From Open Access to Open Science

A European legal perspective with a specific focus on licensing

OpenAccess Week

OpenAire Seminars

Dr. Thomas Margoni
Senior Lecturer in Intellectual Property and Internet Law
Director of the LLM in Intellectual Property and the Digital Economy
School of Law – CREATe Centre – University of Glasgow
Legal Coordinator OpenMinTeD

openMINTED
Open Mining Infrastructure for Text & Data

thomas.margoni@glasgow.ac.uk

Open Access

"… free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited" (Budapest Open Access Initiative)

Possible Issues

1) Copyright exists in most cases where articles, publications, datasets, etc are created;

2) SGDR and other rights even in absence of originality;

3) Limited and fragmented presence of ELCs, absence of broad standards such as fair use in US

4) Other legal hurdles

5) Licences may work but are not the perfect solution

6) Examples

Legal barriers

Copyright and rights related to copyright (e.g. Sui generis database right (SGDR))

- These rights usually restrict the reproduction (copy) and distribution of protected works and databases with substantial investment (e.g. Art 2 InfoSoc Directive and Arts. 5 and 7 Database Directive)
- Problem: reproduction is defined very broadly by EU law (any temporary or permanent copy of the whole or part of a work, etc); SGDR restricts copies of **substantial parts** and repeated copies of insubstantial parts
- Therefore any TDM (or any other act) which requires any temporary copy of the original work or DB or part thereof infringes protected works and/or SGDR

    - **Privacy/data protection**
      Protects personal data (e.g. databases containing names, addresses, age, sex, etc).
      One of the most important elements is the concept of **consent:** data subject can give consent for treatment of his/her data (e.g. in a DB). But such consent needs to be specific for a purpose. Consent cannot be given for any type of use (like e.g. copyright licences). Therefore, all data subjects may have to give their consent for every new use, something difficult to foresee in an open research environment (Open Science)
    - **PSI**
      Public Sector Information legislation is based on a different paradigm than other approaches (e.g. U.S. where works of Federal Government are not protected in the U.S.). PSI 2013 has an “open by default” approach but copyright and other similar rights and personal data are object of specific exclusion and therefore PB are under no obligation to make them accessible and/or reusable. Plus, FoA remains MS competence.
    - **Contracts/terms of use**
      Even when no rights exist on a specific BD (because there is no originality, no substantial investment, no personal data, etc) terms of use of data provider may restrict use and redistribution of DB. This limitation is based on a contractual relationship but is still an enforceable obligation (although there are differences). See ECJ in Ryanair v PR Aviation

Exceptions to legal barriers:

- **Copyright and rights related to copyright**
  - Exception and limitations to copyright (ELC), fair dealing, fair use. ELC are only partially harmonised (e.g. in EU 1 mandatory plus 20 at discretion of MS). Internationally, even more differences.
  - For TDM in EU possible exception for research and teaching. Problem: it is not uniformly implemented in all MS and it is often limited to partial copies. It is also limited to non commercial activities and only for illustration for teaching and research. Art. 5(1) is mandatory but limited in scope. Absence of general open norm (e.g. US fair use; UK fair dealing is narrower)
  - Recently, UK introduced a limitation to copyright and related rights for acts of TDM for non commercial purposes and for legally accessed sources on the basis of the EU ELC for research. In draft for a Directive for Copyright in DSM EC has introduced a mandatory TDM exception, not limited by contracts (but yes by TPM) which is only available to research organisations (contrast this with e.g. US where most TDM are considered “transformative” uses, therefore covered by fair use).

- **Privacy/data protection**
  Anonymisation of data (removal of personal data) but this is time/money consuming and may reduce the usefulness of DB
- **PSI**
  PSI legislation does not affect FoA (Freedom of Access) legislation which is MS power. But if MS empower FoA legislation then PSI “reusable by default” rule applies. However, limitation regarding copyright and personal data still applies
- **Contracts/terms of use**
  These are private agreements so there are no real exceptions. However, certain regulations (antitrust, abusive clauses, consumer protection) could under certain circumstances invalidate specific terms. This is however a case per case issue and does not seem to constitute a sound course of action.

Licences and licence compatibility

- Licences are permissions/authorisations (contract or otherwise based) that allow one or more parties to perform certain activities.
- Licences (so called esp. in the field of copyright) may be directed to a plurality of subjects and be drafted in standard forms or had hoc
- Some licences are usually called public licences (e.g. CCPL = Creative Commons Public Licence, GPL = General Public Licence, etc).
- In certain fields Open Content Licences (e.g. CCPL, CC0, EPL, etc) are used to grant a permission to perform acts (copy, redistribute, modify, etc) in relation to a work of authorship or other subject matter (e.g. a DB), under certain conditions (Attribution, Non Derivatives, Share Alike, Non Commercial, etc).
- A possible problem is “licence proliferation”, i.e. too many (and possibly incompatible) licences. Therefore, there is a general consensus that new licences should not be created unless really necessary.
- Some projects (e.g. OpenMinTeD, OpenAire) promote Legal Interoperability through analysis of legal documents and compatibility matrix.

Inner limits of licences

- Licences are a powerful instrument but not perfect…

  - “Private ordering tool” i.e. can we entrust a private law tool with a function that should be a matter of public interest/intervention (wider access to knowledge)?
  - Licences are a voluntary tool, i.e. only if the owner of Work/DB is willing to grant you access, licences work. If work owner says no, there is no remedy based on contracts that can force him/her to deal with you.
  - Even if DB is willing to employ licences, very often there are problems of correct labelling (legal code, metadata, etc) of resources. This is a very serious issue faced in many projects in TDM and in science/academia.

Policy recommendations best practices

- Through proper policy choices some of other disadvantages can be fixed.
  - Recommending 1 or a very limited no. of licences which are **compatible** (fixing problem of licence incompatibility)
  - Crucial importance that **data providers, funding agencies, scientific and public institutions** require use of correct licences and subject grants or funding to the **correct implementation** of those licences (fixing problems of “voluntarity” and “labeling”)
  - **Influence public debate** so that legislative intervention in the field is appropriate (e.g. definition of right of reproduction, harmonisation of ELC, need of a broader standard for ELC, limit of non commercial exception such as in UK).
  - Many projects in EU (e.g. OpenMinTeD) focus on OA resources given the complex legal issues (market failure?) connected with TDM.

Policies, best practices and OA requirements

Examples:

- H2020 funded projects must be published in OA
- H2020 has also an OA data pilot which should become non optional
- National funding bodies and assessing bodies only consider OA publications for grant applications or for tenure, scientific assessment, promotions, etc.
- Scientific foundations require OA publishing.

Open Access

Often this contrast with traditional academic publishing where publishers commonly require a copyright transfer/exclusive licence from the authors in order to build a business model based on paid distribution of hard copies or access to online versions

Open Access

Some countries (e.g. DE, NL) create a termination of transfer of rights in order to republish in OA (although with limitations)

Other countries pass laws that are hard to assess due to the fact that it is hard to understand the intended legal effect (don’t address IP) and the recipients of the legal obligation (e.g. IT, ES).

Open Access

But there is more, e.g.

- Open Methodology (reproduceability of scientific results and preregistration)
- Open Peer-review (biases in the composition of reviewing committees and influence of “schools”)
- Open Citations (lock-in of scientific databases and lack of transparency)
- Open Data (SGDR and non protected DB and protection of non original data)
- FLOSS (software as results and as tools)

Is this still just about (open) Access?

This is much more, not only access to science but about science itself:

Open Science

From Open Access to Open Science

Open Science includes all these features:
Open Access, open methodology, open peer review, open citations, etc.

And has a number of goals:

- Efficiency
- Transparency
- Accountability
- Impact
- Diffusion
- Access
- Innovation

Open Science

Is it/should it be more than an umbrella concept?

Propositive concept that not only collects the concepts aforementioned but offers guidance and value-based normative concept about how rules and norms within science should be regulated

Open Science and IP

In the field of IP (copyright) we should consider the following:

- 1 Subject matter (copyright at all?)
- 2 Authorship and Ownership (who should own scientific outputs and results)
- 3 Rights (should right of reproduction be as broad as it is now? Communications to the public? Modification?)
- 4 Exceptions and limitations (new paradigm?)
- 5 Relationship between copyright, contracts and scientific norms
- 6 Reconceptualisation of the relationship between authors' rights and users' rights

Example: OpenMinTeD

- The global research community generates over 1.5 million new scholarly articles per annum.

  The STM report (2009)

- ... some 90% of papers ... are never cited.
  ... 50% of papers are never read by anyone other than their authors, referees and journal editors

  Lokman I. Meho, The rise and rise of citation analysis, 2007

- ... one paper published every 30 seconds

  Spangler et al, Automated Hypothesis Generation based on Mining Scientific Literature, 2014

From: OpenMinTeD 2016

thomas.margoni@glasgow.ac.uk

Example: OpenMinTeD

Machine reading

process textual sources, organise and classify in various dimensions, extract main (indexical) information items,

… and “understanding”

identify and extract entities and relations between entities, facilitate the transformation of unstructured textual sources into structured data

… and predicting

enable the multidimensional analysis of structured data to extract meaningful insights and improve the ability to predict

thomas.margoni@glasgow.ac.uk

Example: OpenMinTeD

	GPLv3	GPLv2	Apachev2	EPLV1	LGPLv3	LGPLv2	AGPLv3	GNUAllPermissive	GFDLv1.3	MPLv2.0	ModifiedBSD / 3-Clause BSD	SimplifiedBSD / 2-Clause BSD	Expat / MIT
GPLv3	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
GPLv2		Yes	No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Apachev2			Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
EPLV1				Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
LGPLv3					Yes	Yes	Yes	Yes	https://docs.g...t?usp=sharing		Yes	Yes	Yes
LGPLv2						Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
AGPLv3							Yes	Yes	No	Yes	Yes	Yes	Yes
GNUAllPermissive								Yes	Yes	Yes	Yes	Yes	Yes
GFLDv1.3									Yes	No	Yes	Yes	Yes
MPLv2.0										Yes	Yes	Yes	Yes
ModifiedBSD / 3-Clause BSD											Yes	Yes	Yes
SimplifiedBSD / 2-Clause BSD												Yes	Yes
Expat / MIT													

https://openminted.github.io/releases/license-matrix/

openMINTED
Open Mining Infrastructure for Text & Data

creative commons UK

FACT SHEET ON CREATIVE COMMONS & OPEN SCIENCE V.0.1

This information guide contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licences. It is intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licences in their work.

What is Open Science?

Open Science is the movement to make scientific research and data accessible to all for knowledge dissemination and public reuse.

How should I licence my data for the purposes of Open Science?

We recommend you use the CC0 Public Domain Dedication, which is first and foremost a waiver, but can act as a licence when a waiver is not possible.

CC ZERO LICENCE, 'NO RIGHTS RESERVED' LOGO

By applying CC0 to your data you enable everyone to freely reuse your data as they see fit by waiving (giving up) your copyright and related rights in that data.

You should keep in mind that there are many situations in which data is *not* protected as a matter of law. Such data can include facts, names, numbers – things that are considered 'non-original' and part of the public domain thus not subject to copyright protections. Similarly, your database (which is a structured collection of data) might be considered 'non-original' and thus ineligible for copyright, and it might additionally be excluded from other forms of protection (like the EU *sui generis* database right, also known as the 'SGDR', for non-original databases).

In these cases, using a Creative Commons licence such as a CC BY could signal to users that you claim a copyright in the non-original data despite the law, and perhaps despite your real intention.

Finally, if your data is in the public domain worldwide, you might state simply and obviously on the material that no restrictions attach to the reuse of your data and apply a Public Domain Mark.

PUBLIC DOMAIN MARK LOGO

When in doubt, consider which use may be appropriate according to the chart below:

CC0 & PUBLIC DOMAIN LICENCES
WHICH LICENSE TO USE AND WHEN

CC ZERO	PUBLIC DOMAIN
'Creative arrangement' of data is original, but any copyright has been waived and content is made available copyright-free	'Creative arrangement' of data is not original; the author acknowledges this and communicates the data is in the public domain

https://zenodo.org/record/841086#.WYwTWYpLdE4

https://zenodo.org/record/840652#.WYwTcopLdE6

openMINTED
Open Mining Infrastructure for Text & Data

But I would like attribution when others use my dataset. In that case, shouldn't I use a CC BY licence?

We recommend that you avoid using a CC BY licence. Here's why:

While attribution is a genuine, recognisable concern, not only might using a CC BY licence be legally unenforceable when no underlying copyright or SGDR protects the work, but it may also communicate the wrong message to the world. A better solution is to use CC0 and simply ask for credit (rather than require attribution), and provide a citation for the dataset that others can copy and paste with ease. Such requests are consistent with scholarly norms for citing source materials.

Legally speaking, datasets that are ***not*** subject to copyright or related rights (and are thus in the public domain) cannot be the object of a copyright licence. Despite this, agreements based in contract law may be enforceable. Creative Commons licences, however, are copyright licences. Therefore, where the conditions for a copyright or related right are not triggered, copyright licences, such as the CC BY licence, are unenforceable.

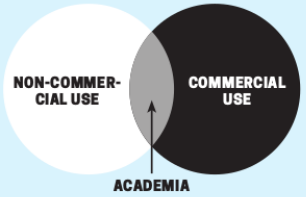
In some cases, however, rights may exist (like the *sui generis* database right previously mentioned), and permission for others to use your dataset may be legally required. These rights are meant to protect the maker's investment, rather than originality. As such, database rights do not include the moral right of attribution. So by using a CC BY licence, you signal to users that you restrict access to your dataset beyond the protections provided by the law. We are not saying that this cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

I'm uncomfortable with others using my research for commercial purposes. Should I use a non-commercial licence for my dataset?

We recommend you avoid using a non-commercial licence. Here's why:

For legal purposes, drawing a line between what is and is not 'commercial' can be tricky; it's not as black and white as you might think. For example, if you release a dataset under a non-commercial licence, it would clearly prohibit an organisation from selling your dataset to others for a profit. However, it might also prohibit someone using the dataset in their research if they intend to eventually publish that research. This is because most academic journals are commercial businesses that charge some sort of fee for access to their content, hence, such use could qualify as 'commercial'. Consequently, using a non-commercial licence prevents researchers from using your data in work destined for publication. This can subsequently affect the dissemination, recognition, and impact of your dataset.

BECAUSE WHO DOESN'T LOVE A GOOD VENN DIAGRAM?



Please also consider that the current definition of 'Open Access' in the relevant international declarations states that limiting reuse to non- commercial activities does ***not*** comply with 'Open Access' (see the Berlin Declaration, Bethesda Statement on Open Access Publishing, and Budapest Open Access Initiative).

Ultimately, the decision is yours. However, the better open science practice is to avoid restricting use of your dataset to only non-commercial use.

I'm uncomfortable permitting use of my research for any and all purposes. Should I use a 'No Derivatives' (ND) licence for my dataset?

We recommend you avoid using a 'No Derivatives' licence. Here's why:

Similar to how a non-commercial licence might restrict meaningful reuse of your dataset, a ND licence can have the same effect: it may prevent someone from recombining and reusing your data for new research. For data to be truly Open Access, it must permit these important types of reuse.

What happens if I use 'Share Alike' (SA) licensed material in my work? Does that mean I have to make my work available under the same SA licence?

Not necessarily, but it depends on how you use the SA licensed content.

A 'Share Alike' CC licence applies only to the content licensed as SA that you have used. It does not require you to also make your work available under a SA licence, so long as you have not combined the independent works into one new work (known as a 'derivative' work).

When using SA content in your work, be sure to maintain the SA licensing information in regards to the content used. This can be done by providing the SA licensing information next to the content in your work and by designating it as SA when listing the other restricted content in your rights statement.

For example, if you include a CC BY-SA dataset in your research, you do not have to licence the entire body of work under a CC BY-SA, but the CC BY-SA dataset must retain the original licence. However, if you create a new dataset by combining two existing datasets, one of which belongs to you and the other is licensed under a CC BY-SA, then the new work (a derivative work) must be licensed CC BY-SA.

We understand that might be confusing, so here's an illustration to help:

NAVIGATING MULTIPLE LICENCES AND MAINTAINING RIGHTS INFO

It sounds like you're really pushing for the use of CC0 for open science datasets.

Exactly. Data is only open if anyone is free to use, reuse, and distribute it. This means it must be made available for both commercial and non-commercial purposes under non-discriminatory conditions that allow for it to be modified.

When data is made available for all reuse, others can create new knowledge from combining it. This leads to the enrichment of open datasets and further dissemination of knowledge. Accordingly, CC0 is ideal for open science as it both protects and promotes the unrestricted circulation of data.

And remember, it's bad science not to cite the source of data you use. To help others cite your data include a citation that users can copy and paste to give you credit for your hard work.

For example, the citation for this document is:

'Fact Sheet on Creative Commons and Open Science', Creative Commons UK, DOI: 10.5281/zenodo.840652, CC BY 4.0, https://creativecommons.org/licenses/by/4.0/

After reading this document, should you still wish to use CC BY make sure to include the citation for your dataset so others may cite your work with ease.

'Fact Sheet on Creative Commons and Open Science', Creative Commons UK, DOI: 10.5281/zenodo.840652, CC BY 4.0, https://creativecommons.org/licenses/by/4.0/

Support for this publication was provided through the University of Glasgow's College Strategic Research Major Initiatives Fund (ES/M500471/1). This guide is for informational purposes only and may not apply to your specific case. It does not constitute legal advice.

The font used is Cooper Hewitt, an open source typeface designed by Chester Jenkins and commissioned by the Cooper Hewitt museum.

Example: Open Science check list for repositories

1) Apply the right licence to **your repository**

2) Don’t forget the **metadata**

3) Apply the right licence **also to the content** of your repository (not the same thing as point 1)!

4) In particular, **CC BY 4.0** for works such as papers, articles, monographs, creative images,

etc)

5) Data and dataset should be under a **CC0** (or a Public Domain Dedication)

6) **Require** that uploaders choose a licence when they upload their content

7) Suggest which licence **should be chosen in order to meet OS** requirements (see above)

8) Explain why what you recommend is the best choice and why other choices are not good **but let uploaders choose**

openMINTED
Open Mining Infrastructure for Text & Data

Open Science

Thanks!

thomas.margoni@glasgow.ac.uk

@openminted_eu

openMINTED
Open Mining Infrastructure for Text & Data