Creative Commons, Creative Commons AU, and Wikimedia AU response to the Australian Government's Productivity Commission Interim Report on "Harnessing data and digital technology"

14 September, 2025
Submitted via https://engage.pc.gov.au/page/make-a-submission

About Creative Commons

Creative Commons is a nonprofit organization that helps overcome legal obstacles to the sharing of knowledge and creativity. To that end, we publish a suite of licenses and legal tools designed to give every person and organisation in the world a free, simple, and standardised way to grant copyright permissions for creative and academic works, ensure proper attribution, and allow others to copy, distribute, and make use of those works. In addition, we work closely with major institutions and governments to create, adopt and implement open licensing and ensure the correct use of CC licenses and CC-licensed content. We also engage in public policy, contributing to shaping the laws and rules governing sharing and the commons, including any technological impacts, with a particular emphasis on copyright reform in the public interest.

About Creative Commons Australia

Creative Commons Australia (CC Australia) is the local Chapter of the CC Global Network. CC Australia promotes and encourages the use of CC licences in Australia. We connect the Australian commons and we connect Australia to "the commons". We coordinate Chapter meetings and working groups, organise engagement events and share global news with the Chapter and our communities.

About Wikimedia Australia

Wikimedia Australia is the Australian chapter of the international Wikimedia Foundation. As an independent, not-for-profit organisation and registered charity, we support our members, the broader community and partner organisations to contribute to Wikipedia, Wikidata and other Wikimedia platforms through events, training and partnerships.

Summary

Creative Commons is pleased to respond to the Australian Government's Productivity Commission Interim Report on "Harnessing data and digital technology." Our comments focus on the copyright aspects of the interim report, and fall under two of the four thematic areas addressed, namely:

- Enable AI's productivity potential
- New pathways to expand data access

In sum, we support the Government's focus on finding ways to appropriately expand data access, including through reforms to copyright, while taking into account the wide variety of interests and rights that may be implicated.

The data that powers AI was created by people and communities. Today, billions of webpages fuel AI systems. This has happened quickly at an unprecedented scale. It has also happened without the involvement of content creators and stewards, reaching beyond people's reasonable expectations for how their copyright material would be used when they shared them publicly. AI development and use raises a number of important concerns, including the impact these tools will have on artists and creators' jobs and compensation, and harmful use of these tools to exploit people's privacy (e.g., their biometric data) or perpetuate biases. These issues particularly impact the open sharing community, which often is producing information, including scholarly and other creative outputs in the public interest, without the same commercial incentives for creation as many other content creators.

The Government should not evaluate this issue as an all-or-nothing choice. It is a mistake to view expanded access to data for text and data mining (TDM), including for AI development, as inherently incompatible with ensuring creators have appropriate choices over use of their works and the ability to be remunerated. AI is outpacing the social contract, but the Government should seek out win-win solutions rather than treat this as a zero sum game, putting the sustainability of our shared information commons and the public interest as its core objective.

Our submission below elaborates on this theme, with a focus on the way copyright law and AI interact. Our argument proceeds as follows:

- A well-designed TDM exception to copyright can be consistent with copyright's purpose and function
- An effective TDM exception can be consistent with ways to provide rightsholders with effective ways to "opt-out" of certain AI training
- A TDM exception can be consistent with a thriving licensing market and other forms of remuneration
- A TDM exception that allows works to be used as *inputs* for AI training need not override copyright holder's ability to address infringing *outputs* of AI systems

In devising the proper balance, Australia can build on and learn from examples around the world on how to best craft such an exception in ways that are clear and fair. As one recent paper notes:

> "Globally, the binary policy debate that assumes that text and data mining and AI training must be categorically condemned or applauded has been eclipsed by a more granular debate about the specific circumstances in which the unauthorised use of copyrighted works for AI training should be allowed or prohibited. Countries that have hesitated until now to modernise their copyright laws in the area of AI training have several templates open to them and little reason for hesitation."[1]

About our approach to copyright, TDM, and AI

A founding insight of Creative Commons is that all creativity builds on the past. Creators necessarily learn from, and train their own skills by, engaging with pre-existing content and artists — for instance, noticing the style in which musicians arrange notes, or building on surrealist styles initiated by visual artists. Likewise, scientists and researchers build on past discoveries and the existing literature to gain a better understanding of how the world works and to progress ideas.

TDM is a way to study and analyse existing works, using machines, in order to create new insights and materials. TDM typically consists of computers analyzing huge amounts of text or data, and has the potential to unlock huge swaths of interesting connections between textual and other types of content. Understanding these new connections can help organize our shared knowledge, and enable new research capabilities that result in novel scholarly discoveries and critical scientific breakthroughs. Because of this, TDM is increasingly important for research. While much of the discourse around TDM as applied to generative AI has focused on the creation of artistic works, TDM and AI have uses that can help generate advances across science, education, healthcare, and other domains that are of significant importance to society.[2]

Another motivation for founding Creative Commons was offering more choices for copyright owners who wish to share their works. Through engagement with a wide variety of stakeholders, we heard frustrations some copyright owners had with the inflexibility they seemed to face with copyright's default framework, i.e. all rights reserved. Instead they wanted to let the public share and reuse their works in some ways but not others. Our engagement revealed that people were

---

[1] Sag, Matthew and Yu, Peter K., The Globalization of Copyright Exceptions for AI Training (October 04, 2024). Emory Law Journal, Vol. 74, 2025, Forthcoming, Texas A&M University School of Law Legal Studies Research Paper No. 24-75, Emory Legal Studies Research Paper Forthcoming, Available at SSRN: https://ssrn.com/abstract=4976393 or http://dx.doi.org/10.2139/ssrn.4976393, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4976393

[2] See examples here: https://www.knowledgerights21.org/news-story/kr21-principles-on-artificial-intelligence-science-and-research/

motivated to share not merely to serve their own individual interests, but rather because of a sense of societal interest. Many wanted to support and expand the body of knowledge and creativity that people could access and build upon — that is, the commons. Creativity depends on a thriving commons, and helping people license their works for more uses is a means to that end.

Similar themes have come through as we have consulted with our community about generative AI. Obviously, the details of AI and technology in society today are different from 2002 when the first CC license was published. But the challenges of an all-or-nothing system where works are either open to all uses, including AI training, or entirely closed, are a through-line. So, too, is the desire of some copyright owners to share in ways that support creativity, collaboration, and the commons.

In turn, we have been engaging actively with stakeholders to avoid “all-or-nothing” choices in the age of AI. Specifically, we have been exploring the development of “preference signals” for AI training – a way for copyright holders and data stewards to make requests to potential users about some uses, not enforceable through the licenses, but as an indication of their wishes. Preference signals raise a number of tricky questions, including how to ensure they are a part of a comprehensive approach to supporting a thriving commons — as opposed to merely a way to limit particular ways people build on existing works, and whether that approach is compatible with the intent of open licensing. At the same time, we do see potential for them to help facilitate sharing in the public interest. We have heard from many different people, including artists and rightsholders, as well as developers of AI technologies, and we believe that it is possible for policy to guide the development of AI in ways that benefit all.

A well-designed TDM exception to copyright can be consistent with copyright’s purpose and function

A TDM exception can be consistent with copyright’s purpose and function – providing sufficient incentives for creativity in order to benefit the public. No creativity happens in a vacuum, purely original and separate from what’s come before; people observe the ideas, styles, genres, and other tropes of past creativity, and use what they learn to create anew. Copyright law has always reflected this fact – by differentiating between protectable expression and uncopyrightable ideas and facts, through appropriate exceptions, and by ensuring broad freedoms associated with the public domain, among other limitations to copyright.

Today, people are using TDM, including in its application to develop AI, including general purpose and generative AI, to study, analyse, and learn from past works. Just as people learn from past works by studying them directly, they can use computers to study works – for example, “training” generative AI by analyzing past materials in order to extract uncopyrightable information and make non-infringing outputs.

Given how digital technologies function, engaging in TDM may involve making intermediate copies of copyrighted works in order to analyse them. These uses are sometimes called non-expressive uses – that is, uses that involve copying, but do not communicate the protectable expressive aspects of the underlying work so that it can be read or otherwise enjoyed. Treating this sort of copying as per se infringing copyright can, in effect, shrink the commons and impede others' creativity, by restricting legitimate uses only because they are machine-enabled and involve copying as an intermediate step. It can extend creators a monopoly over ideas, genres, and other concepts not limited to a specific creative expression, as well as over new tools for creativity.

In general, we think that analysis of existing works in order to derive uncopyrightable elements or make otherwise non-infringing uses can be accomplished consistent with the purpose and function of copyright law, even if it involves making a copy of a whole work as an intermediate step, such as through TDM.

An effective TDM exception can be consistent with ways to provide rightsholders with effective ways to "opt-out" of certain AI training

At the same time, we recognize the broad concerns around AI today and, in particular, how they are trained on existing works. The data that powers AI was created by people and communities. Today, billions of webpages fuel AI systems. This has happened quickly at an unprecedented scale. It has also happened without the involvement of content creators and stewards, reaching beyond people's reasonable expectations for how their copyright material would be used when they shared them publicly.

Devising effective copyright that serves the public interest requires careful consideration of the many interests at play. Expanding copyright to control AI training risks stifling innovation and access to knowledge, with negative knock-on effects on science, research, and education. The future depends on shared expectations and responsible reuse. Any viable solution needs to be legally grounded, technically interoperable, and backed by the collective action of humans.

We are also supportive of preference signals that help copyright owners indicate how they wish for their copyright material to be used. It is critical that such signals are narrowly tailored to serve the public interest, and to recognise that the development of these signals is also not a binary choice. They could be crafted to differentiate different types of uses and users. Like robots.txt, they could be developed as voluntary standards and norms, providing flexibility and adaptability for the signals to evolve over time and be interpreted contextually.

For our part, Creative Commons released an initial concept for CC Signals, a first step to giving creators back more agency in how their works are used in the context of AI.[3] Unlike the CC licenses, they are explicitly designed to signal expectations even where copyright law is silent or unclear, when it does not apply, and where it varies by jurisdiction. We have listened to creators who want to share their work but also have concerns about exploitation and the viability of the commons. CC signals provide a way for creators and data stewards to express those nuanced expectations. The CC signals will likely build on top of developing standards for expressing AI usage preferences (e.g., via robots.txt). CC signals are for those who want to keep sharing, but with some expectations that use of their works will sustain a thriving commons. We hope that the spirit of this tool is taken into account in different jurisdictions that are designing TDM exceptions to copyright, including in Australia.

Alternatively, similar to the EU's existing TDM exception, they could be part of a legally binding "rights reservation" or "opt-out" scheme. That said, a more rigid, prescriptive approach instantiated in law makes it even more essential to carefully address implementation and operational challenges at the outset.

As the Australian Government considers this issue and possible "opt-out" systems, it should consider key criteria and implementation challenges, including:

- Differentiate and broadly protect research uses: In the EU's 2019 copyright directive there are two prongs to the TDM exception. Article 3 provides a TDM exception for certain research uses, and this protection is not subject to the rights reservation also referred to as "opt-out." Meanwhile, Article 4 provides an exception for all users, except that respect for the rights reservation is required. The AU should follow this basic structure, and consider a broader, clearer exception for research, such that, in addition to research and cultural heritage institutions, nonprofits and other research labs unaffiliated with research institutions can also easily avail themselves of the exception. This should clearly cover, for instance, educational institutions and educational uses. As researcher advocacy organization KnowledgeRights21 has argued, "public research institutions and knowledge valorisation is held back by the artificial and unworkable distinction that EU / UK copyright law makes between commercial and non-commercial research. Despite European governments' claimed strong support for public-private partnerships in research, the reality is that the distinction between commercial and non-commercial research that European copyright law makes means that working with and sharing information between partners in the context of knowledge valorisation projects is impossible."[4]
- Require machine readable rights reservations by rightsholders: Opt-outs should be provided via clearly understandable and machine readable formats, such that it is easy

[3] https://creativecommons.org/ai-and-the-commons/cc-signals/

[4] https://www.knowledgerights21.org/news-story/kr21-principles-on-artificial-intelligence-science-and-research/

for TDM users to readily determine the copyright status of a work. Moreover, it should only be rightsholders of a particular copyright-protected work that can exercise the reservation as, after all, they are the ones who control the pertinent rights.

- Ensure flexibility for new technologies: TDM and its uses will continue to evolve. The exception should be able to accommodate TDM not just in today's context of generative AI but also in future contexts. As such, any legislative framework should be crafted in such a way that it is as future proof as possible or at the very least includes a mechanism for review, without undermining legal certainty.
- Ensure feasibility: in line with the above, it is crucial that rights reservations be readily identifiable and available. They must not be highly difficult or costly to implement for rightsholders or users. To that end, developers should not be held liable where the validity of a rightsholder claim is uncertain, such as when multiple entities offer conflicting preferences or when someone fraudulently or inadvertently or mistakenly claims to be the rightsholder.
- Facilitate interoperability: where appropriate, rights reservation systems should build on and complement existing tools like robots.txt. Where practical, it is also important to develop systems that work across media types and across jurisdictional borders.

A TDM exception can be consistent with a thriving licensing market and other forms of remuneration

The consultation asks about how to support effective licensing. Here, it is important to recognise that the creation of a TDM exception is not antithetical to licensing as a whole. Rather, TDM is already facilitating the creation of new types of tools and services, which in turn are opening up new licensing opportunities in *other* uses beyond protected TDM. For instance, while generative AI developers may train their base models on copyrighted works under national copyright limitations and exceptions, many are still working to obtain licenses over works for uses that go above and beyond permitted uses, such as displaying significant expressive text from a particular news article or expressing images with a high degree of similarity to protected works. This is not a new phenomenon, but rather consistent with the history of copyright – for instance, uses of the video camera and video cassette recorder (VCR) were built on copyright's limitations and exceptions, but gave rise to a huge market for film rentals and sales.

A TDM exception that allows works to be used as inputs for AI training need not override copyright holder's ability to address infringing outputs of AI systems

It is possible for AI-generated works to implicate the exclusive rights of rightsholders in some cases. We cannot enumerate all cases specifically; however, there are many possible ways this

can happen, and we urge caution and flexibility in determining when an output has infringed exclusive rights. It is possible for an output to be infringing; it is also possible that an output including material under copyright may be using it in a way that is fair, legitimate, and lawful.

Liability for any infringement in this case should depend heavily on the facts of the situation, as the nature of generative AI tools and usage of those tools can vary greatly. For example, where a user has made a knowing attempt to get copyrighted material they have previously had access to as output from an AI tool (by directing prompts specifically toward that output, for example), that user has simply used the tool to reproduce the material, in the same way that one might use other tools to create a copy of an existing work. This may be non-infringing, for example if said use falls under an exception or limitation to copyright or if another defense can be invoked.

We would caution against considering a simple attempt to make reference to a rightsholder's work as an infringement—for example, a user requesting artwork in the style of an existing artist or a user quoting from a literary work in order to comment or parody. Copyright law already contemplates such cases, and already deals with some ambiguity that depends on facts: was a creator simply being inspired by the style of an artist, which is not an infringement and is critical to free expression and the development of artistic genres? Or was a creator attempting to reproduce copyrightable elements in an infringing manner? Likewise, quotation for purposes of criticism, commentary, parody or pastiche is largely considered fair and legitimate, even though it involves use of works protected by exclusive rights. Private use is another example of use that may implicate one or more of the exclusive rights, but which is considered lawful under copyright law.

The debate about how copyright should apply to generative AI has often been cast in all-or-nothing terms—does it infringe on pre-existing copyrights or not? The answer to this question is certainly that generative AI *can* infringe on other works, but just as easily it may not. The criteria that already exist in copyright to determine substantial similarity between works are apposite to address these questions with regard to AI outputs. Our concern is that overregulation of generative AI could easily tip the carefully crafted balance achieved within the copyright system, undermining AI tools' ability to help people realise their creative visions, advance research, improve productivity, and positively transform society.