Copyright & Licensing: Current context and considerations for researchers and libraries using AI in research today

Alex Fenlon Maria Rehbinder

2024-06-04

This guide aims to provide library professionals at European research institutions, particularly those supporting or undertaking activities that combines AI tools and methods and digital cultural heritage collections and data, with a brief overview of the current copyright and licensing context for such research today.

Introduction

This guide aims to provide library professionals at European research institutions, particularly those supporting or undertaking activities that combines AI tools and methods and digital cultural heritage collections and data, with a brief overview of the current copyright and licensing context for such research today.

If it's been awhile since you've had a look at copyright rules closely, let's take a minute to cover the basics so we have a good foundation for understanding the more complex considerations of copyright in an AI context today.

Equally, if you feel well versed in the basics already and are keen to dive straight into the AI context, **feel free to hop on down to the Relevance to the Library section**.

Copyright basics

"Copyright is a legal term used to describe the rights that creators have over their literary and artistic works. Works covered by copyright range from books, music, paintings, sculpture, and films, to computer programs, databases, advertisements, maps, and technical drawings." WIPO

Copyright is a form of intellectual property intended to protect original works as they are created and recorded. As soon as an original work is created, it is protected by copyright and

there is no need to formally register it. In some countries registration may increase authors' rights for compensation in case of infringement.

The purpose of copyright is to protect and reward creators for the works they have produced. Copyright consists of *economic rights* - the right of reproduction and the right to make a work available to the public- and *moral rights* - the right to be identified as the author and the right to prevent mistreatment.

To qualify for protection the work must be original in the sense that it is an author's own intellectual creation, i.e. it reflects the author's personality, the author was able to express their creative abilities in the production of the work by making free and creative choices.

Ideas vs expression

Copyright protects the artistic or literary form, but not the subject matter. The work that is created is protected, but not to the ideas or theories/ concepts represented. The ideas vs expression debate is a central part of the copyright discourse when looking to build on or reuse the work or ideas of another. If you are using the work itself this is likely to involve copyright issues, but if you are looking to use the ideas rather than the work this may not involve copyright issues at all.

Public domain

Copyright is finite and lasts for a set duration based on the type of work and where it was created.

In Europe and USA this period is 70 years after the author dies. According to the international Berne Convention the minimum protection period is the life of the author plus 50 years after their death. Copies or reproductions of works where copyright expired do not get protection in EU or USA:

When copyright expires it becomes Public Domain however given the differences in duration across borders a work may be Public Domain in one country yet still under copyright in another.

For example, Canada has a copyright duration of 50 years after death for literary works whereas in France it's 70. In Canada the work would not require permission for certain uses but in France those same uses would still require a licence or an exception.

Another example is the UK. "Some works are protected in the UK until 31st December 2039, even where the author died perhaps hundreds of years ago. This is also known as the '2039' rule and applies to literary, dramatic and musical works, but not artistic works other than photographs and engravings. Other rules apply for photographs, films and sound recordings."

When a work enters the Public Domain it is free to be used by anyone in any way without the permission of the owner. When copyright expires also the moral rights of the creator expire,

so works can be altered and used without crediting the author. However, many countries have given public authorities, such as the Ministry of Education/ Culture the possibility to protect the 'educational and cultural value' of works that are in the public domain, Italy for example.

Ownership

Ownership is important as it determines who has control over how and when a work is used, as well as who can access and use it. Owners, also known as rights holders, have the right to control who can access the work, how they can use it, if they can make copies, how it is performed or communicated to the public. They can licence these permissions to use to a licensee or they can sell these rights completely to another party. Ownership is a requirement for licensing, for example for licensing with Creative Commons licences.

Copyright is an intangible asset which is separate to possession of a physical work such as a painting or book and transferring physical possession does not include any transfer of copyright.

Using a work without permission from the copyright owner, usually in the form of a licence, is a copyright infringement, unless an exception applies.

Copyright ownership is granted to the creator of the work in the EU automatically by copyright legislation. However, this can be changed by contract. In the UK for example, the law says the creator is the owner but if a work was created in the course of employment, the employer will own the copyright unless the employment contract says otherwise. The US has a work-for-hire clause in the copyright legislation so that copyright is born to the employer if not otherwise agreed. Even ownership is not quite straightforward and local knowledge is important.

Open Access & Open Data

Owners of copyright works are able to apply any licensing conditions they wish to their work, however research funders may make it a condition of a grant that certain outputs arising from funded research must be released openly. Where this is the case, they will communicate their preferred licensing choice.

The most common licences used are Creative Commons licences. Creative Commons licences are intended to pre-approve certain uses meaning a user can make use of a work within the terms of those licence conditions. If a user wishes to make other uses they will need to contact the owner in the usual way and negotiate separate permissions. If a user breaches the Creative Commons terms they can be sued for infringement by the owner.

Creative Commons licences do not impact on the ownership status of the work- they are simply licences. Only the copyright owner of a work has the authority to grant a Creative Commons licence.

Creative Commons licences can be very open permitting any reuse provided the creator and the licence are cited, or they can be closer to a traditional 'all rights reserved' approach which might enable access to a work but any reuse is limited in nature. The licence terms are binding so any user must mention the name of the author and the work as shown in the licence, and if there are additional terms such as Non-Commercial or Share-Alike these have to be complied with.

Creative Commons has a comprehensive website with guidance but it is highly likely that research institutions and libraries will have other additional guidance on this too.

Research and copyright

Copyright protected materials are central to research activity whether that is in the form of text, music, images or code. Data itself is not protected by copyright, but databases can be protected as catalogue or sui generis database, as noted above.

Researchers will produce copyright protected works when writing articles, monographs or code. Researchers will need protection for the right to be recognised as authors and also as owners of copyright. Research ethics requires contributors to be recognised in a research output, even if they are not owners or authors in a copyright sense. Good scientific research requires contributor roles to be defined and mentioned in research outputs and CRediT – Contributor Role Taxonomy (niso.org) describes 14 roles that can be used when defining contributor roles. Contributor roles should be agreed before sending a manuscript to a publisher.

Researchers will use works created by others- either as a fundamental part of their research methodology or during publication. Where exceptions like those for non-commercial research or non-commercial data mining do not apply, ensuring that licences are in place that allow for the uses within the research activity is essential. The use of research data and the management of data related rights should be considered and outlined within the data management plan which must also take into account how the research data are to be archived and used, as well as how the authors of the data are attributed in compliance with research ethics.

Exceptions

In order to use the copyright protected work, either permission from the owner, agreements covering the uses, or legislation allowing the use is needed.

Copyright legislation tries to strike a balance between protecting the freedom of sciences, expression, and information, and the protection of the copyright owner. This means that while the owner of the rights can control it, there are exceptions to those rights which enable some uses without the need for the owner's permission. These are known as exceptions, or user rights.

Copyright exceptions include for example certain educational uses, the provision of inter library loans by libraries, text and data mining for non-commercial scientific research purposes and commercial purposes, creating accessible copies of works for those with disabilities

Copyright laws vary by country and, while EU law is largely harmonised, there are differences between member states concerning the copyright exceptions. Understanding local national law exceptions for research and education is important.

Anomalies

Using archival materials can raise some unusual situations which mean the normal copyright rules don't apply. Below are a couple of examples of unusual copyright scenarios:

Works of art on public display The rules around works of art on public display vary by country. In some EU countries the so-called 'Freedom of Panorama' rules dictate what can and cannot be done with such works for commercial purposes. Italy for example controls commercial uses of works of cultural importance regardless of their copyright status.

Contractual issues

While works in the public domain may be free from copyright access to them and their use may be controlled via contract law. Many museums and galleries apply terms and conditions to the use of digital surrogates which apply irrespective of the copyright status.

Database Rights

The EU sui generis database protection and the Nordic catalogue protection intend to protect the investment in a database. A database usually does not have the original artistic or literary form that protection of a database as an original work would require but it does involve creative choices in the structure layout and data fields being collected. While the data items themselves may not have copyright the overall database will be protected by the database right. Ownership of database rights may vary by law or contract. For example in the Nordic version these rights are given to the university as employer, not to the researcher as the employee.

Unpublished works

There are specific rules surrounding the copyright of works that have never been 'communicated to the public' or they have never been publicly accessible.

Orphan works

Where the owner of the rights in a copyright work is known but cannot be traced they are classed as orphan work. Within the EU the Orphan works Licensing Scheme is a EU wide licensing mechanism that gives assurance to users of these works subject to the payment of a fee.

Out of commerce works

Where a work is no longer commercially exploited by the owner, or publisher for example, there are EU rules around how a user might use these items.

Ethical authorship

Copyright authorship and who are authors in a scientific publication differ. When listing authors and those persons who have contributed to the research and to the publication should be mentioned, contributor roles should be defined, and a tool for this is the CRediT – Contributor Role Taxonomy (niso.org).

Licensing basics

Licensing agreements are a central type of legal contract concerning copyright works. Use of published content within institutions will be covered by licences. Access to ebooks, e-journals, databases etc are all covered by licences. It is important to know and understand the terms under which access to content is provided so that users can use the content without risk of breaching the licence terms.

Many of the permitted uses within licences closely mirror the copyright exceptions mentioned above, however they provide a clarity and certainty that the exceptions may not.

In general, licences give institutions, their staff and students, permission to use the licensed content for specific purposes. These purposes are usually limited to education and non-commercial research activities, i.e. the students, researchers and educators can use the content for their purposes but HR or financial teams that are not directly engaged in research or teaching delivery cannot.

The terms of the licences will vary but they frequently allow for saving or printing of parts of the licensed works. Some licences will allow users to include extracts within teaching materials or even within publications. How much can be used will also vary.

It is important that these purposes and licence terms are clearly understood and that the use they expect to make of the content is expressly permitted within the licence terms. For example if an institution has an active research interest in data mining, licences that seek to prevent or restrict this activity will be problematic. Where licence terms are unclear this should be raised with the provider. Where they conflict with the legal exceptions in law this should also be queried.

Institutions may also rely on licences provided by collective licensing societies- official bodies that represent a group of authors, publishers or rights holders. The licences may cover things like photocopying and scanning of printed works, showing broadcast television programmes or playing recorded music. Again understanding the terms, the uses and obligations is important.

Relevance to the Library Sector (Case Studies/Use Cases)

Libraries and cultural heritage institutions today don't just support academics to undertake computational research by providing guidance and increasingly computational access to digital collections, they undertake digital research in their own right as part of library work, using existing models and developing new ones for analysing digital collections at scale for metadata improvement and enhancement and a whole host of other applications. For European institutions understanding TDM rights granted by the Directive on Copyright in the Digital Single Market (later the DSM or Directive (EU) 2019/790) and the EU AI Act is important when undertaking this work, as well as keeping up to date on your local contexts which aligns with these but in some cases may differ, such as in the UK.

This is an evolving area and our aim here is to provide a brief overview of the current copyright and licensing context for European researchers and institutions using AI tools and methods in research today.

Text and Data Mining (TDM)

Text and data mining (TDM) is a common research technique that allows researchers, and research organisations to analyse large volumes of data using modern computing power. Under EU law the Directive on Copyright in the Digital Single Market (later the DSM or Directive (EU) 2019/790) introduced TDM copyright exceptions. TDM is defined Article 2(2) as:

> 'text and data mining' means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations;

TDM is a core part of machine learning and artificial intelligence (AI) technologies. It could include the harvesting and scrapping of online data sources, or digitising printing items so that they can be read by computers.

DSM directive Article 3 allows for the use of copyright works in TDM activities for the purposes of non-commercial scientific research by research organisations and cultural heritage institutions provided they have lawful access to the content. This exception is mandatory and rights holders can not override it using agreements or technical measures when EU legislation is applicable. If the agreement is done with an organisation outside EU and governed by legislation other than EU legislation, researchers should contact their legal department for advice on how EU legislation exceptions can be applied to contractual obligations (see Regulation (EC) N o 593/2008 of the European Parliament and of the Council of 17 June 2008 on the law applicable to contractual obligations (Rome I). It is advisable to define in agreements with US/other non-EU companies that mandatory exceptions to copyright in national EU member state legislation apply to the agreement, despite being otherwise governed by US legislation.

DSM directive Article 4 allows for text and data mining for any purpose by any organisation or person, provided they have lawful access. However, right holders are permitted to opt out of this broad exception as defined using a machine-readable opt-out, as defined in Article 4(3):3.

> the exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as m_achine-readable means in the case of content made publicly available online.

What is the way to opt out in "appropriate manner, such as machine-readable means" is now to be discussed in a court case in Germany Machine readable or not? - notes on the hearing in LAION e.v. vs Kneschke - Kluwer Copyright Blog (kluweriplaw.com). District Court of Hamburg, Germany has on 27.9.2024 made a decision in the first European case that examines the relying on the the TDM exception for the purpose of training generative AI models

As the DSM is a directive, member states were able to implement certain elements as they see fit and this has led to some disjointed approaches across the EU with some countries taking different approaches. Ireland for example requires that an author is entitled to be informed that the copy has been made for text and data mining purposes and ask for details about the steps taken to ensure the security of the works copied (see Copyright and Related Rights Act, 2000 (as amended) sections 53A and 53B). Similar requirement of transparency regarding the materials used for AI training of generative AI models is a key point in the AI Act; the sections concerning generative AI are applicable from August 2025.

Example 1

The library is a partner on a non-commercial collaborative research project at a university where a research team wants to engage in linguistic analysis, using computational methods, of EU newspaper articles the library has made available publicly online. In order to complete the research they need to extract all of the articles during a certain period to build a corpus of data sourced from different EU newspapers. Once the corpus is complete the researcher wants to use a computer program to perform the analysis. As the data is sourced from different newspaper websites, there is some data cleaning required.

All of the copying of the articles for the purposes of this research is permitted as is any data cleaning under the terms of the TDM exception above. The research team is permitted to carry out any steps necessary to obtain and format the data to enable them to complete the analysis using computational methods. If the researchers want to use printed articles those could be digitised and made machine-readable too.

Example 2

Now the corpus is complete the researcher wants to use a free online AI service to complete the analysis.

The model's terms say that the user of the service declares owning all of the input data they provide and grants a licence to the model provider which allows the model provider to retain the input data and use it as a part of the training data. The DSM directive Article 3 exception or its national legislations do not allow this granting of licence to commercial companies to the input data, so the researcher can only use paid services, which the university has purchased, and which have terms that allow the input data to stay on the university VPN.

While the mandatory exception would cover the collection and analysis of the data for scientific research purposes, it does not enable the researcher to own the data nor does the exception give the researcher the authority to grant permissions for non-scientific training data uses. If the researcher would give the material to the AI system provider, contrary to the scope of the exception legislation, then the AI system provider would not have acquired the data lawfully, so the general text and data mining exception would not apply.

AI Act and Copyright

The EU AI Act legislates obligations for providers and deployers of artificial intelligence (AI) systems and AI models and includes articles concerning the copyright protected content as training data of AI models. The AI Act confirms that text and data mining exceptions support the use of content for training AI models. AI Act Article 53 contains obligations for providers of general-purpose AI models:

- Article 53(c) *the provider of a general purpose AI model must put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights (opt-out) expressed in Article 4(3) of Directive (EU) 2019/790 (DSM directive).*
- Article 53(d) *the provider of a general purpose AI model must draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template that will be provided by the EU AI Office.*

Article 53 and other general-purpose AI sections of the AI Act are applied from August 2025 onwards.

AI models that are used to create content, such as text or images, for example Midjourney or ChatGPT, are considered general-purpose AI models in the recitals of the AI Act.

- AI Act Article 3 (4) defines *'provider' as a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.*

However, the AI Act is not directly applied to scientific research:

- Article 2(6) *The AI Act does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development.*

But if the result of a research study is a general-purpose AI model, and it is placed on the market or put to use, other than for the sole purpose of scientific research, then the *provider* requirements apply, e.g. the provider needs to be able to list a sufficiently detailed summary of the content used for training of the general-purpose AI model. According to Article 2 (8) the AI Act does not apply to testing or development prior to placing it on the market, but it applies to real world testing of the general-purpose AI model.

AI Act does not regulate artificial intelligence, which is a philosophical rather than legal term, but it regulates AI systems that are are defined in Article 3 (1):

- *'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*

The autonomy of the AI systems makes it different from other tools that are controlled by the author and this autonomous aspect of generative AI has an effect when regarding the human authorship of the generative AI system output.

AI Act Article 3 (33) defines that *'input data' means data provided to or directly acquired by an AI system on the basis of which the system produces an output. Both input data and output should be considered in the framework of copyright, so that neither are infringing copyright of third parties.*

When applying the exception rules allowing text and data mining, users of third party works as training data must bear in mind that the three step test limits all exception rules.

- Directive 2001/29 Article 5 (5) states that *the exceptions and limitations of copyright can only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rights holder.* For example a

generative AI model that would produce illustrations in the style of an illustrator could unreasonably prejudice the legitimate interests of that illustrator as rights holder.

Considerations for libraries and research projects using AI

If a research project intends to use copyright protected works as training data for AI models, researchers should consider how the text and data mining exception in EU legislation would allow the intended reproduction of works as training data. The recent court case from Hamburg District Court, considers scientific research to be a wide concept. If the research outputs are to be commercialised, consideration must also be given to the new EU AI Act and its requirement in Article 53 to document all copyright protected training data.

If researchers are using copyright protected works in existing third party AI tools, use should be balanced against the exceptions or licences covering the content they wish to use. Researchers should consider whether they are required to grant the third party tool permissions to use the content, and whether this is possible within the scope of the licence or exception. As the researcher may not own the content they wish to input, they may not have the authority to grant those permissions. This is especially important where third party tools develop their model using input data.

Increasingly, publishers are seeking to restrict the use of the licensed content as training data to train AI models or as input used in AI systems, while in other instances they are creating licencing agreements to allowing large technology companies access to scholarly content (see Generative AI Licensing Agreement Tracker - Ithaka S+R and An academic publisher has struck an AI data deal with Microsoft – without their authors' knowledge). The ICOLC Statement on AI in Licensing offers a useful template to begin to push back to ensure researchers and institutions are able to use licensed content for non-commercial research at least. Of course institutions are heavily involved in the translation of research into commercial activity so the ICOLC statement is only of limited use, but it's a start.

Consideration should also be given to the impact of AI tools on library systems and the ingestion of content, openly licensed or otherwise, by generative AI tools. For instance, the KB restricts access to collections for training commercial AI | KB, National Library of the Netherlands and in January 2024 issued a Statement on commercial generative AI | KB, National Library of the Netherlands outlining their position "that commercial parties who crawl digital resources on websites on a large scale for training models, using applications such as ChatGPT, are not complying with the AI principles established by the KB in 2020."

The harvesting of vast volumes of online content as training data for AI models is under increasing scrutiny, as is the harvesting of personal data. It should be noted that some right holders are actively seeking to prevent their content being analysed, accessed, or processed by any AI tools, regardless of whether they are "inhouse", third party, secure/ protected/

ring fenced or otherwise. Different jurisdictions have differing legislation and court cases will further define AI and copyright legal questions over time.

AI and original work protected by copyright

AI is bringing a new urgency to the requirement of originality, which is a fundamental aspect of copyright. According to the Berne Convention for the Protection of Literary and Artistic Works, a copyright protected work has to reflect/contain the intellectual creation of the author. This requirement of originality is a fundamental part of what copyright protects and is a feature of national and EU legislation endorsed in many court cases.

Example 3

In the US the person or company wishing to register copyright can apply for copyright registration from the US Copyright Office (USCO). USCO received and examined an application containing images that were created by generative AI program and service Midjourney. Users of Midjourney operate through "prompts" (text commands) which include the description of what Midjourney should generate. Midjourney does not interpret prompts as specific instructions to create a particular expressive result, but simply converts words and phrases into smaller tokens that are used for the training of data and to generate an image. The process is not controlled by the user because it is not possible to predict what Midjourney will create ahead of time.

USCO concluded that the images generated by Midjourney are not original works of authorship protected by copyright, since Midjourney generates images in an unpredictable way. The fact that Midjourney´s specific output cannot be predicted by users makes Midjourney different for copyright purposes from other tools used by artists for example editing tools and assistive tools that allow the choice of specific changes and include specific steps to control the final image by the user.

In the US District Court Of Columbia copyright case Thaler v. Perlmutter, the Court decided that human authorship is a fundamental requirement for copyright claim.

In the EU, courts have the same requirement of human authorship and require the ability of the author to control the creative process by making free choices. The Court of Justice of the European Union has considered the originality requirement in the cases C-5/08 Infopaq and C-145/10 Painer defining that copyright can only apply to a work which is original in the sense that it is its author's own intellectual creation. An intellectual creation is an author's own if it reflects the author's personality. That is the case if the author was able to express his creative abilities in the production of the work by making free and creative choices. Simple facsimile copies of original works, then would not be copyright works in their own right.

Example 4

Digital copies of historic artworks that are in the Public Domain are created to replicate the original as far as possible. The items are photographed and then the images are uploaded to a website.

These digital surrogates would not contain any originality as defined and would not attract a new, distinct copyright. The creator of the photograph is trying to copy the original, they are not making free creative choices. They are not creating their own creative outputs.

There is a parallel here with generative AI perhaps.

Example 5

In the Czech decision by the Prague Municipal Court 11.10 2023, a person (claimant) filed a lawsuit against a law firm (defendant) after it published an image on its website without the claimant's permission. The claimant had created an image with the generative AI service DALL-e. The image shows two hands signing a business contract, and it had been created by the person using the following text prompt : "create a visual representation of two parties signing a business agreement in a formal environment; for example, in a conference room or a law office in Prague. Show only the hands."

In the Court's opinion, the image cannot as a matter of principle be protected by copyright, because such an image is not the result of creative activity of a natural person – an author. The Court concluded that the image is not a copyright-protected work. The claimant argued that the image was created on the basis of his specific prompt, which justified the claimant's copyright to the image. The Court considered that the prompt itself could only be regarded as a theme or idea for a work, neither of which can be protected by copyright. The Czech Copyright Act specifically lists themes, ideas and similar more abstract concepts as excluded from copyright protection, which is a principle generally recognised internationally.

Hands-on activity and other self-guided tutorial(s)

This guide contains a ton of resources for learning about copyright but for something a little bit different, check out the Copyright Card Game or online quizzes such as European Copyright 10 Questions.

Recommended Reading/Viewing

Copyright and AI is a constantly moving area and it can be hard to keep up with the latest developments but we can recommend following the The IPKat (ipkitten.blogspot.com) and Kluwer Copyright Blog (kluweriplaw.com) and News from the LIBER Copyright & Legal Matters Working Group for excellent updates.

Much guidance is based on national law, so you will need to keep an eye on copyright legislation, which will provide an overview of the local picture. These international networks and organisations are also useful places for getting more in depth information:

- WIPO
- EU Intellectual Property Office
- Creative Commons
- KR21
- Communia Association

Finding Communities of Practice

If you are employed in an institution that is a member of LIBER do join or reach out to the LIBER Copyright & Legal Matters Working Group, a group of librarians, lawyers, professors and communications professionals who monitor current European law and react to proposed changes, on behalf of libraries, archives, researchers and students.