Text and Data Mining Exceptions in the New Directive on Copyright in the Digital Single Market

CLARIN ERIC
Thursday the 28th of October 2021
Organisers

This edition of the CLARIN Café is organized by

**Paweł Kamocki** chair of the CLARIN Legal and Ethical Issues Committee (CLIC).

and

**Vanessa Hannesschläger**, vice-chair of the CLARIN Legal and Ethical Issues Committee (CLIC).

CLARIN host is **Francesca Frontini** (CLARIN ERIC)
Plan

• 14:00 - 14:05 Opening and CLARIN 101 - Francesca Frontini (ILC-CNR and CLARIN )

• 14.05 - 14.35 TDM exceptions in the DSM Directive and how we got there
  – Dr. iur. Paweł Kamocki, IDS Mannheim, chair of the CLARIN Legal and Ethical Issues Committee

• 14.35 - 15.05 Challenges in transposing the TDM exceptions: national perspectives from selected Member States
  – Prof. Dr. Aleksei Kelli, University of Tartu, member of the CLARIN Legal and Ethical Issues Committee

• 15.05 - 15.35 TDM exceptions and language resources: some pressing issues from Czechia
  – Prof. Dr. Jan Hajič, Charles University Prague, member of the Standing Committee for CLARIN Technical Centres

• 15.35 - 16.00 Questions and discussion
  – Moderator: Dr. Vanessa Hannesschläger, DLA Marbach, vice chair of the CLARIN Legal and Ethical Issues Committee
Housekeeping Rules

The event is recorded for further dissemination purposes.

Keep your microphone switched off if you are not speaking.

Questions and comments? Put them in the chat box.
CLARIN 101
https://www.clarin.eu/content/clarin-in-a-nutshell
CLARIN ...

- is the *Common Language Resources and Technology Infrastructure*
- has the ESFRI **ERIC** status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing**
- is an integral part of the **European Open Science Cloud**
  - See [clarin.eu/eosc](http://clarin.eu/eosc)
Ready-To-Use Language Resources

The term language resource refers to a broad range of speech and language data types in machine readable form, as well as tools and services for the processing of language data.

Examples of language resources are: written or spoken corpora and lexicons, multi-modal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, multimedia databases, etc.

Following a longstanding tradition (Godfrey & Zampolli 1997), language resources also include software tools for the preparation, collection, management, or use of other resources. Examples of such tools are corpus management and exploration systems, OCR systems, pipelines, speech processing systems, machine translation systems, environments for manual annotation and evaluation.
CLARIN Resource Families

**Corpora**
- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

**Lexical Resources**
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

**Tools**
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

See also the info on the CLARIN Resource Families initiative: [https://www.clarin.eu/resource-families](https://www.clarin.eu/resource-families)
CLARIN in Data Types

- Newspaper archives
- Literary texts
- Social Media data
- Parliamentary records
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- …
CLARIN in Communities of Use

- Digital Humanities
- Linguistics and Philology
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- Teachers
- Industry and Professionals
- General Public
- .....
CLARIN today

- **68 centres**
- **22 members**: (AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI)
- **3 observers**: FR, UK, ZA
The Technical Infrastructure

Harvested Metadata → Indexing → Metadata Search
Virtual Language Observatory (VLO) → Selection → Data Analysis
Language Resource Switchboard (LRS)

Findable Accessible Interoperable Reusable

clarin.eu/fair vlo.clarin.eu switchboard.clarin.eu
The Knowledge Infrastructure

- A network of Knowledge Centres
  - [https://www.clarin.eu/content/knowledge-centres](https://www.clarin.eu/content/knowledge-centres)

- Dedicated Funding and Support
  - [https://www.clarin.eu/funding](https://www.clarin.eu/funding)

- The "Teaching with CLARIN" initiative
  - [https://www.clarin.eu/content/teaching-clarin](https://www.clarin.eu/content/teaching-clarin)

- Committees
  - Standards Committee
  - Legal and Ethical Issues Committee (CLIC)

[https://www.clarin.eu/content/clarin-for-researchers](https://www.clarin.eu/content/clarin-for-researchers)
Some Recent Activities on Related Topics

• 30 March 2021 - **CLARIN Café on the Rights of Data Subjects in Language Resources**

• 27-29 September 2021 - **CLARIN conference**
  – 2 sessions **Legal Issues Related to the Use of LRs in Research** (overall 5 papers)
  – Panel **The Role of Corpora for the Study of Language Use and Mental Health Conditions** (legal and ethical considerations discussed)

• Yesterday, 27 October 2021 - **Data Management for FAIR CMC corpora**
  – 11:00 - 11:30 Intellectual Property Rights and Legal Issues with CMC Corpora (Pawel Kamocki)
The café
CLARIN – CLIC

Legal and Ethical Issues Committee
https://www.clarin.eu/governance/legal-issues-committee

Legal Information Platform
https://www.clarin.eu/content/legal-information-platform

https://www.clarin.eu/content/bibliographyfurther-reading-legal-and-ethical-issues
TDM Exceptions in the DSM Directive and how we got there

Paweł Kamocki, IDS Mannheim
A long time ago in our Galaxy...

1710: 1st Copyright Act (Statute of Anne), An act for the encouragement of learning...

1787: US Constitution, Art. I, Section 8 clause 8:
[The Congress shall have power...] To promote the *Progress of Science and useful Arts*, by securing for limited Times to Authors (...) the exclusive Right to their respective Writings.

1879: US Supreme Court, *Baker v. Selden*
“[W]hilst no one has a right to print or publish [the author’s] book (...) any person may practice and use the art itself which he has described and illustrated therein”. 
More recently, still in our Galaxy

~1970s: copyright becomes ‘interested’ in research
  - first research exception in the Berne Convention, v. 1967

Late 1990s: beginning of computational analysis of unstructured data (not in relational databases)

1999: Marti Hearst ‘Untangling Text Data Mining’
Data Mining listed among ‘Emerging Technologies that will change the world in 2001’ by MIT

Copyright in the Information Society (InfoSoc) Directive

Broad definition of exclusive rights...

- authors shall have the exclusive right to authorise direct or indirect, temporary or permanent reproduction of their works by any means and in any form
- TDM (like any digital use) necessarily involves reproduction

...counterbalanced by a large catalogue of optional exceptions

- Member States may provide exceptions in case of use for the sole purpose of scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved [Art. 5(3)(a) InfoSoc]

- CJEU’s tendency to enlarge the scope of copyright (Infopaq 2009)
A while (20 human years) has passed...

2013: Stakeholder dialogue ‘Licenses for Europe’ (failed)
2013/2014: Public consultation on copyright reform

TDM exceptions based on InfoSoc (non-commercial research only!) adopted in some EU Member States:

- **UK** (2014)
  (lawful access, no sharing, overrides contracts)

- **France** (2016)
  (lawful source, public research only, no application decree = dead letter)

- **Germany** (2017)
  (sharing allowed for joint research or verification of results; at the end of the project, copies to be deleted or transferred to an archive; overrides contracts; fair compensation to be paid to a collecting society)
Meanwhile in other countries

US: TDM, also for commercial purposes, is *fair use*
- key elements:
  - TDM is a highly transformative use (added value)
  - TDM does not harm the market value of source material (note: limited public sharing in Google Books)

Japan: TDM is covered by a flexible exemption for ‘non-enjoyment’ purposes

14 September 2016: first draft published by the EC, accompanied by an Impact Assessment

- Options to address the TDM problem:
  1. No legislative change (self-regulation)
  2. Mandatory TDM exception for non-commercial research
  3. Mandatory TDM exception for research by public interest research institutions → Art. 3 DSM
  4. Mandatory TDM exception for research by anyone with lawful access → Art. 4 DSM

until April 2019: turbulent legislative process (controversies mostly unrelated to the TDM exceptions)

7 June 2019: entry into force (publication)

7 June 2021: deadline for transposition
Text and Data Mining - definition in the DSM Directive

‘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’ (Art. 2 DSM)

- very broad definition
- covers all sorts of NLP applications

SIDE NOTE: the wording of the German exception from 2017 referred specifically to the NLP terminology (‘corpus’).
Art. 3 DSM - TDM for scientific research (I)

**Mandatory** exception (Member States *shall* provide…)

**Beneficiaries** (who…):

- research organisations (universities, research institutions and their libraries):
  - carry out research on a non-profit basis or re-invest profits in research OR
  - public-interest mission recognised by the State (e.g. through funding)
  - public private partnerships expressly covered (Recital 11)
  - NOT: institutions controlled by commercial undertakings
- cultural heritage institutions (publicly accessible libraries and museums; archives, film or audio heritage institutions)
Art. 3 DSM - TDM for scientific research (II)

Permitted acts:

- reproductions of copyright-protected works (excl. software!) and databases
- reproductions of subject matter of related rights (e.g. scientific and critical editions, non-original photographs)
- extractions from databases (*sui generis* database right)
- NOT: communication to the public (sharing)

**Purpose:** TDM for scientific research

**Requirement:** *lawful access* to the source material

- subscription- or license- based access (incl. OA licenses)
- OR free availability online (‘gratis OA’)
- “the right to read is the right to mine”
- redundant? (CJEU, ACI Adam 2014 -- private copies can only be made from lawful sources)
- NOTE: access for *institution*, not individual researcher
Art. 3 DSM - TDM for scientific research (III)

Copies (corpora):

- should be stored with *appropriate level of security* (protection against unauthorised access)
- can be retained for research purposes, incl. verification

_Rightholders_ (publishers) can apply *proportionate measures* to ensure the security and integrity of their networks and databases.

Member States should encourage stakeholders to agree on best practices on both these aspects (stay tuned…?)
Art. 4 DSM - General TDM exception (I)

**Mandatory** exception (Member States *shall* provide…)

**Beneficiaries:** everyone (no limits)

**Permitted acts:**
- reproductions of copyright-protected works (incl. software) and databases
- reproductions of subject matter of related rights (e.g. scientific and critical editions, non-original photographs)
- extractions from databases (*sui generis* database right)

**Purpose:** TDM (for any purpose)

**Requirement:** lawful access
Art. 4 DSM - General TDM exception (II)

**Condition:** “the right to mine” has not been expressly restricted by rightholders in an appropriate manner (opt-out)

- “mineable by default”
- for content publicly available online - by machine-readable means, e.g. *robots.txt*

**Important:** copies (corpora) can only be retained for *as long as necessary for TDM*, and deleted afterwards

- no long-term storage for verification?
- very bad idea e.g. for AI applications (no access to data used to train the algorithms!)
Common provisions for Art. 3&4 DSM

General rule: contracts override exceptions

BUT Both TDM exceptions (Art. 3&4) override contracts

- very desirable feature, as access to most Internet content is contract-based (Terms of Service)
- BUT: Art. 4 DSM can be easily overridden by ‘appropriate measures’, (probably) including contracts...

Both TDM exceptions can override technological protection measures (TPM)

- the mechanism for requesting access to TPM-protected content is cumbersome and rarely used in practice
Challenges in transposing the TDM exceptions: national perspectives from selected Member States
Transposition of EU Directives - general remarks

- Article 288 TFUE:
  
  A directive shall be binding, as to the result to be achieved, upon each Member State to which it is addressed, but shall leave to the national authorities the choice of form and methods.

- Usually modification of national law is required.

- EC may sue a Member State in CJEU for non-transposition or inadequate transposition.

- A MS may be sued for damages by an individual who suffered a loss due to non-transposition or inadequate transposition (rare).
Transposition of the DSM Directive

Deadline for transposition: **7 June 2021**

NOTE: Art. 25 DSM: “Member States may adopt or maintain in force broader provisions, compatible with the exceptions and limitations provided for in [the InfoSoc Directive] for uses or fields covered by the exceptions or limitations provided for in this Directive”.

- e.g. combine Art. 3 DSM (TDM exceptions for research) and the ‘general’ research exception in Art. 5(3)(a) InfoSoc (permits reproduction and communication to the public for non-commercial research, with attribution of the source unless impossible)
Selected challenges in transposing the TDM exceptions (I)

- Is adaptation (e.g. digitisation) of source material permitted?
  - adaptation right is not harmonised in the EU, but it can be seen as a component of the reproduction right
- Use a different term / a different definition of TDM?
  - risk of inadequate transposition...
- Can the TDM output (e.g. a translation model) be freely shared? Is it a derivative work?
  - Probably to be answered by courts (CJEU?).
- Does “lawful access” need to be further defined?
  - Also: probably to be answered by courts...
Selected challenges in transposing the TDM exceptions (II)

- Under Art. 3 (research TDM), should mining of software be allowed? (It is under Art. 4…)
- Specific requirements for secure storage of copies under Art. 3 (research TDM)? For measures that publishers may implement to protect their databases and networks?
  - Probably not in the law, the Directive encourages a bottom-up approach.
- Under Art. 3 (research TDM), can copies be shared for verification purposes?
- Specify ‘appropriate measures’ for rightholders to opt-out from Art. 4 (general TDM)?
- More details about the retention of copies under Art. 4 (“for as long as necessary for the purposes of TDM”)
Austria

- Not yet implemented (currently under parliament revision)

- §42h Text- and Data- Mining
  - Everyone may, on behalf of a research or cultural heritage institution, reproduce a work in order to use it to automatically evaluate texts and data in digital form for scientific or artistic research and to obtain information about, among other things, patterns, trends and correlations.
  - prerequisite: lawful access
  - limited to reproduction
  - storage for as long is justifiable by the research purpose (including evaluation; but presumably only for specific research purpose/project)
  - appropriate safeguards (e.g. "standards and best practices agreed upon by representative associations and institutions")
  - individual use possible if lawful access and no explicit ban on reproduction ("Nutzungsvorbehalt"); in this case, the storage limitation only relates to the analysis and information retrieval duration!
Italy

- Not yet implemented (currently under parliament revision)

- In the draft law proposed by the government artt. 70-*ter* and 70-*quater* are introduced to the Italian copyright law in order to rule TDM exceptions
- The wording of artt. 70-*ter* and 70-*quater* is substantially identical to that one of artt. 3 and 4 of the digital copyright directive
- Relevant remarks:
  - contractual provisions in contrast with the TDM exceptions for scientific research are void
  - copies of works or other subject matter made for purposes other than scientific research shall be stored with an appropriate level of security
Slovenia

- Not yet implemented (currently under parliament revision)

Public debate since 2019 - Ministry of Culture invited interested stakeholders (National university Library, Research Institutes, Universities)

Suggestions:
- TDM → Data Analysis (broader)
- Respond from rights holders within 72 hours
- No regulation on data storage specifics
- Remote access for TDM
- permission for public release of TDM results
**Germany (I)**

- Implemented by law of May 31, 2021 (entered into force: 7 June 2021)
- TDM definition slightly modified:
  - DSM: “analysing text and data in digital form”
  - German law: “analysing individual or multiple digital or digitised works”
- Art. 3 DSM (+ Art. 5(3)(a) InfoSoc) → 60d UrhG
  - ‘citizen scientists’ pursuing non-commercial research can also benefit from the exception (5(3)(a) InfoSoc)
  - sharing “with individual persons for verification” or “with a limited circle of persons for joint scientific research” allowed, but only by non-commercial institutions (5(3)(a) InfoSoc); upon completion of joint research, sharing must end
  - if sharing → obligation to mention the source, unless impossible (5(3)(a) InfoSoc)
- Art. 4 DSM → 44b UrhG
  - specifies that for online content, opting-out requires machine-readable form (the Directive does not seem categorical about this)
- ‘solely technical’ modifications of source material for TDM allowed (‘carved out’ from the adaptation right, 23 UrhG)
TDM exceptions and language resources: some pressing issues from Czechia

Jan Hajič, Charles University Prague
The legal situation

Copyright Directive 2019/790 (by Ministry of Culture)
- **Not** adopted yet (deadline June 7 2021 missed)
- Sitting in the Parliament since May 2021
- Parliament dissolved after regular elections Oct. 20, 2021
- (Old) government sent identical version (with formal changes only) to (new) parliament again Oct. 25, 2021
- (New) parliament convenes Nov. 8, 2021
- There will be ~100 government-sponsored waiting in line
- No expected date for floor discussion

Until then:
- Current Copyright Act 121/2000
- Research exception only
  - … plus the usual ones: citation, libraries, disabled, education etc.
The Copyright Directive transposition (CZ)

Implemented as amendment of the Copyright Act 121/2000

TDM-related paragraphs (translation on the next slides):

§ 39c Licence k rozmnožování díla pro účely automatizované analýzy textů nebo dat

(1) Do práva autorského nezasahuje ten, kdo zhotoví rozmnoženinu díla za účelem automatizované analýzy textů nebo dat v digitální podobě, prováděné za účelem získání informací, zahrnujících mimo jiné vzory, tendence a souvztažnosti; takto zhotovenou rozmnoženinu je oprávněn uchovat pouze po dobu nezbytnou pro účely této automatizované analýzy textů nebo dat.

(2) Ustanovení odstavce 1 se nepoužije pro rozmnoženiny díla, jehož autor si užití podle odstavce 1 výslovně vyhradil vhodným způsobem; v případě díla zpřístupněného podle § 18 odst. 2 strojově čitelnými prostředky.

(3) Ustanoveními odstavců 1 a 2 není dotčeno ustanovení § 39d.

§ 39d Licence k rozmnožování díla pro účely automatizované analýzy textů nebo dat k vědeckému výzkumu

Do práva autorského nezasahuje

a) vysoká škola, která jako součást své činnosti provádí vědecký výzkum, nebo právnická osoba, jejímž hlavním cílem je provádět vědecký výzkum nebo vykonávat vzdělávací činnost zahrnující rovněž vědecký výzkum, jestliže je vědecký výzkum této vysoké školy nebo právnické osoby prováděn tak, aby přístup k jeho výsledkům nebyl přednostně umožněn tomu, kdo na tuto vysokou školu nebo právnickou osobu vykonává rozhodující vliv, a současně tak, aby výzkum byl prováděn ve veřejném zájmu nebo na neziskovém základě nebo tak, že všechny zisky jsou zpětně investovány do vědeckého výzkumu této vysoké školy nebo právnické osoby, nebo

b) instituce kulturního dědictví,

zhotoví-li pro účely vědeckého výzkumu rozmnoženinu díla za účelem automatizované analýzy textů nebo dat v digitální podobě, prováděné za účelem získání informací, zahrnujících mimo jiné vzory, tendence a souvztažnosti; takto zhotovenou rozmnoženinu je povinna uložit s vhodnou úrovni zabezpečení a může ji uchovávat pro účely vědeckého výzkumu, včetně ověření výsledků výzkumu.
§39c Licences to reproduce a work for the purpose of automated text or data analysis
(1) A person who makes a reproduction of a work for the purpose of automated analysis of texts or data in digital form, carried out for the purpose of obtaining information including, among other things, patterns, tendencies and correlations, shall not interfere with copyright; he shall be entitled to keep the reproduction thus made only for the time necessary for the purpose of such automated analysis of texts or data.
(2) The provisions of paragraph 1 shall not apply to copies of a work the author of which has expressly reserved use pursuant to paragraph 1 in an appropriate manner; in the case of a work made available pursuant to § 18 par. 2, by machine-readable means.
(3) The provisions of subsections 1 and 2 are without prejudice to the provisions of Section 39d.
§39d Licence to reproduce a work for the purpose of automated analysis of texts or data for scientific research
It does not interfere with copyright law when
(a) a university which carries out scientific research as part of its activity, or a legal person whose main objective is to carry out scientific research or to carry out an educational activity involving also scientific research, if the scientific research of that university or legal person is carried out in such a way that access to its results is not preferentially afforded to the person exercising decisive influence over that university or legal person and at the same time that the research is carried out in the public interest or on a non-profit basis, or in such a way that all profits are re-invested in the scientific research of that university or legal person, or
(b) a cultural heritage institution,
when, for the purposes of scientific research, it makes a copy of a work for the purpose of automated analysis of texts or data in digital form, carried out for the purpose of obtaining information including, inter alia, patterns, tendencies and correlations; the copy thus made shall be stored with an appropriate level of security and may be kept for the purposes of scientific research, including the verification of research results.
What will be solved (hopefully)

In research:
- the problem of getting licenses (permissions) from 1,000s of authors of large collections, such as those on the web
- use of work where authors are unknown or hard to find without working hard to find them
- use of modern language corpora in research (not only 70 years after authors’ demise)
- use of continuous texts (not shuffled or otherwise pseudo-non-copyrighted modifications) - very important

In industry:
- (at least theoretical) possibility to use data for training their models (in case they acquire them themselves)
What will not be solved (unfortunately)

In research:
- reproducibility (if we understand the wording “for verification purposes” as **NOT** implying sharing(?)
  - !! I would argue that “**verification**” implies that it must be done by **someone else**, not the one who collected the data, otherwise it cannot called “verification”
- modifiability (e.g., are we entitled to add annotation as part of the “text and data mining” process)?
  - I would argue yes (it is “**automated analysis**”)
  - but what about manual annotation (to train models)?
- distribution of (trained) models created based on the TDM process on the copyrighted data:
  - are models covered by copyright?
    - I don’t know the definite answer (is there one?)
  - or are models just digital objects, not copyrightable and thus free of the original restrictions
    - who is then the owner of these digital objects?
    - who owns a text (or any output), processed by the trained models?
What will not be solved (unfortunately)

In industry:

- Model ownership: see the previous slide
- Sharing of the texts (non-research institutions cannot even keep the data, not even for (internal) verification)
  - Not that companies share their data much, but some are willing - they can’t, even with the CD in force
- Current licences
  - if something is shared under a known license (such a CC), is it overridden by the CD?
    - i.e., for CC-...-NC, can it be used by the industry without asking the copyright holder?
- Technology transfer between research and industry
  - Data (collected under the TDM by industry) cannot move to research organizations - impractical
What would be needed from the Language Technology research and technology transfer point of view

Clarification (or a precedent… anyone :)?)

- model ownership and “definition”
- what overrides what and what cannot be overridden
- technical means of protection and their workarounds

Addition

- possibility to share texts (w/annotation) (special license?)
  - absolutely crucial for reproducibility
    - no reproducibility -> no publication!
  - and for industry, too (technology transfer, contractual development, ...); AI Regulation - must keep data!!?
  - credits (research, industry)

EU/non-EU compatibility

- transposition compatibility
Questions?

legal@lists.clarin.eu
Discussion
https://consent.dariah.eu/

Call for translations: https://eldah.hypotheses.org/425
Getting involved in CLARIN

• Join our NewsFlash
  – https://www.clarin.eu/content/newsflash

• Check out our events
  – https://www.clarin.eu/events

• Open calls
  – https://www.clarin.eu/content/funding-opportunities
  – New call https://www.clarin.eu/content/seed-grants-support-horizon-europe-proposal-preparation

• Follow us on Twitter @CLARINERIC

• And stay tuned for the next cafés
  – https://www.clarin.eu/content/clarin-cafe
  – #clarincafe
See you at the Next Café

December 1st 2-4 pm CET

Towards guidelines for integrating CLARIN into teaching. Lessons learnt from UPSKILLS

By Iulianna van der Lek and Darja Fišer

...registration page soon online!