicam no acesso a diferentes fontes (em especial bancos de dados e armazéns de dados, como veremos mais adiante), bem como a coleta, extração, reprodução, adaptação, transformação, armazenamento e exposição do material ali contido, além do descarte de conteúdo irrelevante. Pode-se, inclusive, reutilizar o conteúdo extraído e produzido na análise em um momento posterior, seja

de forma isolada ou em combinação com dados novos (GEIGER, FROSIO e BULAYENKO, 2018, p. 6-7).

Tudo isto demanda uma arquitetura composta não só pelas próprias fontes dos dados extraídos, como também os servidores responsáveis pela busca dos dados relevantes, a base de conhecimento que informa os parâmetros a serem utilizados no processamento, busca e avaliação de padrões, o mecanismo de mineração (que contém módulos responsáveis por realizar as correlações e previsões), um módulo de avaliação dos padrões e, finalmente, a interface com o usuário, pela qual o usuário interage com o sistema e obtém as informações que são apresentadas ao final da análise (HAN, PEI e KAMBER, 2011, p. 7-9).

A presença das fontes como parte necessária nesta estrutura também torna implícito que, a despeito de sua importância, a mineração de dados e textos é apenas uma parte do processo de análise - para que ela seja viável, é necessário antes extrair os dados de algum lugar e armazená-los em outro, onde serão devidamente organizados, filtrados e preparados. Este "lugar" é um registro a partir do qual os dados podem ser preservados e acessados de uma maneira mais organizada e permitir o cruzamento de várias informações com mais eficácia: o *banco (ou base) de dados digital*.

BANCOS DE DADOS E A APROPRIAÇÃO DOS CONJUNTOS

Kroenke *et al* (2016, p. 3-32) definem bancos de dados como sendo conjuntos de tabelas e outras estruturas criadas a partir de informação coletada de diversas fontes. Na prática, eles servem para que os dados coletados sejam preservados e acessados de uma maneira mais organizada, permitindo o cruzamento de várias informações com mais eficácia. Esta organização inclui a criação e armazenamento de metadados[1], bem como índices e descrições dos aplicativos usados. Esses agrupamentos ou coleções de dados e metadados fazem parte de um sistema mais amplo, que inclui programas de computador dedicados a criar, processar e administrar essas bases – os sistemas de gerenciamento de bancos de dados (SGBD) –, os aplicativos que servem de interface entre usuários e o SGBD, e os próprios usuários.

Esses bancos de dados podem assumir duas formas. A primeira e a mais comum delas é a relacional, a qual, conforme o nome indica, trabalha com tabelas que representam relações atributo-entidade e é comumente gerenciada por *softwares* que operam em linguagens como o SQL (daí o outro nome usado para se referir a essas bases: bancos de dados SQL). Essas tabelas são tipicamente representadas através de linhas (referentes a todos os atributos de uma mesma entidade, ex.: dados a respeito de um indivíduo) e colunas (que indicam um mesmo atributo em indivíduos distintos, como idade, altura ou peso) (HAN, PEI e KAMBER, 2011, p. 10-12; KELLEHER e TIERNEY, 2018, p. 7-8; KROENKE et al, 2016, p. 3-17).

A realidade de se operar com dados em volume e variedade crescentes, contudo, tem levado ao desenvolvimento de bancos de dados não-relacionais (ou NoSQL), nos quais os dados são representados não em listas, mas como objetos separados, com atributos particulares. Isto dá uma flexibilidade maior ao armazenamento, uma vez que cada

[1] Metadados são dados a respeito da estrutura e propriedades de outros dados, ou "dados sobre dados", como o horário e data em que um certo arquivo foi coletado (KELLEHER e TIERNEY, 2018, p. 243). Também podem se referir à estrutura do próprio banco de dados, como os nomes e propriedades de tabelas, linhas e colunas (e a que tabela estas últimas pertencem) (KROENKE et al, 2018, p. 19).

entidade pode existir de forma independente, sem a necessidade de se adequar a uma lista ou tabela. Trata-se de um formato útil para lidar com conjuntos de dados que possuem características muito diversas (como arquivos de áudio, vídeo e texto) e que, portanto, dificilmente seriam compatíveis com um banco de dados relacional. Contudo, ainda é necessário converter esses dados para um formato legível – ou seja, aplicando etiquetas, ou "*tags*", para cada atributo relevante -, a fim de tornar possível o resgate e o tratamento posterior do material coletado (KELLEHER e TIERNEY, 2018, p. 9-10).

Normalmente, os bancos de dados surgem duas vezes no processo de coleta e análise de dados, com propósitos distintos: inicialmente, são usados para simples armazenamento, consulta e outras atividades operacionais, servindo como pontos de origem comuns de muitos dos elementos que são filtrados e extraídos na mineração. Já durante o processo de limpeza, integração e transformação, estes elementos são comumente reunidos dentro dos armazéns de dados, ou *data warehouses*, que são bases de dados criadas com o propósito específico de concentrar o conteúdo de outros repositórios em um só local para que ele possa ser modelado e analisado, além de armazenar as informações resultantes do processo de mineração de dados (HAN, PEI e KAMBER, 2011, p. 105-107; KELLEHER e TIERNEY, 2018, p. 8-9; KROENKE et al, 2016, p. 492-494).

Desta maneira, seja para fins operacionais ou para uso como *data warehouses*, a montagem, aquisição e manutenção de grandes bases de dados constituem um componente vital dentro da estrutura que é construída para que a mineração de dados seja possível. Entretanto, quando tratamos de grandes volumes de dados, isto demanda uma capacidade de processamento que ultrapassa os limites humanamente possíveis, tornando inviável operar uma grande base de dados ou minerar seu conteúdo sem assistência. Assim, cresce o uso de sistemas de inteligência artificial (IA) para auxiliar nessas tarefas ou executá-las inteiramente, através de um processo chamado de *aprendizado de máquina*, ou *machine learning*.

INTELIGÊNCIA ARTIFICIAL E TREINAMENTO DE MÁQUINAS

O grande volume de dados hoje produzido e a existência de tecnologias capazes de promover o processamento de tal volume de dados foram elementos fundamentais para a sofisticação dos resultados obtidos e para a opção pela adoção de tecnologias de IA nas mais diversas áreas (DREX, HILTY *et al*, 2019, p. 4), inclusive no combate ao Coronavírus.[2] Publicações recentes propõem o uso de sistemas de IA no desenvolvimento de modelos preditivos do crescimento de casos (ABHARI, MARINI, CHOKANI, 2020), desenvolvimento de medicamentos (HO, 2020) e novas formas de diagnosticar a COVID-19, seja através da análise de informações (BATISTA *et al*, 2020), ou de imagens (WANG *et al*, 2020).

A relação entre dados e IA é ambivalente. De um lado, sistemas de IA são necessários para o processamento de grandes volumes de dados (SAMUEL, 1959; ALVARENGA, 2019). De outro, os dados, sua seleção e classificação são fundamentais para a aprendizagem de um algoritmo e, por conseguinte, para que se obtenha o resultado

[2] Mais informações a respeito das recentes pesquisas envolvendo tecnologias de IA e a sua aplicação no combate ao Coronavírus podem ser encontradas na Seção "Telemedicina e Inteligência Artificial" do "Observatório de Tecnologias Relacionadas ao Covid-19" do INPI (Disponível em: https://www.gov.br/inpi/pt-br/servicos/patentes/tecnologias-para-covid-19/Telemedicina) e no "Observatório COVID-19" da Fiocruz (Disponível em: https://portal.fiocruz.br/observatorio-covid-19).

pretendido de sua operação (DREX, HILTY *et al*, 2019). Se valendo do trabalho de Russel e Norvig (2013, p. 25), é possível afirmar que os dados, hoje, representam elementos centrais no que concerne à operação e treinamento de um sistema de IA, compartilhando do protagonismo que antes era exclusivo do algoritmo.

Se o *machine learning* representa o aprendizado de máquinas, este pode se dar de determinadas maneiras, dentre as quais se destacam o aprendizado supervisionado, não-supervisionado e o aprendizado por reforço (DREX, HILTY *et al*, 2019, pp.4-8; HAYKIN, 2001). Este primeiro envolve uma maior intervenção por parte do ser humano, que "rotula" os dados de treinamento daquela determinada rede neural, o que irá viabilizar um resultado mais preciso (DREX, HILTY *et al*, 2019). Para melhor ilustrar o funcionamento de um aprendizado supervisionado, transcreve-se o exemplo ilustrado por Drexl, Hilty *et al* (2019, p.5):

> Um modelo de aprendizado de máquina pode ser usado para reconhecer gatos em imagens. No caso de aprendizado de máquina supervisionado, o modelo é treinado em um conjunto de dados contendo dados rotulados (ou seja, cada imagem é acompanhada pela informação de que existe um gato na imagem), permitindo que ela se torne mais precisa. Uma vez concluído o treinamento, o modelo deve, em princípio, ser capaz de reconhecer, a partir de uma imagem não identificada, se um gato aparece nela (saída). Esse modelo pode finalmente ser implementado em um carro autônomo, permitindo, por exemplo, frear quando confrontado com um gato (aplicativo).[3]

O aprendizado não-supervisionado, por sua vez, não demanda que os dados sejam rotulados em um primeiro momento, o que não afasta a participação humana no que concerne o processo como um todo, uma vez que se faz fundamental ao final deste, quando da análise de seu resultado (DREX, HILTY *et al*, 2019, p.8). Nessa modalidade de aprendizado, ao invés de se valer de um conjunto de dados devidamente rotulado, o sistema é exposto a um grande conjunto de dados sem qualquer identificação, sobre o qual, mediante a verificação de padrões, passa a dividi-los em grandes conjuntos por meio, por exemplo, do que é denominado *clusterização* (DREX, HILTY *et al*, 2019, p.8). Importante ressaltar que a opção por uma forma de aprendizado não anula a outra, uma vez que "diversos métodos [...] usam aprendizado não-supervisionado adicional para facilitar o aprendizado supervisionado".[4]

Por fim, o aprendizado por reforço, ao contrário das duas modalidades anteriores de aprendizado, "não depende de conjuntos de dados pré-existentes, mas reúne dados de simulações ou jogos. O algoritmo determina as regras com base no feedback contínuo das ações executadas durante o treinamento" (DREXL, HILTY *et al*, 2019).[5] Um dos exemplos citados por Drexl, Hilty *et al* (2019, p. 8) é o caso do jogo Go, em que

---

[3] Grifos do original. Tradução nossa. Texto original em Drexl, Hilty (et al, 2019, p. 5): "A machine learning model might be used to recognise cats in pictures. In the case of supervised machine learning, the model is trained on a data set containing labelled data (i.e., each picture is accompanied by the information whether there is a cat in the picture), allowing it to become more accurate. Once the training is completed, the model should in principle be capable of recognising from an unlabelled picture whether a cat appears in it (output). This model could finally be implemented in a self-driving car, allowing it, for instance, to brake when confronted with a cat (application)."

[4] Tradução nossa. Texto original em Schmidhuber (2015, p. 89): "The main focus of current practical applications is on Supervised Learning (SL), which has dominated recent pattern recognition contests [...]. Several methods, however, use additional Unsupervised Learning (UL) to facilitate SL [...]." Grifos do original.

[5] Tradução nossa. Texto original em Drexl, Hilty et al (2019, p. 8): "does not rely on pre-existing data sets, but rather gathers data from simulations or games. The algorithm determines the rules based on continuous feedback on the actions it takes during the training."

o algoritmo foi capaz de aprender sem qualquer *input* por parte dos programadores no que tange às regras do jogo ou estratégia, se desenvolvendo a partir da simulação de diversas partidas que jogava contra si mesmo, de onde extraiu o conhecimento capaz de aprimorar a sua forma de jogar. A eficácia desse modelo de aprendizagem é demonstrada pelo fato de que um sistema de IA chegou a vencer um ser humano em um jogo (DREX, HILTY *et al*, 2019, p.8).

Contudo, a vitória que se busca agora é outra, mais urgente e global. Com as diferentes aplicações de sistemas de IA no combate ao coronavírus, não apenas a escolha pelo(s) modelo(s) de aprendizado desejado, mas também a seleção, e robustez, das bases de dados que serão utilizadas no treinamento e teste de um determinado sistema representam fatores determinantes no desenvolvimento de uma vacina, por exemplo. Para tanto, além de profissionais competentes, são necessários muitos dados - o que pode demandar investimentos consideráveis.

DIREITOS AUTORAIS, ACESSO À INFORMAÇÃO E PESQUISA

Embora iniciativas como o NextStrain e o Repositório da Universidade John Hopkins, por razões de interesse público, façam uso de dados disponíveis publicamente e operem com uma plataforma aberta para uso por outros projetos (NEXTSTRAIN, 2020), outros agentes podem se ver incentivados a buscar a implementação de medidas que visem limitar ou proibir o acesso de seus bancos de dados por concorrentes ou, ainda, impedir que terceiros utilizem o conteúdo dessas bases sem autorização, a fim de preservar os interesses de seus titulares (STUCKE e GRUNES, 2015, p. 3). E as bases de dados são protegidas por uma estrutura técnico-jurídica que opera em diversos níveis, dentre os quais se destacam três: (i) direito autoral; (ii) medidas tecnológicas de proteção (*tecnhological protection mesaures*); (iii) e medidas anti-burla (*anti circumvention measures*).

No primeiro nível, temos a proteção aos bancos de dados (sejam eles operacionais ou data *warehouses*) por meio do direito autoral, bem como dos dados que sejam em si expressões criativas protegidas – como fotos, músicas, livros, artigos científicos e os próprios programas de computador, entre outros. Este é o caso do Brasil desde 1998, quando bancos de dados e programas de computador foram incluídos entre as obras protegidas pela Lei 9.610/98 (Lei de Direitos Autorais), artigo 7º, XII, XIII, §2º. Destaca-se, contudo, que só recebem proteção as compilações que, pela seleção ou arranjo de seu conteúdo, constituam criações intelectuais, dotadas de originalidade – aqui entendida como possuindo suficiente 'distinguibilidade', não podendo ser banal ou comum, devendo trazer algo novo por si só de forma a ser inconfundível com outras obras do mesmo gênero. Portanto, não estão cobertas por este regime as compilações cuja seleção ou organização de seu conteúdo atenda a critérios puramente técnicos ou funcionais. É o caso dos bancos de dados cujo conteúdo é coletado e armazenado de forma exaustiva e, muitas vezes, organizado segundo critérios comuns, como ordem cronológica ou alfabética (DERCLAYE, 2008, p. 45).

O segundo nível, de caráter tecnológico, se refere à implementação de medidas tecnológicas de proteção (MTP) e de ferramentas de gestão de direitos digitais (digital Rights Management, ou DRM) para controlar o acesso e uso de obras protegidas. Neste caso, os próprios mecanismos utilizados na criação e manutenção das bases de dados exercem tal papel, expandindo enormemente o escopo da proteção às bases de dados, mesmo nos casos que não satisfazem os requisitos da lei autoral – podendo, por exemplo, impedir o acesso e uso de dados não-protegidos ou obras em domínio público (BRANCO, 2011, p. 269).

Por fim, o terceiro nível é constituído por medidas anti-burla (*anti-circumvention measures*), dispositivos legais que vedam a alteração, supressão, modificação ou inutilização desses mecanismos tecnológicos, a exemplo do art. 107 da LDA. Tais disposições não se confundem com a proteção estendida aos bancos de dados ou aos software propriamente ditos, apenas protegem quaisquer dispositivos tecnológicos que previnam o acesso ou cópia não autorizados de obras protegidas – formam uma espécie de '*paracopyright*', que opera à margem do objeto central do direito autoral, mas ainda dentro do sistema (BROWN, 2003).

Esta tripla camada de proteção se torna especialmente problemática quando analisamos os procedimentos necessários para a mineração de dados: processos de coleta e análise de dados que, como já vimos, tipicamente envolvem técnicas de cópia, extração e modificação do conteúdo existente em bases de dados alheias, implicando em atos que dependeriam de autorização prévia dos titulares. Assim, no caso de os dados envolverem conteúdo protegido (como livros, fotografias, etc.), há potencial violação de direitos de reprodução, especialmente se for copiada uma parte substancial do acervo – o que é bastante comum, dado que muitos processos de TDM habitualmente visam obter o máximo de informação relevante possível. Por outro lado, caso estejamos lidando com um banco de dados original, tanto a reprodução de material relevante como o descarte de conteúdo irrelevante para a análise também podem constituir uma violação de direito autoral, já que podem replicar ou alterar a seleção ou arranjo do banco de dados de onde se extraiu o material, implicando em violação tanto do direito de reproduzir como de adaptar a obra (GEIGER, FROSIO e BULAYENKO, 2018, p. 6-7).

Esta multiplicidade de proteções possibilita impor, com respaldo dos direitos autorais, bloqueios ao acesso ou uso de conteúdo que não cumpre os requisitos para a cobertura por este regime protetivo ou que já se encontra em domínio público. À exceção de bancos de dados abertos, é difícil encontrar uma base digital que não possua uma ou mais dessas camadas de proteção, por mais que só reúna dados "brutos", sobre os quais não recai qualquer criatividade. Além disto, no campo acadêmico, é comum encontrar periódicos ou outros veículos de divulgação científica que cobram pelo acesso aos textos que divulgam, o que leva, não raro, à necessidade de estar filiado a alguma instituição capaz de negociar licenciamentos com múltiplas editoras. Tudo isto é intensificado por um contexto de perseguição às limitações de direito autoral, que existem justamente com a finalidade de equilibrar e flexibilizar a exclusividade decorrente do sistema autoral, em razão do interesse público. Prova disto é que a extensão do direito autoral para os bancos de dados, ao menos no Brasil, não veio acompanhada de uma limitação própria que estipule condições nas quais o acesso e uso de bases de dados e seu conteúdo seria permitido, mesmo que somente para fins de pesquisa científica.

Neste cenário, qualquer projeto de pesquisa que dependa da mineração de dados e não queira se sujeitar a possíveis dissabores judiciais se vê diante de duas opções: a primeira é evitar fontes proprietárias, o que pode comprometer os resultados da pesquisa em virtude de deficiências no material separado para análise; a segunda é entrar em contato com os detentores dos direitos sobre as obras e bancos de dados envolvidos e obter autorização para uso, onde se tem um labirinto burocrático extremamente demorado e dispendioso, considerando que a quantidade e variedade de dados utilizados em um projeto costumam implicar em uma vastidão de titulares diferentes. Neste último ponto, frisa-se que os pesquisadores do próprio NextStrain investem no uso de plataformas de divulgação científica abertas, sob o argumento de que o sucesso do projeto depende de análises velozes e ampla divulgação dos

resultados, algo que as práticas recorrentes de publicações científicas que usam um modelo mais proprietário de disseminação do conhecimento não seriam capazes de permitir a contento (NEXTSTRAIN, 2019).

Desta maneira, o sistema de direito autoral sobre bases de dados, com excessivos poderes aos titulares e pouca clareza, se torna uma ferramenta de imposição de interesses particulares sobre demandas públicas, impondo obstáculos à pesquisa científica e à inovação tecnológica. No caso específico dos estudos sobre a COVID-19, a questão se torna ainda mais crítica, pois os custos em recursos e tempo dedicados a processos de licenciamento representam atrasos na adoção de iniciativas essenciais para o combate à pandemia. Cabe, mais do que nunca, trazer à frente a importância da função social da propriedade e da defesa do interesse público, mediante um resgate das limitações de direito autoral e do direito à pesquisa.

CONSIDERAÇÕES FINAIS

Ao longo deste artigo, buscamos avaliar se o sistema de direito autoral, tal como estruturado hoje, é capaz de satisfazer sua função social e o interesse público, especialmente no que se refere ao direito à pesquisa. Tal indagação se faz premente diante do interesse público de acesso a informação e da necessidade de produzir e compartilhar, com liberdade, conhecimentos capazes de contribuir decisivamente para o combate ao coronavírus – algo tornado viável graças à mineração de dados e textos. Para tanto, fizemos uso de uma análise documental e bibliográfica, utilizando casos concretos como o repositório da Universidade Johns Hopkins e o projeto NextStrain para ilustrar os pontos levantados.

Concluímos que a estrutura de proteção aos bancos de dados pelo direito autoral se encontra, hoje, em um descompasso com o interesse público, especialmente no que concerne o acesso à informação e à pesquisa científica. A inexistência de dispositivos que permitam o uso de processos de mineração de dados e textos cria um ambiente de incerteza e de precariedade para o pesquisador, o que pode inviabilizar pesquisas capazes de promover resultados determinantes a cenários como o que hoje se vive. Isto se agrava quando constatamos que, cada vez mais, tais empreitadas necessitam do emprego de sistemas de inteligência artificial – os quais são capazes de processar e produzir uma vastidão de informação a velocidades humanamente inviáveis, mas que, ao mesmo tempo, não conseguem operar de maneira satisfatória sem ter acesso a uma quantidade vasta de dados disponível.

Em um momento no qual necessitamos produzir conhecimento sobre a COVID-19 da forma mais rápida e completa possível, a existência desses entraves não implica apenas em custos financeiros ou temporais, mas também limita o direito de acesso a saúde e contribui para o aumento de perdas humanas. Neste caso, a continuidade de projetos baseados em modelos de ciência aberta e seu papel no combate à pandemia tem demonstrado a necessidade de readequar a proteção autoral para que ela efetive, de fato, o seu papel principal: servir de intermediário entre os interesses dos titulares e o interesse público.

Embora a solução não esteja limitada a alterações na legislação autoral, sendo fundamental a avaliação e a (re)construção de um arcabouço institucional adequado para a pesquisa e orientado pela abertura e compartilhamento, a discussão e a inclusão de limitações que permitam a mineração de textos e dados e o exercício do direito à pesquisa são necessárias e urgente.

Como expõe Quintais (2017, p.203):

> "As limitações são ferramentas essenciais para equilibrar a exclusividade de direitos autorais com o interesse público e os direitos fundamentais. Eles possibilitam a promoção do acesso e a disseminação da cultura, educação e conhecimento. No ambiente online, elas fornecem um contrapeso necessário à expansão de direitos exclusivos tecnicamente definidos para atividades digitais fora do núcleo comercial do direito autoral."

No Brasil, a discussão sobre o papel das limitações e exceções aos direitos autorais, em especial no ambiente digital, ganhou robustez a partir da primeira década deste século. A necessidade de harmonizar os direitos autorais com outros direitos igualmente fundamentais, com a imposição de limites à exclusividade autoral para satisfação de outros direitos que, diante de diversas circunstâncias, devem ser privilegiados, reflete a obrigação jurídico normativa de satisfação da função social dos direitos autorais (SOUZA, 2006). E, como visto, desde então, decisões judiciais no Superior Tribunal de Justiça que culminaram no Enunciado 115 da III Jornada de Direito Comercial promovido pelo Conselho da Justiça Federal consolidaram o entendimento de que "as limitações de direitos autorais estabelecidas nos arts. 46, 47 e 48 da Lei de Direitos Autorais devem ser interpretadas extensivamente, em conformidade com os direitos fundamentais e a função social da propriedade estabelecida no art. 5º, XXIII, da CF/88."

E, diante dos desafios contemporâneos, o exercício da pesquisa e seu reconhecimento enquanto direito fundamental são essenciais para o avanço do conhecimento em todas as áreas do saber. E será só com o aprofundamento da harmonização com os demais direitos fundamentais que os direitos autorais serão capazes de satisfazer sua função social, alcançando o constitucionalmente garantido equilíbrio entre a proteção e o direito de acesso.

APOIO E FINANCIAMENTO

Artigo produzido no âmbito do Instituto Nacional de Ciência e Tecnologia (INCT) Proprietas e do Núcleo de Pesquisas em Direitos Fundamentais, Relações Privadas e Políticas Públicas (NUREP), com o apoio da CAPES, CNPq e FAPERJ.

REFERÊNCIAS

ABHARI, Reza S.; MARINI, Marcello; CHOKANI, Ndaona. COVID-19 Epidemic in Switzerland: Growth Prediction and Containment Strategy Using Artificial Intelligence and Big Data. *medRxiv*. 2020. Disponível em: https://doi.org/10.1101/2020.03.30.20047472

ALVARENGA, Miguel Bastos. *Mineração de dados, Big Data e Direitos Autorais no Brasil*. 2019. Dissertação (Mestrado em Políticas Públicas, Estratégias e Desenvolvimento). Instituto de Economia. Universidade Federal do Rio de Janeiro (UFRJ). Rio de Janeiro, 2019.

ARANTES, J. T. *Artificial intelligence to track news of COVID-19*. Agência FAPESP, 20 mai. 2020. Disponível em: https://agencia.fapesp.br/artificial-intelligence-to-track-news-of-covid-19/33174/. Acesso em 23 ago. 2020.

BANTERLE, F. *Data ownership in the data economy*: a European dilemma. EU Internet Law in the digital era (edited volume based on the REDA 2017 conference). Springer,

2018 (no prelo). Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3277330. Acesso em 16 jun. 2020.

BATISTA, Andre Filipe de Moraes; MIRAGLIA, Joao Luiz Miraglia; DONATO, Thiago Henrique Rizzi; FILHO, Alexandre Dias Porto Chiavegatto. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*. 2020. Disponível em: https://doi.org/10.1101/2020.04.04.20052092

BRANCO, S. V. O Domínio Público no Direito Autoral Brasileiro – Uma Obra em Domínio Público. Rio de Janeiro: Lumen Juris, 2011.

BRASIL. Conselho da Justiça Federal. *III Jornada de Direito Comercial: Enunciados aprovados em 7/6/2019*. 2019. Disponível em: https://www.cjf.jus.br/cjf/noticias/2019/06-junho/iii-jornada-de-direito-comercial-e-encerrada-no-cjf-com-aprovacao-de-enunciados/copy_of_EnunciadosaprovadosIIIJDCREVISADOS004.pdf. Acesso em 17 ago. 2020.

BRASIL. *Lei nº 9.610, de 19 de fevereiro de 1998***.** Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. 1998b. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9610.htm. Acesso em 16 jun. 2020.

BRASIL. Ministério da Saúde. *Sequenciamento do coronavírus possibilita o desenvolvimento de vacinas***.** Blog da Saúde, 16 mar. 2020. Disponível em: http://www.blog.saude.gov.br/index.php/perguntas-e-respostas/54104-confira-a-entrevista-sobre-o-sequenciamento-do-coronavirus. Acesso em 06 ago. 2020.

BRASIL. Superior Tribunal de Justiça. 3ª Turma. *Recurso Especial nº 964404/ES (2007/0144450-5)*. Recorrente: Mitra Arquidiocesana de Vitória. Recorrido: Escritório Central de Arrecadação e Distribuição (ECAD). Relator: Min. Paulo de Tarso Sanseverino. Brasília, 15 de março de 2011. Lex: Diário de Justiça Eletrônico, Brasília, v. 815, 23 mai. 2011.

BROWN, K. Digital Rights Management: Trafficking in Technology That Can Be Used to Circumvent the Intellectual Property Clause. 40 *Houston Law Review*, vol. 803, 2003, p. 803-836.

CASTELLS, Manuel. A era da informação: economia, sociedade e cultura. Vol. 1: Sociedade em rede. São Paulo: Paz e Terra, 1999.

CHEN, H; CHIANG, R. H. L.; STOREY, V. C. Business Intelligence and Analytics: from Big Data to Big Impact. *MIS Quarterly: Management Information Systems*, vol. 36 (4), pp. 1165-1188, dez. 2012.

DEAN, J. *Big Data, Data Mining, and Machine Learning*: Value Creation for Business Leaders and Practitioners. Wiley, 2014. ProQuest Ebook Central. Disponível em: http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=1687540. Acesso em 16 jun. 2020.

DERCLAYE, E. The Legal Protection of Databases: A Comparative Analysis. Edward Elgar, 2008.

DONG E.; DU H.; GARDNER L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, vol. 20 (5): 533-534. Disponível em: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext. Acesso em 23 ago. 2020.

DREXL, Josef; HILTY, Reto M.; BENEKE, Francisco; DESAUNETTES, Luc; FINCK, Michèle; GLOBOCNIK, Jure; OTERO, Begoña Gonzalez; HOFFMANN, Jörg; HOLLANDER, Leonard; KIM, Daria; RICHTER, Heiko; SCHEUERER, Stefan; SLOWINSKI, Peter R.; THONEMANN, Jannick. Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective. *Max Planck Institute for Innovation and Competition Research Paper Series – Research Paper No. 19-13.* Research Group on the Regulation of the Digital Economy. October, 2019. Disponível em: https://ssrn.com/abstract=3465577.

FUNDAÇÃO OSWALDO CRUZ. *Observatório COVID-19*: Informação para ação. 2020. Disponível em: https://portal.fiocruz.br/observatorio-covid-19. Acesso em 19 de jul de 2020, às 11:51.

GEIGER, C.; FROSIO, G.; BULAYENKO, O. *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects*. Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02, 2018. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3160586. Acesso em 16 jun. 2020.

GREGORY, M. AI Trained on Old Scientific Papers Makes Discoveries Humans Missed. *Vice*, 9 jul. 2019. Disponível em: https://www.vice.com/en_us/article/neagpb/ai-trained-on-old-scientific-papers-makes-discoveries-humans-missed. Acesso em 17 jul. 2020.

HAN, J.; PEI, J.; KAMBER, M. *Data mining*: concepts and techniques. [S.l.] Elsevier, 2011.

HAYKIN, Simon. *Redes neurais*: princípios e práticas. Trad. Paulo Martins Engel. – 2.ed. – Porto Alegre: Bookman, 2001.

HO, Dean. Addressing COVID-19 Drug Development with Artificial Intelligence. *Advanced Intelligent Systems*. vol. 2. 5. 2020. Publicado por WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202000070.

HUGENHOLTZ, P. B. *Data property*: Unwelcome Guest in the house of IP. In: REDA, J. (ed.). Better Regulation for Copyright: Academics meet Policy Makers. TheGreens/EFA, p. 65-77, 2017. Disponível em: https://juliareda.eu/wp-content/uploads/2017/09/2017-09-06_Better-Regulation-for-Copyright-Academics-meet-Policy-Makers_Proceedings.pdf. Acesso em 16 jun. 2020.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL (INPI). *Observatório de Tecnologias Relacionadas ao Covid-19*. Telemedicina e Inteligência Artificial. Disponível em: https://www.gov.br/inpi/pt-br/servicos/patentes/tecnologias-para-covid-19/Telemedicina

JOHN HOPKINS UNIVERSITY. Center for Systems Science and Engineering (CSSE). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. Disponível em: https://github.com/CSSEGISandData/COVID-19. Acesso em 23 ago. 2020.

KELLEHER, J. D.; TIERNEY, B. *Data Science***.** Cambridge: MIT Press, 2018.

KROENKE, D. M. *et al*. *Database Concepts*. 8ª ed. Nova York: Pearson, 2016.

MARTENS, B. *The importance of data access regimes for artificial intelligence and machine learning*. JRC Technical Reports: JRC Digital Economy Working Paper 2018-09, dec. 2018.

NEXTSTRAIN. *Nextstrain*: analysis and visualization of pathogen sequence data. Disponível em: https://nextstrain.org/docs/getting-started/introduction. Acesso em 17 jul. 2020.

PINHEIRO, A. M.; TIGRE, P. B. (eds.). *Inovação em serviços na economia do compartilhamento*. Rio de Janeiro: Saraiva, 2019.

QUINTAIS, João Pedro. Rethinking Normal Exploitation: Enabling Online Limitations in EU Copyright Law. *AMI : Tijdschrift voor Auteurs-, Media- & Informatierecht*. 41 (6). 2017. Pp.197-205.

ROWLEY, J. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, vol. 33 (2), pp. 163, 2007.

RUSSEL, Stuart; NORVIG, Peter. *Inteligência Artificial*; tradução Regina Célia Simille- Rio de Janeiro: Elsevier: 2013. (Tradução de Artificial Intelligence, 3rd. ed.)

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, vol. 3 (3), pp. 210–229, jul. 1959

SAUTOY, Marcus du. *The creativity code*: art and innovation in the Age of AI. The Belknap Press of Harvard University Press. Cambridge, Massachusetts. 2019.

SCHIRRU, Luca. *Direito Autoral e Inteligência Artificial*: Autoria e Titularidade em Produtos da IA. 2020. Tese (Doutorado em Políticas Públicas, Estratégias e Desenvolvimento). Instituto de Economia. Universidade Federal do Rio de Janeiro (UFRJ). Rio de Janeiro, 2020.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. Review. *Neural Networks*, 61. 2015. Pp. 85–117.

SOUZA, Allan Rocha. *A função social dos direitos autorais*: uma leitura civil-constitucional das limitações aos direitos autorais. Rio de Janeiro: Editora da Faculdade de Direito de Campos, 2006.

STUCKE, M. E.; GRUNES, A. P. Debunking the Myths Over Big Data and Antitrust. In: *CPI Antitrust Chronicle*, 2, mai. 2015.

WANG, Shuai; KANG, Bo; MA, Jinlu; ZENG, Xianjun; XIAO, Mingming; GUO, Jia; CAI, Mengjiao; YANG, Jingyi; LI, Yaodong; MENG, Xiangfei; XU, Bo. A deep learning

algorithm using CT images to screen for Corona Virus Disease (COVID-19). *medRxiv*. 2020. Disponível em: https://doi.org/10.1101/2020.02.14.20023028.

WEBSENSORS. *Um poderoso framework de Inteligência Analítica*. Disponível em: https://www.websensors.net.br/websensors/. Acesso em 23 ago. 2020.