Paving the way for text and data mining in science

To TDM or not to TDM?

OpenMinTeD
Thursday, May 24, 2018 Brussels, Belgium

Dr. Thomas Margoni
Senior Lecturer in Intellectual Property and Internet Law
School of Law – CREATe
WG3 – Legal Interoperability - OpenMinTeD
University of Glasgow

thomas.margoni@glasgow.ac.uk

Not to TDM

Current EU legal framework

Broad definition of right of reproduction (and redistribution, communication to the public, etc) NOT counterbalanced by broad set of exceptions and limitations (fragmented, not mandatory, if mandatory of unclear scope e.g. 5(1), narrow interpretation, etc)

thomas.margoni@glasgow.ac.uk

Not to TDM

Future (sic!) EU legal framework

Goal: moderinse EU copyright law and make it fit for the digital age in the DSM

How: In Draft Directive on Copyright in the DSM with a number of provisions (in particular 8 proposals that will change EU copyright law).

thomas.margoni@glasgow.ac.uk

Not to TDM

The most problematic are:

- **TDM exception (Art. 3);**
- Protection of press publications concerning digital uses (Art. 11);
- Use of protected content by information society service providers storing and giving access to large amounts of works and other subject-matter uploaded by their users (Art. 13).

thomas.margoni@glasgow.ac.uk

Or just a little bit TDM

1) Text and Data mining: any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations;

2) **Scope**: exception to the right of reproduction;

3) **Beneficiaries**: research organisations with lawful access for research purposes;

4) **Relationship to contracts**: Cannot be limited by contract;

5) **Relationship to technology**: Can be limited by technological measures (integrity measures and TPM)

thomas.margoni@glasgow.ac.uk

For some purposes but not for others

1) **Text and Data mining**: any automated analytical technique aiming to analyse text and data in digital form in order to generate <u>information such as patterns, trends and correlations</u>;

Comment: definition is broad enough to cover current TDM practices.

thomas.margoni@glasgow.ac.uk

For some uses but not others

2) **Scope**: exception to the right of reproduction;

Comment: Problematic. It does not cover rights of redistribution/communication to the public and adaptation (derivative works). It means that all the times that the results of TDM are a copy in part of a protected work (Art. 2 Infosoc as interpreted by CJEU in Infopaq says that even 11 consecutive words can infringe) or when the results can be an adaption (derivative) of the original (thumbnails?) the exception is not available.

thomas.margoni@glasgow.ac.uk

For some beneficiaries and not for others

3) **Beneficiaries**: research organisations with lawful access for research purposes;

Comment: Problematic. Individuals, micro and SMEs, industry, etc cannot benefit even if acting non commercially. Purposes other than research (e.g. journalism, criticisms, review, etc) are not covered. Why? Potential contrast with fundamental rights?

thomas.margoni@glasgow.ac.uk

Overridable by TPM

4) **Relationship to contracts**: Cannot be limited by contract;

5)Relationship to technology: Can be limited by technological measures (integrity measures and TPM).

Comment: 4) is good. But 5) is contradictory. It creates imbalance and uncertainty with regards to the medium through which a prohibition is expressed. If “exception not available” is expressed in human/legal language (contract) this is not enforceable, but if the same condition is expressed in computer language (DRM or TPM) then it is allowed. 5) basically circumvents 4) in a way that is unreasonable, not proportionate and harmful for consumers.

thomas.margoni@glasgow.ac.uk

TDM and Copyright Law: the absurtidy

TDM normally extracts principles, facts, data, correlations, etc which are not protected by copyright law (Art. 2 WCT, 9(2) TRIPs, but also generally in Berne and most legal traditions).

Thus the extraction of those unprotected elements from protected works should not need an exception if copyright framework was properly designed.

Main problem with EU copyright design is that it is not properly designed: it harmonises broadly rights (reproduction, redistribution, communication to the public, etc), but does not do the same with exceptions (exhaustive but not mandatory list, narrow interpretation, etc). The current proposal does not fix this design problem.

thomas.margoni@glasgow.ac.uk

TDM and Copyright Law: the possible (<u>but not ideal</u>) solution

- **Now**: Implement a TDM exception not limited to research organisations for research purposes (i.e. “option 4” of the impact assessment p. 108 – 109).

  Comment: This will only fix some of the problems identified above, but it could be technically be done in the present draft (although it seems that none of the proposed amendments is in this direction).

- **Tomorrow**: A better drafted EU copyright law clearly marking the boundaries between protection and PD, e.g. through an open and flexible norm that will cover TDM but also future technological advancements.

  Comment: This will allow courts to readily balance investment and innovation needs without having to wait for legislative intervention. The latter has caused a major delay in EU development of TDM and connected technology sectors in comparison to other more innovation oriented jurisdictions (US, Canada, Singapore, Japan, etc).

thomas.margoni@glasgow.ac.uk

Yes TDM

OpenMinTeD!

thomas.margoni@glasgow.ac.uk

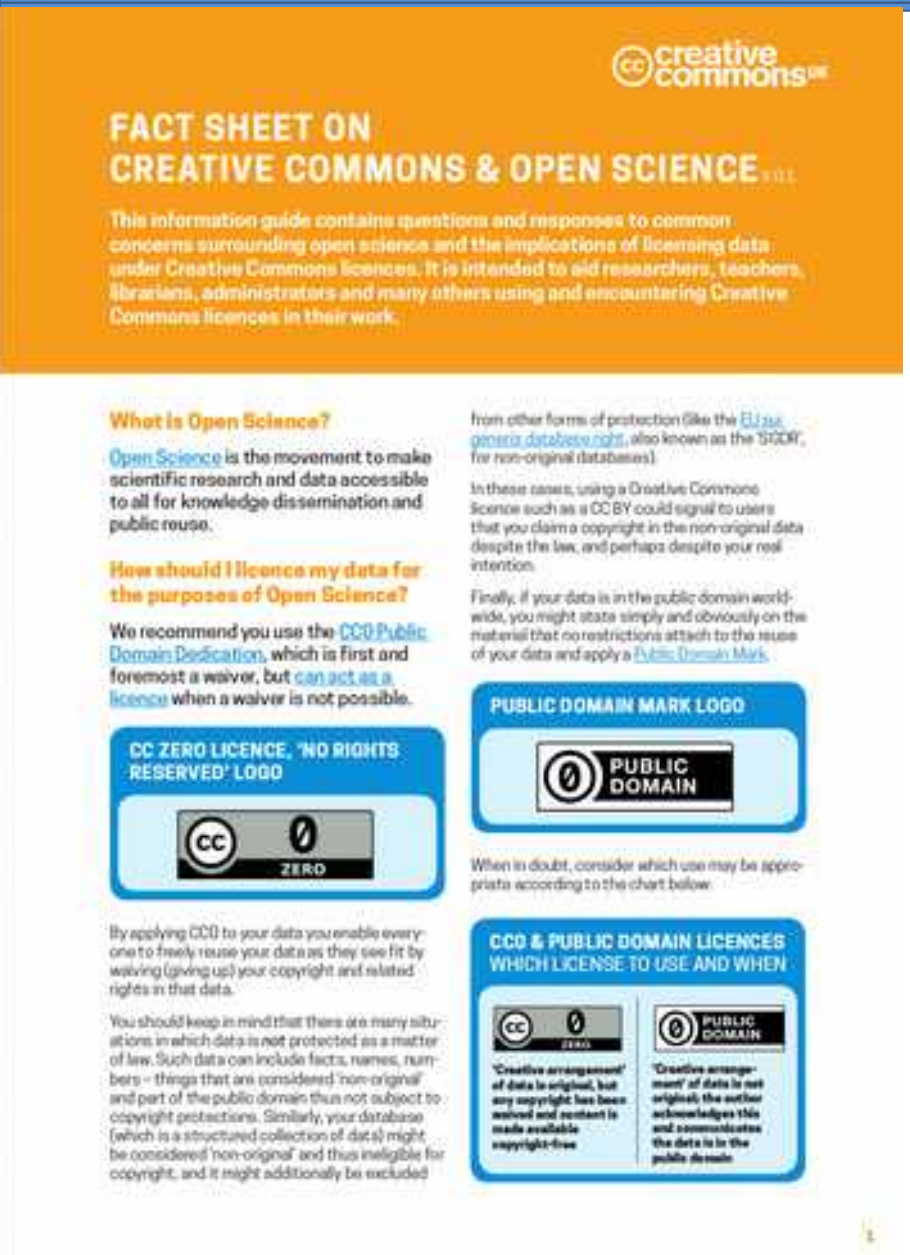
• Example: OpenMinTeD licence compatibility tools

	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...(?)upgrade-sharing LC		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

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

thomas.margoni@glasgow.ac.u

Example: Fact-sheet and FAQs on licensing, OA, OS, etc.



creative commons

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 'SIGDR', 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

'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

thomas.margoni@glasgow.ac.u

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/M600471/1). 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.

thomas.margoni@glasgow.ac.u

Example: OpenMinTeD Guidelines

- GitBook WE ARE HIRING! Pricing Explore About Blog Sign In Sign Up

pennyl67 > OpenMinTeD guidelines

Updated a day ago

ABOUT 42 DISCUSSIONS 0 CHANGE REQUESTS ★ Star 2 Subscribe 1

Read

OpenMinTeD guidelines

Welcome to the OpenMinTeD Guidelines!

This is where you'll find information on the following:

- how to make your resources interoperable with other resources for Text and Data Mining (TDM) purposes
- how to share your resources through the OpenMinTeD platform [1]
- how to build new TDM applications with existing components
- how to give us feedback and contribute to the guidelines.

What is OpenMinTeD?

OpenMinTeD enables the creation of an infrastructure that fosters and facilitates the use of **Text and Data Mining (TDM) technologies** in the world of **scholarly content**, builds on existing TDM tools and platforms, and renders them discoverable and interoperable through a dedicated registry and a standards-based interoperability layer.

openMINTED

• 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**

thomas.margoni@glasgow.ac.u

openMINTED
Open Mining Infrastructure for Text & Data

• 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**

thomas.margoni@glasgow.ac.u

openMINTED
Open Mining Infrastructure for Text & Data

Additional info

More info including full references and data is in forthcoming paper. A draft-preview in blog form is available here:

http://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/

Additional information about the other provisions (especially press publishers rights and intermediary liability) is here:

http://www.create.ac.uk/blog/2018/04/26/eu_copyright_directive_is_failing/

A recent paper for a natural language processing conference briefly discussing whether current Art. 5(1) (temporary acts of reproduction) can be used for TDM purposes

thomas.margoni@glasgow.ac.uk