Legal Issues in Digital Humanities: Analysis of Recent Advocacy and Continuing and Emerging Issues

Ketzan, Erik
ketzane@tcd.ie
Trinity College Dublin

Nayyer, Kim
kpn32@cornell.edu
Cornell Law School

Dombrowski, Quinn
qad@stanford.edu
Stanford University

Tilton, Lauren
ltilton@richmond.edu
University of Richmond

de Smedt, Koenraad
desmedt@uib.no
University of Bergen

Kamocki, Paweł
kamocki@ids-mannheim.de
Leibniz Institute for the German Language

Trollip, Benito
Benito.Trollip@nwu.ac.za
South African Centre for Digital Language Resources

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities

Introduction

Erik Ketzan, Trinity College Dublin
Kim Nayyer, Cornell Law School

Legal issues, especially copyright and personal data protection, are an important consideration in many areas of digital humanities: project planning, data/corpus acquisition, how data may be lawfully processed, and how research outputs may be distributed. With a global perspective, this panel features speakers with expertise in legal issues in DH, to discuss longstanding legal issues which can aid or restrict DH research and update the DH community with analysis of new legal developments.

Changes to law can come from law-making bodies informed by advocacy and official reports. Quinn Dombrowski and Lauren Tilton analyze the successful advocacy in the United States to allow the circumvention of technical protection measures for DH research. Koenraad de Smedt shares the outcomes of a committee review of the legal framework for sharing and re-use of research data in Norway.

Once changes to law are enacted, or judicial opinions delivered in common law jurisdictions, legal issues as they pertain to DH are often far from settled, but have only begun to be interpreted and applied. Paweł Kamocki discusses how questions remain about the interpretation and application of the 2019 Directive on Copyright in the Digital Single Market, as well as the continuing unclarity regarding circumvention of technical protections measures in the European Union. Kim Nayyer analyzes recent court cases and decisions from the Supreme Court of Canada and Supreme Court of the United States, and the state of restrictions on DH activities in these common law jurisdictions. Benito Trollip discusses issues in the interpretation and application of recent laws in South Africa in personal information, indigenous knowledge protection, and copyright, and how these affect DH research activities. Kiyonori Nagasaki will provide an analysis of the first few years after the implementation of copyright reforms pertaining to DH in Japan.

New DH research aims and methods inevitably raise new legal conflicts and solutions… or do they always? Ever-growing research in ML/AI models trained on the open web has generated considerable discourse on the legality and ethics of such data acquisition and processing, but Erik Ketzan analyses how many of the legal questions and proposed solutions around this cutting-edge technology are largely the same as those raised in earlier periods of the Web.

Format: 90 minutes total, 10 minute talk by each speaker/co-presentation, joint Q&A at the end.

Proceeding with Caveats: The Evolving US Legal Landscape for Text and Data Mining

Quinn Dombrowski, Stanford University
Lauren Tilton, University of Richmond

The legal landscape for text and data mining in the United States is evolving rapidly, following the granting of a 2021 exemption to the Digital Millennium Copyright Act (DMCA) that supports scholars who wish to computationally analyze DVD videos and/or ebooks. For the first time, it is legal to circumvent encryption on these materials in order to use them for computational research, and consult them to a limited extent to confirm computational findings. In this presentation, the co-authors of the ACH's letter of support for this DMCA exemption will share a brief overview of the current legal terrain in the United States for computational analysis of DVDs and ebooks, noting some of the significant constraints and limitations in the exemption as it was granted. They will share examples of pilot research projects that are testing the feasibility of the exemption in practice, and describe the plans coming into focus around simultaneously advocating for the renewal of the exemption after its initial 3-year period, and petitioning for a new exemption that better suits the needs of digital humanities scholars. Finally, with an eye towards fostering potential international collaboration, they will cover some of the most notable opportunities in the fine print of US law that differ from comparable language in EU.

How should we share research data?

Koenraad de Smedt, University of Bergen

The title is from an eponymous Norwegian report on licensing and making research data available. To answer this question, important distinctions need to be made, such as raw vs. processed data, source data vs. generated data, personal data and data about the world at large. Furthermore, making data open and making them FAIR-compliant are not the same. Norway, like most European countries, is affected significantly by EU regulations, such as GDPR, but also the Open Data Directive, the Database directive and the Digital Single Market directive, which to some extent attempt to define what research data is and how national policies should shape conditions for sharing. Both on the European and the national levels, there are clear socio-political expectations for research data to be shared as much as possible, but these expectations are not reflected in any coherent legislation regulating the rights to, or sharing and re-use of, such data.

The issues are particularly challenging for many humanities researchers whose objects of research are published works, enriched with various forms of analysis (such as linguistic corpora of literary works). Other humanities researchers face restrictions in working with personal data (as in sign language and spontaneous language production). Yet another group of humanities researchers operate on the fuzzy border between research and artistic expression (such as art exhibits and performances), with potentially different rightsholders in different realms.

Thus it is difficult for individual researchers, research communities and institutions to comply with political and research ethics expectations for research data to be made available while navigating a fragmented and complex legal landscape. Non-legal barriers to sharing are also identified. Licenses which clarify rights, roles and restrictions on sharing are an important step forward. The Norwegian report proposes a number of concrete recommendations for governments, ministries and research institutions, and six license etiquette rules.

The Ongoing and Future Impact of the 2019 Copyright Directive on DH Research in the EU

Paweł Kamocki, Leibniz Institute for the German Language

This talk presents an analysis of the ongoing and future impact of the 2019 Directive on Copyright in the Digital Single Market on DH research in the European Union. The Directive was intended to be transposed in all EU Member States by June 2021, but not all States have met this deadline. Article 3, a copyright exception for text and data mining for scientific research purposes, provides positive steps forward for DH research, but questions about its interpretation and application remain, particularly regarding its "lawful access" requirement and application to technological protection measures (TPM, also known as Digital Rights Management). Article 14, a rule regarding how copies of works of visual arts whose term of protection have expired should remain in the public domain, is a direct policy response to a judicial decision in Germany, where the Federal Court of Justice ruled that photographs of public domain paintings are in principle protected by a related right. I discuss the impact of these Articles on DH research activities including corpus acquisition and data analysis.

Copyright Law Developments, Their Implications for Digital Knowledge Exploration, and Opportunities for Moving Forward

Kim Nayyer, Cornell Law School

A flurry of recent legal developments and emerging issues lay a stage for both profound, exciting possibilities, and sombre evaluation of present realities that require us to reckon with the constraints of digital knowledge capitalism. Often, research resources available to scholars are licensed and not owned, with the effect that licences, or contract law, govern many DH activities with information resources more than copyright law does. Digital exploration may be controlled or limited, or, with risk mitigation in mind, may be presumed to be so. I will present capsule reviews of three recent rulings of the Supreme Court of Canada, each of which offers some degree of vision and promise in respect of copyright and droit d'auteur scope, fair dealings, and technological neutrality, tempered with the inevitable caution necessitated by histories and immediate futures of uneven or imbalanced digital relationships. I will also present capsule reviews of United States court cases underway in respect to fair use, academic collections, and digital formats, and I will share my thoughts on their potential implications, both exciting and cautionary, for scholarly exploration and digital access to knowledge. This discussion will rest against the backdrop of the present relationships that create legal uncertainties about scholars' digital exploration activities, and about the implications of Berne-centred copyright's intersections with Indigenous knowledge and cultural expression protections. I will conclude with illustrations of some strategic and possibly liberating actions in the institutional, policy, and advocacy arenas.

The possibilities of safeguarding online datasets containing personal information: A South African perspective

Benito Trollip, South African Centre for Digital Language Resources

Recent South African legislation governing aspects that include personal information, indigenous knowledge protection, and copyright — namely the *Protection, Promotion, Development and Management of Indigenous Knowledge Act 6 of 2019* (IKA), the *Protection of Personal Information Act 4 of 2013* (POPIA) and the *Copyright Amendment Bill 2022* (CAB) — are going to or are already affecting areas of digital humanities research. With specific reference to POPIA, challenges regarding the generating of research outputs where respondents are involved, are apparent. The use and production of, for example recordings and questionnaires, are possible minefields for researchers in a context where identifiable information could be relevant for the interpretation of their data.

This talk analyses the possible cases of different types of datasets that could contain identifiable information and that have been submitted to an online repository. Implications of making these datasets available, with the relevant legal implications, will be fleshed out. The proposal of safeguards, before datasets that could contain identifiable information is uploaded to and published on an online repository, is discussed with the relevant legal principles in mind.

The situation of intellectual property rights for DH in Japan

Kiyonori Nagasaki, International Institute for Digital Humanities

Japan's Copyright Act was amended in 2018-2019, which introduced new exemptions for copyright protection including, perhaps most notably for DH, an exemption for "minor use", or actions that may cause only minor harm to copyright owners (a concept which may be compared with non-consumptive use). This has been interpreted to mean that even copyrighted works may be made public within the scope of such use and that scanning, OCR, text search, text analysis, and other common acts in DH processing are lawful as long as the content of the work is only processed on a computer. This presentation will report on the impact and present analysis of this amendment on DH research in Japan in recent years.

Training AI on the open web: Are the legal issues as novel as they appear?

Erik Ketzan, Trinity College Dublin

It is now over 20 years since the first Creative Commons licenses were released in December 2002, and for decades we have relied on these and other open licenses to establish the norms of what people may do with copyright-protected content on the open

web. These norms are now being questioned anew in the growing discourse around AI/ML models trained on web data. Image generators such as DALL-E, code assistants such as GitHub Copilot, and other tools and models trained on the open web are being challenged through lawsuits and other forms of objection by the creators of copyright-protected material, often because the outputs of these models contain material so similar to the training data that authors feel their copyrights are being violated. This talk analyses whether the fundamental legal concepts and proposed solutions to these conflicts are truly novel, or whether current debates around AI/ML models are essentially re-hashing discourses and legal solutions that were thought to be essentially settled. Once again, an obvious solution for authors may be a technical protection measure (TPM, a.k.a. digital rights management or DRM), even one as simple as a "do not train models with this content" tag, although how such a solution would conflict with percolating trends in global lawmaking to allow circumvention of TPMs for research remains an open question.