Revista Ciência Aberta Lusófona
V.2, 2026
E-ISSN: 3051-7842

Pecha kucha

Os Direitos relacionados aos Dados de Investigação na prática de Text and Data Mining: os impactos na investigação

Rights related to research data in the practice of text and data mining: the impacts on research

Derechos relacionados con los datos de investigación en la práctica de la minería de texto y datos: repercusiones en la investigación

Suziane Batista Marinho*
Universidade de Coimbra, Portugal
ORCID: https://orcid.org/0000-0002-1986-2191
sbmarinho@student.fluc.pt

Maria Manuel Borges
CEIS20, Universidade de Coimbra, Portugal
ORCID: https://orcid.org/0000-0002-7755-6168
mmb@fl.uc.pt

* Autor correspondente.

Resumo

A ciência produz hoje um grande volume de dados como resultado direto do modo como se processa a investigação. Este crescimento tem reforçado a importância de práticas que permitam a sua exploração de forma eficiente. Entre essas práticas, o *Text and Data Mining* (TDM) destaca-se pela sua capacidade de extrair conhecimento a partir de grandes volumes de informação. Contudo, a utilização de dados levanta questões jurídicas que devem ser consideradas. Ainda que os dados não sejam geralmente protegidos por direitos de autor, a sua integração em bases de dados ou obras científicas pode introduzir limitações relevantes à sua reutilização. Esta proposta pretende analisar de que forma os direitos de autor e o direito *sui generis* das bases de dados condicionam a utilização de dados no contexto do TDM. Adota-se a definição de dados de investigação como registos factuais utilizados para validação científica (OECD, 2007), distinguindo-se as dimensões de entrada (input) e de saída (output). Em ambas podem surgir limitações

CC BY

associadas a licenciamento, custos e incerteza jurídica. No caso das bases de dados, a dupla proteção – direito de autor e direito *sui generis* – pode restringir a extração e reutilização de dados, mesmos quando estes não são protegidos individualmente. A Diretiva (UE) 2019/790 introduz exceções para TDM baseadas no acesso legal, mas persistem obstáculos práticos. Conclui-se que subsiste uma tensão entre proteção jurídica e ciência aberta, sendo o licenciamento aberto essencial para promover a reutilização de dados e o avanço científico.

Palavras-chave:

Direitos de autor; Direito *Sui Generis*; Dados de investigação; Base de dados; Prospeção de textos e dados; Reutilização de dados.

Abstract

Science today produces a large volume of data as a direct result of how research is conducted. This growth has reinforced the importance of practices that allow for its efficient exploitation. Among these practices, Text and Data Mining (TDM) stands out for its ability to extract knowledge from large volumes of information. However, the use of data raises legal issues that must be considered. Even though data is generally not protected by copyright, its integration into databases or scientific works can introduce significant limitations to its reuse. This proposal aims to analyze how copyright and the *sui generis* rights of databases condition the use of data in the context of TDM. It adopts the definition of research data as factual records used for scientific validation (OECD, 2007), distinguishing between the input and output dimensions. Both dimensions may be subject to limitations related to licensing, costs, and legal uncertainty. In the case of databases, dual protection—copyright and the *sui generis* right—can restrict the extraction and reuse of data, even when the data elements themselves are not individually protected. Directive (EU) 2019/790 introduces exceptions for TDM based on lawful access; however, practical obstacles remain. It is concluded that a tension persists between legal protection and open science, with open licensing being essential to promote data reuse and scientific advancement.

Keywords:

Copyright; *Sui Generis* Law; Research Data; Database; Text and Data Mining; Data reuse.

Resumen

La ciencia actual genera un gran volumen de datos como resultado directo de la metodología de investigación. Este crecimiento ha reforzado la importancia de prácticas que permitan su explotación eficiente. Entre estas prácticas, destaca la minería de texto y datos (MTD) por su capacidad para extraer conocimiento de grandes volúmenes de información. Sin embargo, el uso de datos plantea cuestiones legales que deben considerarse. Si bien los datos generalmente no están protegidos por derechos de

autor, su integración en bases de datos u obras científicas puede introducir limitaciones significativas para su reutilización. Esta propuesta tiene como objetivo analizar cómo los derechos de autor y los derechos *sui generis* de las bases de datos condicionan el uso de datos en el contexto de la MTD. Se adopta la definición de datos de investigación entendidos como registros fácticos utilizados para la validación científica (OCDE, 2007), distinguiendo entre las dimensiones de entrada (input) y salida (output). En ambas pueden surgir limitaciones relacionadas con la concesión de licencias, los costes y la incertidumbre jurídica. En el caso de las bases de datos, la doble protección —derechos de autor y derecho *sui generis*— puede restringir la extracción y reutilización de los datos, incluso cuando estos no están protegidos de forma individual. La Directiva (UE) 2019/790 introduce excepciones para la TDM basadas en el acceso legal; no obstante, persisten obstáculos prácticos. Se concluye que subsiste una tensión entre la protección jurídica y la ciencia abierta, siendo las licencias abiertas un elemento esencial para fomentar la reutilización de los datos y el avance científico.

Palabras clave:

Derechos de autor; Ley *sui generis*; Datos de investigación; Base de datos; Minería de texto y datos; Reutilización de datos.

Introdução

Na prática científica, cada vez mais os dados tem assumido protagonismo, ao fornecerem a base para que a investigação seja desenvolvida e auxiliar em novas descobertas. Muitos são os benefícios atribuídos ao compartilhamento e reutilização dos dados de investigação, nomeadamente, redução de dados duplicados, reprodutibilidade, transparência, entre outros. O volume de dados requer práticas que permitam a sua exploração eficaz, destacando-se o *Text and Data Mining* (TDM) pela capacidade de extrair conhecimento a partir de grandes volumes de informação. Todavia, a utilização de dados em larga escala coloca questões jurídicas que devem ser consideradas. Ainda que os dados não sejam geralmente protegidos por direitos de autor, a sua integração em bases de dados ou obras científicas pode introduzir limitações relevantes à sua utilização.

Esta proposta pretende analisar de que forma os direitos de autor e o direito *sui generis* das bases de dados influenciam a prática do TDM e respetivo impacto na investigação.

Dados de investigação e enquadramento jurídico

A definição de "dados de investigação" é complexa, refletindo a diversidade dos contextos disciplinares em que são produzidos e utilizados. Adota-se a definição proposta pela Organisation for Economic Co-operation and Development (OECD) (2007), para a qual os dados de investigação são "registos factuais usados como fontes primárias na investigação, e que são geralmente aceitos pela comunidade científica como necessários para validar os resultados de investigação" (p. 13). Entendidos no sentido lato como

"atos, factos ou informações", eles podem ser incorporados em bases de dados ou apresentarem-se em obras literárias e artísticas que, uma vez protegidas, podem levantar dúvidas quanto à reutilização dos dados de investigação. Assim, importa destacar que os dados, por si só, não são objeto de proteção por direitos de autor, o que se protege é a forma como esses dados são apresentados ou organizados. Esta distinção complexifica-se quando os dados são incorporados em estruturas que beneficiam de proteção legal, como é o caso das bases de dados.

Neste contexto, a análise dos dados de investigação pode ser feita a partir de duas dimensões complementares: a dimensão de entrada (*input*), ou seja, os *datasets* (conjunto de dados) usados na investigação e a dimensão de saída (*output),* que diz respeito aos dados gerados como resultados da investigação. Se as fontes de entrada tiverem algum tipo de direito de autor, deve-se verificar se há permissões legais que permitam aos investigadores prosseguirem com a reutilização; caso contrário, os dados só poderão ser obtidos com a autorização do autor, o que pode acarretar custos e um orçamento adicional para pagamento de taxas de licenciamento, por exemplo. No caso da dimensão de saída, os próprios *datasets* gerados durante a investigação ou como resultado podem atrair para si tais direitos (de autor, conexos ou proteção *sui generis*). Logo, o compartilhamento e a reutilização, acarretam questões de direito de autor, e, mais uma vez, os autores carecem de perceber a respeito de permissões, obrigações de autorização, que, por sua vez, podem exigir tempo e novamente custos relacionados (Senftblen, 2022). Assim, nas duas situações podem existir restrições associadas à necessidade de autorização, licenciamento ou cumprimento de condições específicas de utilização.

A proteção das bases de dados

As bases de dados científicas podem conter quase todos os tipos de dados, artigos de periódicos ou resumos, metadados, dados brutos de investigação, etc. (Guibault & Wiebe, 2013). Quando estes dados são recolhidos numa base de dados não original, o fabricante pode reivindicar alguns direitos que impedem a extração e a reutilização de partes substanciais de dados que de outra forma seriam dados desprotegidos, este é o chamado Direito *Sui Generis* de Banco de Dados (Labastida & Margoni, 2020).

O direito *sui generis*, estabelecido na Europa por meio da Diretiva 96/9/CE, confere uma dupla camada de proteção às bases de dados, e os dados nelas contidos poderão ser protegidos tanto pela proteção do direito de autor, que envolve a estrutura da base sendo necessário que a seleção ou a disposição da base de dados represente a criação intelectual do próprio autor que represente originalidade e criatividade, quanto, a proteção do direito de base de dados, ou seja, o direito daquele que realizou um investimento substancial na sua elaboração, ambos direitos são cumulativos. Esta proteção dupla tem implicações relevantes, na medida em que pode limitar a extração e reutilização de dados, mesmo quando estes não são protegidos individualmente.

Grande parte do conhecimento produzido pelos investigadores é coletado e distribuído por meio de base de dados. Ao compilar uma vasta gama de informações científicas, as bases de dados agilizam o

compartilhamento de informações entre os investigadores e, assim, facilitam a investigação e ajudam a tornar as descobertas disponíveis mais rapidamente, portanto, são fundamentais no processo de investigação (Chapman, 1998).

Entretanto, os dados de investigação podem estar presentes em obras ou base de dados protegidas por estes direitos e, indiretamente, acabam por atrair esta proteção, uma vez que, se não houver uma autorização por parte do autor ou uma licença para o uso, eles não poderão ser reutilizados, portanto, enquanto não há acesso aberto aos dados por omissão, a adoção de medidas que retire as barreiras de acesso para fins de investigação é fundamental.

O TDM e dos desafios jurídicos

O *Text and Data Mining* (TDM), tal como definido pela Diretiva (UE) 2019/790, art. 2º, nº 2, consiste em "qualquer técnica de análise automática destinada à análise de textos e dados em formato digital, a fim de produzir informações, tais como padrões, tendências e correlações, entre outros".

Nesse sentido, as técnicas de TDM se utilizam tanto dos dados de investigação quanto das obras inseridas em bases de dados. Assim, os dados, amplamente gerados e utilizados no decorrer da investigação, auxiliam na descoberta de conhecimento, elaboração de novos produtos, contribuem para o progresso científico em favor da sociedade, em outras palavras, são imprescindíveis para o avanço da ciência.

Para a sua execução, o TDM se baseia em trabalhos anteriores que precisam de estar licenciados para que não haja infração de direitos. Mesmo que os dados individuais não estejam protegidos por direitos autorais, a sua reutilização estará sujeita às condições de licenciamento da base de dados, uma vez que, o acesso aos dados individuais só é possível por meio da base de dados que os hospeda (Parolini, 2026).

Para funcionar, o TDM implica a realização de várias operações — como a cópia, a extração e o processamento de dados — que podem interferir com direitos exclusivos dos titulares de direitos de propriedade intelectual.

Na prática, isto significa que investigadores que pretendam recorrer a técnicas de TDM devem ter em conta não apenas o acesso aos dados, mas também as condições jurídicas associadas à sua utilização. Em muitos casos, estas condições traduzem-se em licenças, restrições contratuais ou custos adicionais, o que pode limitar o alcance destas práticas.

A exceção para TDM na Diretiva (UE) 2019/790

Na antiga legislação da União Europeia (UE) sobre direitos autorais Diretiva 2001/29/CE, não havia clareza se o *Text and Data Mining* (TDM) poderia estar coberto por exceções e limitações disponíveis, e a sua aplicação era incerta (Geiger et al., 2018).

CC BY

A Diretiva (EU) 2019/790 trouxe algumas mudanças relevantes, dentre elas, uma exceção explícita para a prospeção de textos e dados para fins de investigação científica. No nº 1 do seu artigo 3º estabelece a obrigação de os Estados-Membros criarem uma exceção aos direitos de autor e alguns direitos conexos relativamente às reproduções e extrações realizadas por organismos de investigação com o objetivo de realizar, para investigação científica, TDM de obras ou outras prestações a que tenham "acesso legal", que deve ser entendido como o acesso a conteúdos baseados numa política de acesso aberto ou através de acordos contratuais entre titulares de direitos e organismos de investigação, bem como os conteúdos livremente disponíveis em linha[1].

O facto de se tratar de uma exceção obrigatória, para além do benefício contido no artigo em favor do TDM e da investigação científica, também favorece a harmonização jurídica entre os Estados-membros da UE e contribui para o livre acesso aos dados de investigação e consequentemente à construção do conhecimento. A exceção fornecida pela Diretiva 2019/790 consiste em um avanço para o acesso aos dados de investigação, entretanto, os direitos de autor e o direito *sui generis* ainda podem gerar um impacto negativo na investigação, caso não haja o licenciamento adequado para que os dados de investigação contidos em obras e/ou bases de dados, possam ser reutilizados nas práticas de TDM e assim, gerar conhecimento.

Considerações Finais

Os dados por si só, não são protegidos por direitos de autor, o que é protegido na obra científica, não são as ideias, os métodos ou os dados, mas a sua expressão formal, portanto, grande parte dos dados de investigação não são abrangidos por direitos de autor, mas a sua utilização pode ser condicionada por enquadramento jurídico. No caso da TDM, estas limitações assumem particular relevo, dado que esta prática depende do acesso e processamento intensivo de dados. A Diretiva (EU) 2019/790 representa um passo importante no reconhecimento destas necessidades. É necessário continuar a refletir sobre formas de equilibrar acesso vs proteção, procurando soluções que permitam explorar o potencial dos dados de investigação na construção da ciência, a fim de contribuir para uma ciência mais aberta, transparente e acessível a todos.

Referências

Chapman, A. R. (1998). A human rights perspective on intellectual property, scientific progress, and access to the benefits of science. In *Intellectual Property and Human Rights* (pp. 127–174). WIPO. https://tind.wipo.int/record/28737/files/wipo_pub_762.pdf

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, CONSIL, EP, 077 OJ L (1996). http://data.europa.eu/eli/dir/1996/9/oj

[1] Considerando 14 da Diretiva 2019/790.

Diretiva (UE) 2019/790 do Parlamento Europeu e do Conselho - de 17 de abril de 2019 - relativa aos direitos de autor e direitos conexos no mercado único digital e que altera as Diretivas 96/9/CE e 2001/29/CE. https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32019L0790

Diretiva 2001/29/CE do Parlamento Europeu e do Conselho, de 22 de Maio de 2001, relativa à harmonização de certos aspectos do direito de autor e dos direitos conexos na sociedade da informação. https://eur-lex.europa.eu/eli/dir/2001/29/oj/eng

Geiger, C., Frosio, G., & Bulayenko, O. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market—Legal Aspects. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3160586

Guibault, L., & Wiebe, A. (Eds.). (2013). *Safe to be open—Study on the protection of research data and recommendations for access and usage*. Universitätsverlag Göttingen. https://doi.org/10.17875/gup2013-160

Labastida, I., & Margoni, T. (2020). Licensing FAIR Data for Reuse. *Data Intelligence, 2*(1–2), pp. 199–207. https://doi.org/10.1162/dint_a_00042

Organisation for Economic Co-operation and Development. (2007). *OECD principles and guidelines for access to research data from public funding*. OECD Publishing. https://doi.org/10.1787/9789264034020-en-fr

Parolini, G. (2026). Intellectual Property Rights, open science and open questions on research data: A biodiversity science perspective. *Research Ideas and Outcomes, 12*, e184961. https://doi.org/10.3897/rio.12.e184961

Senftblen, M. (2022). *Study on EU copyright and related rights and access to and reuse of data*. Publications Office. https://data.europa.eu/doi/10.2777/78973