TEXT AND DATA MINING IN INTELLECTUAL PROPERTY LAW

TOWARDS AN AUTONOMOUS CLASSIFICATION OF COMPUTATIONAL LEGAL METHODS

CREATe Working Paper 2020/1

THOMAS MARGONI

CREATe

Text and Data Mining in Intellectual Property Law: Towards an Autonomous Classification of Computational Legal Methods [1]

Thomas Margoni[2]

I. Introduction

Text and Data Mining (TDM) can generally be defined as the "process of deriving high-quality information from text and data,"[3] and as a "tool for harnessing the power of structured and unstructured content and data, by analysing them at multiple levels and in several dimensions in order to discover hidden and new knowledge."[4] In other words, TDM refers to a set of automated analytical tools and methods that have the goal of extracting new, often hidden, knowledge from existing information, such as textual information (text mining) or structured and unstructured data (data mining), and on this basis annotate, index, classify and visualise such knowledge. All this, which is made possible by the fast advancements in computational power, internet speed, and data availability has the potential to constitute, if not a revolution in the scientific field, certainly a major advancement in the speed of scientific development as well as in its direction. In particular, the impact that TDM may have in the *direction* of scientific enquiry is invaluable. This is because by identifying the correlations and patterns that are often concealed to the eye of a human observer due to the amount, complexity, or variety of data surveyed, TDM allows for the discovery of concepts or the formulation of correlations that would have otherwise remained concealed or undiscovered. Considering this point of view, it can be effectively argued that TDM can create new knowledge from old data.

The first part of this chapter illustrates the state of the art in the still nascent field of TDM and related technologies applied to IP and legal research more generally. Furthermore, it formulates

[1] Forthcoming in Calboli I. & Montagnani L., *Handbook on Intellectual Property Research*, OUP, 2020. The author would like to thank the editors for the great initiative, support and feedback and the many reviewers who commented on early drafts of this paper. This research was supported by H2020 grant ReCreating Europe: Rethinking digital copyright law for a culturally diverse, accessible, creative Europe (grant no 870626).
[2] Senior Lecturer in Intellectual Property and Internet Law, School of Law – CREATe, University of Glasgow.
[3] *See Text Mining*, WIKIPEDIA, https://en.wikipedia.org/wiki/Text_mining (Last visited March 4, 2020).
[4] *See* the OpenMinTeD definition available at http://openminted.eu/about/. Other definitions which appear substantially similar can be found in e.g. Sec. 29A UK CDPA 1988 ("computational analysis of anything recorded in the work") or in Art. 2(2) Directive on Copyright in the Digital Single Market ("any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations").

some proposals of systematic classification in a field that suffers from a degree of terminological vagueness. In particular, this paper argues that TDM, together with other types of data-driven analytical tools, deserves its own autonomous methodological classification as 'computational legal methods.' The second part of the chapter offers concrete examples of the application of computational approaches in IP legal research. This is achieved by discussing a recent project on TDM, which required the development of different methods in order to address certain problems that emerged during the implementation phase. The discussion allows readers to take a detailed view of the technology involved, of the relevant skills that legal researchers necessitate, and of the obstacles that the application of TDM to IP research contributes to overcome. In particular, it demonstrates some of the advantages in terms of automation and predictive analysis that TDM enables. At the same time, the limited success of the experiment also shows that there are a number of training and skill-related issues that legal researchers and practitioners interested in the application of TDM and computational legal methods in the field of IP law should consider. Accordingly, the second argument advanced in this chapter is that law school programmes should include in their educational offer courses on computational legal methods.

II. State of the Art and Systematic Classification

The application of TDM and related techniques, such as Machine Learning (ML), Natural Language Processing (NLP), and text and data analytics (sometimes summarised under the broad term of Artificial Intelligence, or AI) to legal research is a recent methodological approach. The UK reference manuals in legal research methods often still refer to two broad categories: on the one hand, black-letter or doctrinal approaches, and on the other, socio-legal studies, a category that—depending on the authors and manuals—can include approaches as varied as comparative law, law and economics, historiographic analysis, critical studies and empirical methods, to name a few.[5] TDM, or all-encompassing terms such as legal informatics, computational law, or computational legal methods are not usually mentioned.[6] Law schools

[5] *See ex pluris*, Robert Cryer, et al., Research Methodologies in EU and International Law (2011); Dawn Watkins & Mandy Burton, Research Methods in Law (2nd ed. 2018); Mark Van Hoecke (ed.), Methodologies of Legal Research – Which Kind of Method for What Kind of Discipline? (2013); Michael Slater & Julie Mason, Writing Law Dissertations (2007).

[6] Terminology does not appear to be standardised. As a mere illustrative exercise, "computational legal studies" is employed by Danial Katz (*e.g.* https://www.computationallegalstudies.com) and by Ryan Whalen (http://www.lawtech.hk/the-emergence-of-computational-legal-studies-2018/); "computational law" by Genesereth Michael Computational Law: The Cop in the Backseat, White Paper, CodeX—The Stanford Center for Legal Informatics (2015), and by the project https://legalese.com. Legal informatics has been in use at least since 1997, see Erdelez Sanda, O'Hare Sheila, Legal Informatics: Application of Information Technology in Law, in Annual Review of Information Science and Technology (ARIST), 32, 367-402 (1997).

offering undergraduate or graduate courses on how computational tools and methods can be applied to legal research are still rare and are in most cases limited to centres characterised by a noticeable research and agenda setting approach.[7]

A. Background Context and Methods of Legal Enquiry

It is a well-known aspect of legal scholarship that the way in which the law is studied (and taught) varies—sometimes significantly—depending on the legal system and legal culture, on the law school or programme and within the same law school or programme on taught subjects. Such a methodological variety can be predicated on the personal (i.e., the teacher's) or institutional (i.e., the teachers') view of the law's nature or function. Nature and function may be based on the (sometimes only implicit) assumption that the law is a closed, coherent, logical and self-contained system, which possesses an autonomous meaning disconnected from moral values and sometimes also from social variables.[8] Therefore, based on this view, the law has to be studied, researched, and practised mainly focusing on how it is laid down in books, statutes, and court decisions.[9] Other times, such nature is based on the idea that the law is part of a broader and more complex reality and can be measured and evaluated through evidence and empirical methods as much as, or as close as possible to, other natural or social sciences.[10] Based on this point of view, law schools should train legal scientists who will learn to choose from a rich methodological toolbox and select the most appropriate technique for any given research question, problem, or experiment.[11]

There is no shared understanding of what Law is and there probably never will be. As a result, there can never be a unique understanding of how the Law should be studied, researched, or taught. Consequently, the aforementioned views—and a growing number of intermediate and new positions—should not surprise. After all, in legal research there still appears to be some discrepancy not only about what methodologies should be followed in relation to a specific

[7] *See for example* the MIT Computational Law course, https://mitmedialab.github.io/2018-MIT-IAP-ComputationalLaw/.

[8] Also referred to as "legal doctrine" in Paulina Westerman, *Open or Autonomous? The Debate on Legal Methodology as a Reflection of the Debate on Law*, *in* METHODOLOGIES OF LEGAL RESEARCH – WHICH KIND OF METHOD FOR WHAT KIND OF DISCIPLINE? (Mark Van Hoecke ed., 2013).

[9] This is the view traditionally attributed to legal positivism where the usual reference is Herbert L. Hart, THE CONCEPT OF LAW, first published in 1961 by Clarendon Law Series, but could also be linked back to natural law authors such as Grotius or Locke, *see* Westerman, *supra* note 6.

[10] Also referred to as "legal science" *in* Westerman, *supra* note 6.

[11] Legal realism is probably the approach that for the first time suggested that jurisprudence should follow natural sciences methods, e.g. formulation of hypothesis, experimentation, reproducibility of results, the use of empirical methods.

problem, but first and foremost on the very fact that a methodology is required or even possible in the legal field.[12]

That said, it is clear that the attention on 'methodology' in the legal domain is growing. Whereas it may only be partially explicit in how law is taught in law schools, it is becoming a central element in how legal research is conducted[13] and funded.[14] This is further corroborated by the fact that PhD programmes offered by the most competitive law schools in the UK teach first year students courses in research design and both the qualitative and quantitative methods applied to legal studies.[15]

However, even among those UK law schools or research centres that stand out for the importance that they attribute to legal methodologies, courses on TDM, ML, or AI and the law are not yet part of their PhD programmes.[16] The exception being when a specific experiment or research grant allows the involvement of students in research activities, such as the one discussed in the second part of this chapter.

B. From Legal Informatics to Computational Legal Methods

Whereas most legal research methods manuals do not discuss or even identify TDM as an autonomous legal method, at the experimental level, things look different. In fact, the use of computer software to analyse the law is not a novel approach. For instance, the field of legal informatics has promoted the use of software tools in the study, research and application of law for some decades. The reference work of Erdelez and O'Hare dates back to 1997[17] but older sources referring to the use of logic and automation in the legal field are plentiful.[18] Nevertheless,

---

[12] Rob Van Gestel et al., *Methodology in the New Legal World* (EUI, Working Papers 2012/2013).

[13] For an example of a project indexing and classifying legal literature on the basis of the methods employed, *see* CREATe's https://www.copyrightevidence.org/. The literature on specific legal research methods is blossoming, *see e.g.,* THE OXFORD HANDBOOK OF EMPIRICAL LEGAL RESEARCH (Peter Cane & Herbert Kritzer eds.,2010).

[14] Many research projects and grants schemes in the legal field are becoming more and more demanding in terms of legal methods/methodology; this can be viewed for instance in EU's H2020 actions.

[15] This is the case, for instance, of the University of Glasgow, where 1st year Law PhD students are required to take some or all of the mentioned courses.

[16] The University of Glasgow Law School's CREATe (the UK Copyright and Creative Economy Centre www.create.ac.uk) focuses on the use of empirical methods in the study of the law. PhD students in their first year are asked to take qualitative (and depending on their research also quantitative) research methods. Nevertheless, with the exception of the pilot project here described, machine learning techniques applied to the study of law are not yet covered.

[17] *See* Sanda Erdelez & Sheila O'Hare, *Legal Informatics: Application of Information Technology in Law*, 32 ANN. REV. OF INFO. SCI. AND TECH. (ARIST) 367, 367-402 (1997).

[18] *See ex pluris*, Lee Loevinger, *Jurimetrics – The Next Step Forward*, 33 MINN. L. REV. 455 (1948); Layman E. Allen, *Symbolic Logic: A Razor-Edged Tool for Drafting and Interpreting Legal Documents,* 66 YALE L.J. 833 (1957), available at http://digitalcommons.law.yale.edu/fss_papers/4519; Bruce Buchanan & Thomas Headrick, *Some Speculation About Artificial Intelligence and Legal Reasoning*, 23 STANFORD L. REV. 40

over the decades, the speed, availability and sophistication of software, hardware and data transfer have increased exponentially. As a result, early work in the field of legal informatics focused mainly on computer-assisted legal research, law office automation, and the so-called expert systems. Yet, even if equipped with these less refined tools, the results of early experiments combining law and automation have shown immense potential, such as the ability to 'predict' the outcome of selected US Supreme Court decisions with a higher accuracy than 'human experts.'[19]

However, it may be argued that it is only in more recent times that the use of digital technologies have shown a more structural impact on how legal research is conducted. The evolution of legal informatics into what this essay defines as 'computational legal methods,' or perhaps more accurately, the birth within the broader field of legal informatics of a distinguishable and autonomous sub-field called computational legal methods, is certainly in part a reflection in the development of the underlying technology. AI and, in particular, branches such as TDM, ML and NLP (i.e., data-driven and statistical AI[20] as opposed to knowledge-driven and symbolic AI[21]) offer a level of speed, complexity, and the possibility to analyse amounts of data that were simply not imaginable only a few years ago. Yet, an important shift in *method*, not just in speed, can likewise be identified. This should play, as this essay proposes, a defining role in this field's taxonomy.

Accordingly, the defining characteristic of computational legal methods as an autonomous methodological formulation that embraces the use of statistical AI applied to legal studies is identifiable in an *inductive* rather than in a *deductive* method. In other words, a legal researcher employing computational legal methods does not derive certain conclusions from a preconceived set of principles, categories, or rules. Instead, from large sets of data (often many datasets, including very large amounts of interconnected data—the so-called 'Big Data'), which often do not really possess much in common at first sight, it is possible—through machine observation—to identify rules, correlations, and patterns that may point towards new hypotheses, the validation of untested theories or the visualisation of results in a way that can

(1970). Of 1992 is the creation of the journal of Artificial Intelligence and Law, see *From the Editors*, ARTIFICIAL INTELLIGENCE AND LAW 1, 1 - 2, (1992).

[19] *See* Theodore Ruger, et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-making*, 104 COLUM. L. REV. 1150 (2004).

[20] This type of AI, largely based on machine learning, TDM and other statistical approaches fits the so called field of *statistical AI* or *artificial neural networks*, *see e.g.* https://en.wikipedia.org/wiki/Artificial_neural_network. For a discussion of knowledge-driven versus data-driven approaches, see Didier Dubois, et al., *Knowledge-Driven versus Data-Driven Logics*, 9(1) J. OF LOGIC LANGUAGE AND INFO. 65, 65-89 (2000).

[21] This type of AI systems are closer to the so called Good Old Fashioned AI (GOFAI), also known as symbolic AI. The standard reference here is John Haugeland, ARTIFICIAL INTELLIGENCE – THE VERY IDEA (1989).

offer novel insights.[22] This inductive and data-driven approach can be better pursued by recent implementations of AI, such as TDM, and ML algorithms, which employ highly advanced statistical analysis such as artificial neural networks models.

On the contrary, the development of AI systems based on knowledge-driven, symbolic, and semantic frameworks that allow a researcher to infer rules that apply in a specific system from a general set of values or beliefs, would not fit into the systematic classification of computational legal methods. In other words, a deductive approach that is often based on pre-identified sets of concepts, beliefs, symbols, or taxonomies is what characterises legal informatics approaches such as computational law,[23] whereas an inductive, statistical, and data-driven approach is what defines the connected but methodologically different approach of computational legal methods.

Classifying computational legal methods and legal informatics as two related but distinguishable and autonomous fields seems to be a proper solution, especially from a functional point of view. But what is the specific nature of the relationship? In other words, are they just two different methodological approaches applying automation to legal research? Or is the relationship hierarchical where computational legal methods are a *species* of the broader *genus* legal informatics? This second classification should be preferred. It regards legal informatics as the general term that identifies a set of methodological approaches where anything 'computer' is being used to do legal research, and within this genus, there are deductive methods using mostly symbolic and knowledge-driven AI (computational law) and inductive methods, such as TDM, using mostly statistical and data-driven AI (computational legal methods). From a historical point of view, this attributes due credit to a field—legal informatics—that has tackled the complex relationship between law and technology for at least five decades. From a functional point of view, it seems that legal informatics has, until very recently, mostly focused on deductive and knowledge-driven approaches, although it is arguable that this was largely due to the state of technology and thus should not constitute a classificatory obstacle. What should be clear, and

[22] *See e.g.* Nina Varsava, COMPUTATIONAL LEGAL STUDIES, DIGITAL HUMANITIES, AND TEXTUAL ANALYSIS (2018), *as well as* Palmer Palmer, et al., *Needles in a Haystack: Using Network Analysis to Identify Cases That Are Cited for General Principles of Law by the European Court of Human Rights*, *in* COMPUTATIONAL LEGAL STUDIES: THE PROMISE AND CHALLENGE OF DATA-DRIVEN LEGAL RESEARCH (Ryan Whalen, ed., forthcoming 2019), available at https://ssrn.com/abstract=3307084 and https://ssrn.com/abstract=3413518; David Law, *The Global Language of Human Rights: A Computational Linguistic Analysis*, 12 LAW & ETHICS OF HUM. RTS. 111, 111-150 (2018); Laura Pedraza-Farina & Ryan Whalen, *A Network Theory of Patentability*, U. CHI. L. REV. (forthcoming 2019) and available at https://ssrn.com/abstract=3347365.

[23] Computational law is the term employed by e.g. https://legalese.com, which clarifies that in the pursuit of their ambitious goal they engage in symbolic AI not in ML/NLP/TDM. For an overview of how the project intends to achieve the goal, and an extremely interesting list of related resources, *see* https://legalese.com/#ai-nlp.

hopefully convincingly argued, is that there is a new and autonomous method of legal enquiry characterised by an inductive, data-driven and statistical approach that uses tools such as TDM, ML, and neural networks, which this chapter proposes to call 'computational legal methods.'

C. Computational Legal Methods and the Question of 'What is Law'?

As mentioned above, there is a necessary logical correlation between the answer to the question of what is Law and legal methods. Under this point of view, it can be argued that some links between the type of AI methods employed and certain views on the nature, function, and especially methods of legal enquiry can be attempted. In fact, if it is correct to define computational law as a deductive or knowledge-based approach, which through a high level set of symbols or taxonomies attempts to develop a system able to analytically represent a certain legal concept, or more generally, any legal concept or legal area, including the law as a discipline, then perhaps it can be argued that computational law shares, if not the goal, certainly some of the methods of analytical philosophy and legal positivism.

On the other hand, if it is correct to say that computational legal methods are based on an inductive approach that employs TDM and other empirical, statistical, and data-driven methods to develop new and, to some extent, unanticipated knowledge from a variety of legal and non-legal sources, then this field is closer to the broad category of socio-legal studies. Depending on the specific experiment and on the type of algorithm employed, it should be possible to identify within this broad category more specific methodological approaches such as statistical, sociological, ethnographic, comparative or empirical legal methods.

D. Computational Legal Methods as an Autonomous Methodological Approach and the Future of Legal Research

In conclusion, the defining characteristic of computational legal methods as an autonomous field from closely related areas of legal research assisted by computers is a shift in methods and tools: from deductive to inductive, from symbolic and knowledge-driven to statistical and data-driven, and arguably also from positivist to socio-legal. TDM is the latest and most iconic technology currently representing this methodological shift.

However, in this new approach, there also appears to be a displacement in the centrality of the human element. Perhaps it would be excessive at this point in time to state that computational legal methods are those methods where the human and the AI share an, if not equal, at least a comparable amount of responsibility in the experiment design and execution. However, the role of the AI element is no longer that of a supporting or assistive technology, but instead has acquired a completely different centrality that heavily influences the outcome of the research.

Therefore, what this means for the future of legal research needs to be carefully assessed in the years to come, especially in relation to how the AI element is built and trained. The more centrality the AI element acquires, the more crucial a transparent, accountable, and ethically trained AI algorithm must be. This is possibly the single most relevant challenge that the field of computational legal methods will have to confront in the future.[24]

The next section will discuss in detail a concrete example of the application of the tools theoretically presented so far. This will be done by referencing a recently concluded EU H2020[25] funded project on TDM—OpenMinTeD.[26] In doing so, the examples of both kinds of approaches (e.g., deductive and knowledge-driven versus inductive and data-driven) will be presented.

III. Text and Data Mining and Intellectual Property Licences

OpenMinTeD's goal is to allow researchers, research institutions, and data providers to find, use, and combine resources for TDM purposes, thereby boosting science and innovation in the EU. Therefore, the main goal of OpenMinTeD is not to specifically TDM legal sources, but to mine all sorts of scientific knowledge. Nevertheless, in its initial three-year period, the project had a legal working group tasked with addressing some of the legal issues connected with TDM tools and services.[27] As it will be described below, in order to address certain obstacles, the advantages and limits of TDM legal tools were explored.

A. The Identification of the Problem

Initially, the project had to determine whether and how 'resources' (i.e., often literary works sometimes arranged in protected databases) and 'components' (i.e., mostly software or online services/workflows) necessary for the project's TDM objectives were protected by copyright and related rights, and to what extent specific exceptions and limitations applied in the specific case of TDM.[28]

---

[24] For a discussion of the dangers of a theory-less knowledge society *see* Jonathan Zittrain, *Intellectual Debt: With Great Power Comes Great Ignorance*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y. AT HARVARD UNIV. (July 24, 2019), https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c.

[25] EUROPEAN COMMISSION, https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020 (last visited March 4, 2020).

[26] OPEN MINTED, http://openminted.eu/about/overview/ (last visited March 4, 2020).

[27] The author of this chapter was the lead of the legal working group.

[28] The legal basis permitting reuse is not without consequences as it determines the conditions that must be met in order to benefit from it, being it a statutory exception or a contractual authorisation. This aspect, while analysed during the project, is not relevant in this paper. For a more detailed analysis of these aspects *see* Thomas Margoni, *Artificial Intelligence, Machine Learning and EU Copyright Law: Who "Owns" AI?* (CREATe working paper 2018/12), in XXVII AIDA 281, 281-304 (2019); Richard Castilho, et al., *A Legal Perspective on Training Models for Natural Language Processing, LREC2018* (conference paper), available

As a result of the often-unsatisfactory answer to this question, the second part of the project focused on the complementary aspect of licensing, and more specifically licence compatibility across resources and components. Accordingly, the project developed a set of compatibility tools intended to help researchers navigate through licences and licence compatibility.

This chapter's focus is on a specific problem encountered during this second phase, which relates to the issue of 'calculating' the compatibility of non-standardised legal documents that usually regulate the use of components offered as online services. As it will be explained, there were two possible ways to achieve this. On the basis of the classification offered above, they could be summarised as a traditional and rather rudimentary computational law approach and a novel computational legal method approach based on TDM, respectively.

B. Licensing Issues in a Typical Text and Data Mining Workflow

In a typical TDM workflow, there are three main levels at which licence compatibility has to be verified: the content or data level, the tools level, and the service level. Content generally refers to the resources that will be mined. These often consist of literary works (text mining) and are commonly referred to as *corpora* especially in the field of NLP, but may also consist of other types of data such as sounds, images, databases, etc. (data mining), or a mix thereof. Tools are the instruments used to perform TDM activities; for example, computer programmes (software) implementing specific TDM or ML algorithms. The service level refers to the fact that in a growing number of situations, TDM is not performed locally by downloading a specific TDM software as it was perhaps common a few years ago. Nowadays, a normal approach would entail using online services (the almost ubiquitous 'in the cloud') that allow researchers to select or upload an arbitrary number of *corpora* and TDM them using the software of the service provider.

Accordingly, these three levels (content, tools, and service) may or may not come with a specific licence (in the first two cases) or terms of use (in the latter case). Normally, licences are a type of copyright licences that include rights related to copyright, most importantly the European *Sui Generis* Database Right (SGDR). The third category (services) is often characterised by the presence of Terms of Use or Terms of Service (ToS) that regulate the use of a specific service.

---

at: https://www.researchgate.net/publication/323668123_A_Legal_Perspective_on_Training_Models_for_Natural_Language_Processing; Rossana Ducato & Alain Strowel, *Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"*, 50 IIC 649 (2019); Thomas Margoni, & Martin Kretschmer, *Property Rights Over Ideas in an Information Society. Or on the Necessity (and absurdity) of a TDM Exception,* (EPIP conference paper 2018); Christophe Geiger, et al., *Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?*, 49 IIC 814, 814-844 (2018); Elena Rosati, *An EU Text and Data Mining Exception for the Few: Would it Make Sense?*, 13 J. INTELL. PROP. L. & PRAC. 429, 429-430 (2018); Andres Guadamuz & Diane Cabell, *Data Mining in UK Higher Education Institutions: Law and Policy*, 4 QUEEN MARY J. INTELL. PROP. 3, 3-29 (2014).

They may establish certain copyright conditions, but do not normally consist of a standard copyright licence.

Therefore, there are three levels and three corresponding types of legal documents which may or may not allow certain actions necessary in order to perform TDM analysis. For example, a researcher has collected N *corpora* (each with its own licence) and intends to TDM them using a specific software locally or as part of an online service. Whether they can do that and under which conditions is often (i.e., unless a specific exception or limitation applies, something that is not covered by this chapter[29]) a matter of what the corresponding licences establish.

In light of this 'multi-layer' compatibility issue, the project developed a static matrix, or ontology, which 'calculates' the compatibility degree of different licences within the same layer (that is to say among content licences, among tools licences, and among ToS). This matrix can be seen as an attempt to classify ex-ante the conceptual types needed to perform the legal analysis and therefore, in the light of the terminology proposed above, corresponds to a computational law approach. An example of the matrix is shown below.

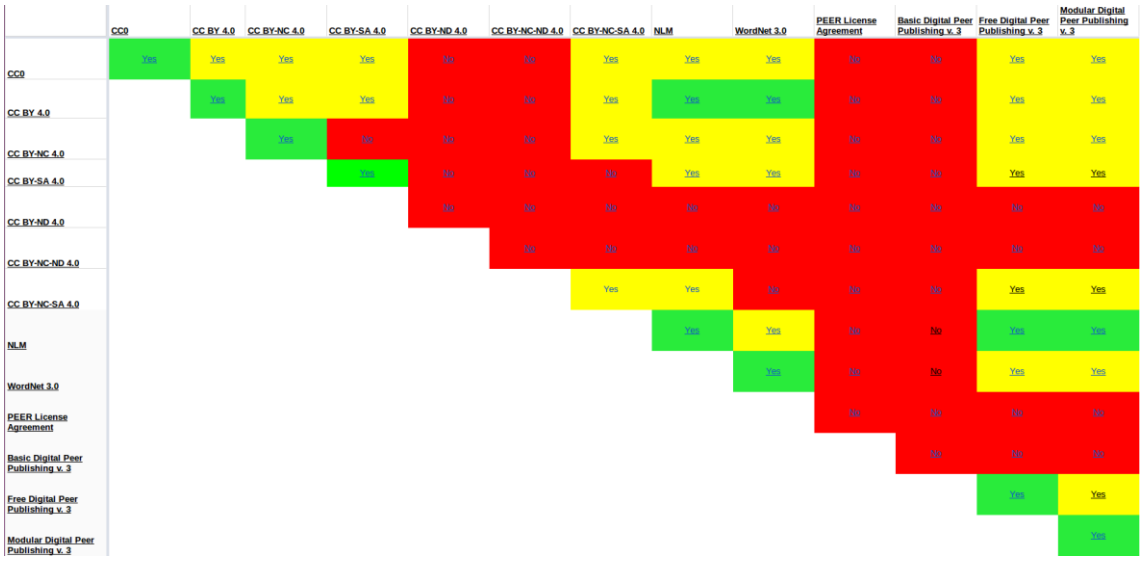


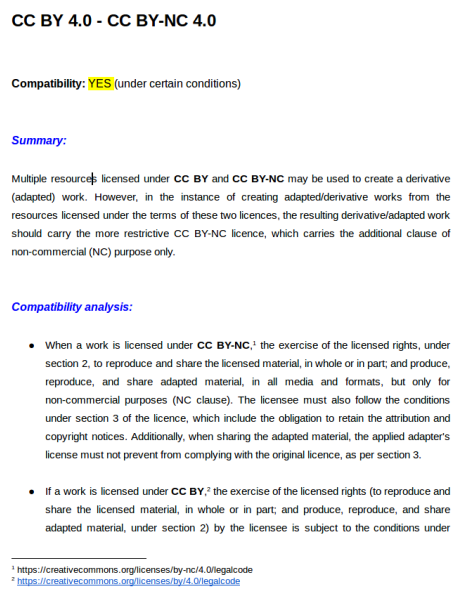
Figure 1. Compatibility Matrix, Sheet 1: Content Licenses

[29] For a detailed analysis of these issues *see* Thomas Margoni, *Artificial Intelligence, Machine Learning and EU Copyright Law: Who "Owns" AI?* (CREATe working paper 2018/12), *in* XXVII AIDA 281, 281-304 (2019).

1. *Pros of the Matrix: Compatibility Analysis*

The matrix clearly identifies three types of sources that are relevant during most TDM processes (content, tools, and services) and for each type, it collects a number of 'most commonly used' licences identified on the basis of the input of the project's partners. It can be observed that for content and tools, all the licences are Open Content or Free and Open Source Software (FLOSS). This is because on the one hand, they are the most popular due to their liberal conditions which favour an open and collaborative scientific environment, while on the other, they are the 'easiest' to analyse.

Ease of analysis is based on two characteristics: all of these licences are 'public licences', which means that they are standardised documents offered to the public usually as standard form contracts. Therefore, it is not necessary to develop a case-by-case analysis of the specific licence developed by vendor X or provider Y. A single licence or a few of them (e.g., CC BY or GNU GPL) normally cover a very high number of providers. On the other hand, this also means that the likelihood of incompatibility is reduced because a few standard licences have many more chances to be compatible among themselves than a higher number of custom licences. Nevertheless, even within the field of 'open licences', there are degrees of incompatibility or of conditional compatibility often caused by the so-called phenomenon of licence proliferation. The matrix makes an important contribution in this direction as it aims to calculate the compatibility degree of the identified licences offering a detailed legal analysis, which is publicly available and can be openly consulted and commented upon (See Figure 2 below).



CC BY 4.0 - CC BY-NC 4.0

Compatibility: YES (under certain conditions)

Summary:

Multiple resources licensed under **CC BY** and **CC BY-NC** may be used to create a derivative (adapted) work. However, in the instance of creating adapted/derivative works from the resources licensed under the terms of these two licences, the resulting derivative/adapted work should carry the more restrictive CC BY-NC licence, which carries the additional clause of non-commercial (NC) purpose only.

Compatibility analysis:

- When a work is licensed under **CC BY-NC**,[1] the exercise of the licensed rights, under section 2, to reproduce and share the licensed material, in whole or in part; and produce, reproduce, and share adapted material, in all media and formats, but only for non-commercial purposes (NC clause). The licensee must also follow the conditions under section 3 of the licence, which include the obligation to retain the attribution and copyright notices. Additionally, when sharing the adapted material, the applied adapter's license must not prevent from complying with the original licence, as per section 3.

- If a work is licensed under **CC BY**,[2] the exercise of the licensed rights (to reproduce and share the licensed material, in whole or in part; and produce, reproduce, and share adapted material, under section 2) by the licensee is subject to the conditions under

[1] https://creativecommons.org/licenses/by-nc/4.0/legalcode
[2] https://creativecommons.org/licenses/by/4.0/legalcode

Figure 2. Compatibility Matrix, Focus on the compatibility analysis

2. Limits of the Matrix: Non-standardised Terms of Use

The main function of the matrix is to provide a compatibility assessment between two licences by offering a concise answer: YES (green), NO (red) and YES with CONDITIONS (yellow). Each of these answers is a clickable link that redirects to a dedicated page with a detailed legal analysis of the compatibility of the two licences (see Figure 2 above). This is an important feature of the matrix. Whereas most researchers will be satisfied with a Yes/No/Conditional answer, it is important that the reasons for a certain compatibility result be publicly available so that they can be openly verified, allowing for correction or updates to possible (but hopefully unlikely) imprecisions.

The static matrix has been implemented in an online tool that automatically calculates compatibility. At the current stage, researchers have to manually select the corresponding licences from a drop-down menu, but in the future, the tool should be able to automatically retrieve the licences from the metadata of the content and services being used. A preview of the automatic tool is available in Figure 3 below.

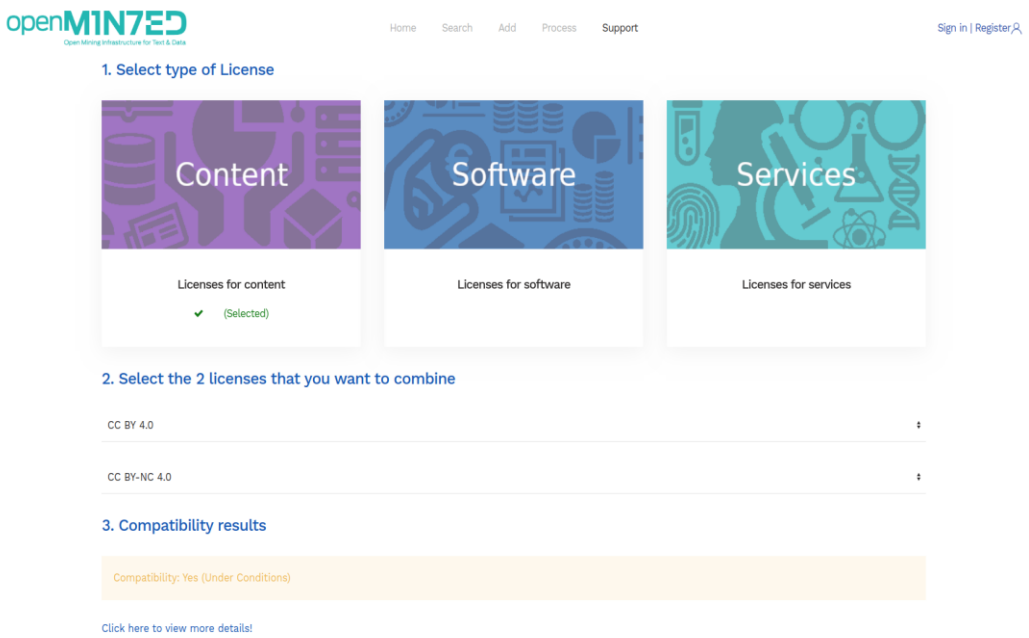


Figure 3. Pre-View of Automatic Licensing Tool

However, the static compatibility calculation is possible only because the selected licences possess two characteristics: they are publicly available and they are standardised (standard form contracts). These conditions are essential in identifying an authoritative public source and version of the licence, and also ensure that the licence is standard (i.e., it does not change depending on the specific user that uses it, and it does not change over time without proper documentation and versioning being provided). This makes it possible to calculate the compatibility of the licences, including versions and clauses, as documented in the matrix. In other words, these variables can be built into the framework design performing the legal analysis.

A problem emerges when the licences or legal documents are not public and/or standardised. This happens for some 'non-open' licences (which for the moment have not been implemented in the matrix) and further applies with much more incidence to the third category considered in the matrix, i.e., the ToS. When using a service, the service provider normally requires the acceptance of the ToS; however, these documents are virtually never standardised. Therefore, the calculation of their compatibility becomes more problematic due to the fact that it would not only be necessary to proceed with a case-by-case analysis of each ToS employed (something that would pose substantial problems with keeping the matrix up to date), but it would also be necessary to keep verifying every possible change that each ToS undergoes at any given time. Considering the fact that the ToS of the most common service providers change many times a year, this would drastically limit the matrix sustainability.

3. Possible Solutions

At this point, two approaches are possible: 1) Identify the most common ToS, deconstruct them, isolate a selection of standard clauses and calculate the compatibility 'statically', or 2) identify a number of clauses of particular interest (e.g., non-commercial, academic use only, mention of the sources, etc.), annotate known licences containing examples of those clauses and train a model on those annotations that will eventually be used by a software/service to identify similar statements on unknown documents (the unstandardised ToS).

Both of these approaches have been followed. Option 1 (currently implemented in the matrix) was chosen because it offered a low risk approach. It should be seen as a suboptimal approach, which is limited to the ToS that are covered (only the ToS that have been previously identified), but it will certainly offer an answer for those ToS. Option 2 presented a higher level of risk in relation to its feasibility due to the need to train legal experts in TDM and ML tools. Nevertheless, it was decided that a pilot experiment should be performed in order to test the feasibility of this type of approach, which has interesting potential because if properly performed, it can calculate

compatibility in virtually any unknown ToS. This second approach was based on a combination of linguistic annotation, text mining and model training and, following the proposed taxonomy, represents a computational legal method approach. The reminder of the chapter will focus on Option 2.

C. The Annotation Project

The lack of standardisation of ToS seems to be one of the main difficulties for interoperability purposes. To illustrate a possible problematic scenario, imagine that *corpora* required for an experiment are all under a liberal licence (e.g., CC BY 4.0), which in particular does not restrict commercial uses, but the mining activity is done using a service whose ToS states that it is only for 'academic' uses. What would be the status of the results obtained from the TDM analysis? How can one promptly calculate this when multiple *corpora* (i.e., multiple licences) and multiple services (i.e., multiple ToS), sometimes hundreds or thousands, are employed and use non standardised language?

In this complex scenario, text annotation, mining, and model training of licences and ToS may represent a valid instrument to detect, analyse, and interpret the linguistic terms by assigning a clearer legal meaning and identifying patterns that are likely to be present in unknown documents.

1. Legal Annotation for Text and Data Mining

The legal annotation of licences and ToS for TDM was done employing the WebAnno software tool (see below). In order to perform this task, it was essential to recruit and train annotators with legal and technical skills. A number of training sessions and tutorials were organised within the annotation project. The overall goal of this exercise was to provide annotators with the essential notions necessary to subsequently annotate legal documents.[30]

The purpose of the experiment was to study the linguistic variety used in legal documents to express licensing terms and conditions in order to be able to 'predict' terms in non-standard licences and ToS. This was done specifically by annotating phrases and expressions with a set of tags that denoted the most popular licensing terms (conditions of use) and annotating phrases and expressions used in licences for defining licensing concepts (e.g., attribution, non-commercial use, etc.). Two main types of 'annotation tag-sets' were identified: the tag-set for

[30] The annotation experiment was coordination by Dr. Giulia Dore (PhD, JD), with the invaluable TDM, ML and NLP support of OpenMinTeD partners Dr. Richard Eckart de Castilho (Technical Lead Ubiquitous Knowledge Processing (UKP) Lab, Computer Science Department, Technische Universität Darmstadt) and Penny Labropoulou (Senior Research Fellow, Institute for Language and Speech Processing, Athena Research Centre), who also prepared the background materials (guidelines, instructions) referred to in this chapter.

licensing terms, which consists of rules for actions (permissions, prohibitions, requirements, etc.) and the tag-set for definitions of concepts, which was open. In the process of phrase annotation, the whole phrase and not just the object of the action had to be annotated. For example, in the following excerpt from CC BY 4.0:

> Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to: reproduce and Share the Licensed Material, in whole or in part; produce, reproduce, and Share Adapted Material.

The entire underlined phrases had to be annotated, not just a single word (e.g., 'grants') in order to acquire a minimum of semantic meaning. The tags that can be associated with them are shown in the following table.

Phrase to be annotated (Statement)			**Tag**
Fragment1	**Fragment2**	**Fragment3**	
the Licensor hereby grants You	to reproduce	the Licensed Material	*permits_reproduction*
the Licensor hereby grants You	to Share	the Licensed Material	*permits_distribution*
the Licensor hereby grants You	to produce, reproduce and Share	Adapted Material	*permits_derivatives*

Table 1

In order to create an annotation, annotators had the following options. Under the first option, if they intended to annotate a *phrase* with one of the pre-defined values of the tag-set, they were asked to click on the value 'Statement' from the drop-down menu item 'Layer', then select the phrase that they intended to annotate and click on the desired value from the drop-down item 'Tag' (e.g., '*requires_copyleft', 'permits_derivatives',* etc.). Under the second option, if they intended to annotate a *discontinuous phrase* with one of the pre-defined values of the tag-set, annotators had to click once again on the value 'Statement' from the drop-down menu item

'Layer', then select the main part of the phrase that they wanted to annotate and click on the desired value from the drop-down item 'Tag' (e.g., '*requires_copyleft', 'permits_derivatives',* etc.). After that, in the 'Fragments' section on the right-hand side menu, five slots labelled as 'Frag1', 'Frag2', 'Frag3', 'Frag4', and 'Frag5' appeared. The annotators had to click on the slot next to 'Frag1' to activate it and then select the remaining part of the phrase in the text. In this way, the phrase was coloured and tagged with the '*statement_fragment'* label over it. This procedure needed to be replicated for all the segments of the phrase by opportunely selecting 'Frag2', 'Frag3', etc. for the remaining parts of the phrase. Finally, if the annotators had to annotate a *phrase* as *containing the definition of a concept*, they had to click from the drop-down menu item 'Layer' on the value 'Statement', then select the phrase that they wanted to annotate and click on the desired value from the drop-down item 'Tag' (e.g., '*defines_attribution', 'defines_non_commercial',* etc.).

A	B	C	D
tag name	**verb**	**act**	**tag description**
	defines	attribution	NEW:
	defines	commercialUse	NEW:
	defines	derivatives	NEW:
	defines	distribution	NEW:
	defines	nonCommercialUse	NEW:
	defines	nonDerivatives	NEW:
	defines	redistribution	NEW:
	defines	share	NEW:
permits_aggregate	permits	aggregate	The Assigner permits the Assignees to use the Asset or parts of it as part of a composite collection
permits_annotate	permits	annotate	The Assigner permits the Assignees to add explanatory notations/commentaries to the Asset without modifying the Asset in any other way
permits_anonymize	permits	anonymize	The Assigner permits the Assignees to anonymize all or parts of the Asset. For example, to remove identifying particulars for statistical or for
permits_archive	permits	archive	The Assigner permits the Assignees to store the Asset (in a non-transient form). Constraints may be used for temporal conditions
permits_commercialUse	permits	commercialUse	Exercising rights for commercial use is not allowed by the licence
permits_communicate	permits	communicate	The Assigner permits the Assignees to communicate the work to the public
permits_derivatives	permits	derivatives	The licence allows sharing of derivative works
permits_distribution	permits	distribution	The licence allows distribution, public display and public performance of the work
permits_education	permits	education	The work can only be used for educational purposes
permits_evaluation	permits	evaluation	Use is allowed for evaluation purposes (cf. ELRA_EVALUATION license)
permits_execute	permits	execute	The Assigner permits the Assignees to run the computer program Asset. For example, machine executable code or Java such as a game or a
permits_extract	permits	extract	The Assigner permits the Assignees to extract parts of the Asset and to use it as a new Asset. A new asset is created and may have very little
permits_incorporate	permits	incorporate	The Assigner permits the Assignees to incorporate the work unmodified into a Collective Work
permits_languageEngineeringRes	permits	languageEngineeringResearch	The work can only be used for research purposes in the Language Enginnering / Language Technology domain; subcase of permits_research
permits_modify	permits	modify	The Assigner permits the Assignees to update existing content of the Asset. A new asset is not created by this action. This action will modify
permits_redistribution	permits	redistribution	NEW: The licence permits the redistribution of the asset
permits_reproduction	permits	reproduction	The licence allows making multiple copies
permits_research	permits	research	The work can only be used for research purposes
permits_sharing	permits	sharing	The licence permits only non-commercial distribution
permits_sublicense	permits	sublicense	NEW: The Assigner permits the Assignee(s) to execute the act of granting a sublicense for using the Asset to a third-party
prohibits_addRestrictions	prohibits	addRestrictions	The assigner prohibits the assignee to impose any further restrictions on the recipients' exercise of the rights granted by the licence
prohibits_aggregate	prohibits	aggregate	The Assigner prohibits the Assignees to use the Asset or parts of it as part of a composite collection
prohibits_annotate	prohibits	annotate	The Assigner prohibits the Assignees to add explanatory notations/commentaries to the Asset without modifying the Asset in any other way
prohibits_anonymize	prohibits	anonymize	The Assigner prohibits the Assignees to anonymize all or parts of the Asset. For example, to remove identifying particulars for statistical or fo
prohibits_archive	prohibits	archive	The Assigner prohibits the Assignees to store the Asset (in a non-transient form). Constraints may be used for temporal conditions
prohibits_commercialUse	prohibits	commercialUse	Use is allowed only for non commercial purposes, such as research and education by academic users
prohibits_communicate	prohibits	communicate	The Assigner prohibits the Assignees to communicate the work to the public
prohibits_derivatives	prohibits	derivatives	Users are not allowed to share derivatives of the work
prohibits_distribution	prohibits	distribution	The licence does not allow distribution, public display and public performance of the work
prohibits_execute	prohibits	execute	The Assigner prohibits the Assignees to run the computer program Asset. For example, machine executable code or Java such as a game or
prohibits_extract	prohibits	extract	The Assigner prohibits the Assignees to extract parts of the Asset and to use it as a new Asset. A new asset is created and may have very litt
prohibits_highIncomeNationUse	prohibits	highIncomeNationUse	Use in a non-developing country is prohibited (from cc-rel)

Figure 4. List of Possible Annotation Tags Used in the Experiment

2. Annotators

The experiment involved ten annotators who all possessed an undergraduate law degree or higher legal qualification and were conducting post-graduate studies in the field of intellectual property. Each also received specific training in text annotation for TDM/ML purposes. While from the very initial stage of the design of the experiment it was clear that properly trained annotators would be instrumental in obtaining high-quality results, their numerical availability, their time availability, and the time necessary to further train them in TDM/ML annotation techniques conditioned the speed and the amount of data that this pilot experiment would cover.

Part of the reason can be certainly attributed to the fact that in order to obtain more consistent results, each licence was independently annotated by three different annotators (randomly selected) because different annotators annotate differently. Ideally, the more annotations there are for a single text, the better the annotation process because this gives more statistical relevance for the model building phase. A licence annotation by a properly trained and experienced annotator should take about one hour, but in fact it often took much longer. The licences were randomly allocated to annotators so that each licence would be annotated by a different mix of annotators. Each annotator was allocated seven licences, except for annotator A who had eight.

In order to obtain consistent results, the annotation phase was standardised as much as possible, which contributed to a more homogeneous performance of the annotators. Illustratively, before the actual annotation on WebAnno was conducted, the annotators were instructed to read the full text of the licence on paper with the tag-set at hand. This helped them get an overall view of the terms and conditions and enhanced their ability to understand the whole structure of the licence. Instructors requested each annotator to focus on: (a) one particular permission, (b) one particular requirement, and (c) one definition. Highlighting these first three elements on paper and choosing the applicable tag/s from the tag-set became handy when the annotation process moved to the screen.

3. The Annotation Software: WebAnno

The chosen annotation software was WebAnno.[31] WebAnno is a general-purpose web-based annotation tool for a wide range of linguistic annotations, including various layers of morphological, syntactical, and semantic annotations. Additionally, custom annotation layers

[31] WEB ANNO, https://webanno.github.io/webanno/ (last visited March 4, 2020).

can be defined, allowing WebAnno to be used for non-linguistic annotation tasks.[32] The choice of WebAnno was also natural because it was developed by one of the project partners who also provided specific training assistance. Nevertheless, WebAnno possesses other important characteristics. In particular, it is a fully web-based tool with a lower entry barrier than many other annotation tools and it is FLOSS under Apache licence 2.0.

D. Results

The experiment proved harder than expected in a few areas. On the one hand, the number of skilled annotators is a first parameter to consider. In order to annotate legal texts, annotators should have a legal background and should receive specific training in the field of copyright licences, TDM, NLP, and annotation procedures. This is a time-consuming activity and requires access to a large and motivated group. As pointed out above, the more annotations per licence, the more accurate the results. This means that licences should receive many more annotations than what was possible. In this experiment, only three annotations were performed per licence, which is quite low. Eight to ten annotations per licence (or even more) would offer more accurate results. However, that would exponentially increase the amount of time that annotators need to invest. The other element to consider is the amount of training material, i.e., of licences, that can be used for annotation purposes. There are only so many standard and public copyright licences, and a few of them (e.g., CC and GPL) have inspired similar licences, meaning that the expressions used in this field may point towards some standardisation. Whereas this is not an issue in itself, this could nonetheless lower the ability of the model to accurately identify the same tag-sets (e.g., non-commercial) in unseen and unstandardised texts (the customs ToS employed by services).

As at the time of writing this report, the accuracy of the model had not been fully tested, and early results appear to point in the direction of a proper recognition of the tag-sets when the target text is not too different from the original one. The low statistical incidence reported above does not probably allow for the proper recognition of excessively different texts.

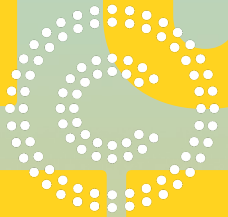
IV. Conclusion

In conclusion, it can be said that the potential of computational legal methods is very promising and only limited by the state of technology, or more accurately, by how much technology legal scholars have learned to use. As it was shown in the discussion of the experiment, a highly interdisciplinary team is a key element. In particular, it is not only necessary that the team be

[32] Richard Eckart de Castilho, et al., A WEB-BASED TOOL FOR THE INTEGRATED ANNOTATION OF SEMANTIC AND SYNTACTIC STRUCTURES (2016), in Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan.

made up of researchers with different scientific backgrounds, but in addition, legal researchers must possess a minimum of interdisciplinary methodological background in order to be able to face the challenges of applying TDM, ML, and statistical methods to legal analysis. This gap in legal research methods and training, and in the full recognition, if not at the undergraduate level then at least at the PhD level, of computational legal methods such as TDM should be urgently filled given the immense opportunities that these approaches can offer to legal scholars and practitioners and to the future of legal research. The imminent risk is that if legal scholars do not fill the gap themselves, they will soon be replaced by their AI equivalent.

TEXT AND DATA MINING INTELLECTUAL PROPERTY AWARENESS



CREATe

UK Copyright and Creative Economy Centre
School of Law / University of Glasgow
10 The Square
Glasgow G12 8QQ
www.create.ac.uk

2020/1 DOI: 10.5281/zenodo3752685

In collaboration with:

University of Glasgow

ReCreating Europe