Open Mining Infrastructure for Text and Data OpenMinTeD

11th International Conference on Open Repositories

Dr. Thomas Margoni
Lecturer in Law – University of Stirling & CREATe – University of Glasgow, UK
OpenMinTeD coordinator of Legal Interoperability WG

UNIVERSITY of STIRLING

openMINTED
Open Mining Infrastructure for Text & Data

thomas.margoni@stir.ac.uk

WG3 Legal Interoperability & Open Repositories

- **Goal**
- **What legal barriers to TDM**
- **Exceptions to legal barriers (and limits)**
- **Licences**
- **Legislative reform**
- **Policy choices**
- **Deliverables: multi-layer compatibility matrix (based on theoretical analysis)**

UNIVERSITY of STIRLING

thomas.margoni@stir.ac.uk

WG3 Legal Interoperability

Goal:

The goal WG3 is to study and identify copyright and related rights (e.g. sui generis database right) restrictions and exceptions to the use and reuse of sources (both textual sources and text-mining services) in TDM activities. On this basis the WG will also identify contractual tools and schemes (e.g. licences) that can best serve the needs of TDM services.

UNIVERSITY of STIRLING

thomas.margoni@stir.ac.uk

What legal barriers to unrestricted TDM

- **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 or databases with substantial investment
  - 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 DB

  - **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 a OA environment
  - **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 a “open by default” approach but copyright and other similar rights and privacy are object of specific exclusion and therefore PB are under no obligation to make them accessible and/or reusable
  - **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)

Exceptions to legal barriers:

- **Copyright and rights related to copyright**
  - Exception and limitations to copyright (ELC), fair dealing, fair use. ELC are limited (in EU 1 mandatory plus 20 at discretion of MS)
    - For TDM only possible one is exception for research and teaching. Problem is that it is not uniformly implemented in all MS and that it is usually limited to partial copies. It is also limited to non commercial activities and only for illustration for teaching and research. Fair dealing (e.g UK) is a broader standard but not as much as fair use (US).
    - Recently, UK introduced a limitation to copyright and related rights for acts of TDM for non commercial purposes and for DB legally accessed.

    - **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&C. and privacy 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. These are usually called public licences (e.g. CCPL = Creative Commons Public Licence, GPL = General Public Licence, etc)
- In the field of OA, Open Content Licences (e.g. CCPL) 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)
- Different type of licences in the OC field. A possible problem is “licence proliferation”, i.e. too many (and possibly incompatible) licences. Therefore, in the “open environment” there is a general consensus that new licences should not be created unless really necessary. One of the main goals of the Legal Interoperability WG in OpenMinTeD is to prepare a licence compatibility matrix.

Inner limits of licences

- Licences are a powerful instrument but not perfect…
- Examples of problems with licences:
  - “Private ordering tool” i.e. can we entrust a private law tool with a function that should be a matter of public interest (wider access to knowledge)?
  - Licences are a voluntary tool, i.e. only of the owner of Work/DB is willing to grant you access, licences work. If data 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 OA licences, very often there are problems of correct labeling (legal code, metadata, etc) of resources. This is a very serious issue faced in many projects in the OA field.

Policy recommendations and licences

- 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 grant of 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, limited amount of ELC, need of a fair use exception, limit of non commercial exception such as in UK).
  - Assessment: only OA publications can be used for assessment

Priority points

- There is no limit to creation of Licences and ToS (licence proliferation). This is generally bad as it raises transactive costs
  - We prioritize OA licences
  - We need a working definition of OA, but don't try to create yet another definition
- Major issue is absence of licences or the use of generic terms such as "this DB is in OA"
- Define the licence in legal terms but also in machine readable (e.g. metata)
  - CC is a good example for OC (Work with WG1)
- How far does SGDR reach, i.e. reuse of SGDR DB especially if you are only referring to statistical results (e.g. how many times a certain word is repeated) and limits of certain clause/conditions (e.g. non commercial).

Deliverables

- Multi-layer compatibility matrix:
  - Static matrix
  - “Calculators”
  - Ongoing study on best graphic representation of information
- Based on theoretical analysis:
  - So far 3 papers in progress
    - WG1&3: “Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview”
    - WG3: “Why we need a TDM exception (but it is not enough)”
    - WG3&FutureTDM: “Implications of the ‘non commercial’ limit in Text and Data Mining”