Nobody puts research in a cage.

Researchers’ perspectives on working with copyright.

2nd edition

Introduction

Access to Knowledge is key to the fundamental **Right to Research**. Resources used in the context of scientific research are often protected by copyright and related rights, and rights holders can prohibit their use for research purposes. Researchers rely on copyright exceptions and limitations to **access, use and reuse** protected data sources in scientific projects. A fair and modern copyright framework is therefore essential to create an enabling environment for scientific research.

In some countries, researchers benefit from broad and flexible copyright exceptions and limitations that allow them to use protected materials in their projects, while in others they face overly restrictive laws that force them to either refrain from using such materials or to work in legal grey zones.

In the European Union (EU), a recent reform attempted to address some of the obstacles copyright law poses to scientific research. Text and data mining (TDM) – a modern technology where researchers use computational methods to analyse mass amounts of text, images and other data sources – is now allowed across all the EU.

While the EU-wide copyright exception for TDM represents a significant improvement to the legal framework for research in the region, it does not respond to all the pressing needs of researchers and their audiences. The mandatory exception only covers the rights of reproduction. It does not cover the right of communication to the public, which is essential to enable researchers, on the one hand, to get **remote access** to research resources and, on the other, to share the research results and underlying resources for purposes of verification, validation and dissemination of results. This poses problems from the perspective of **research transparency,** prevents researchers from complying with **open access** requirements for scientific research and hinders **joint and cross-border** initiatives.

This publication intends to demonstrate some of the problems that EU researchers face due to these constraints. It assembles an initial selection of views of individual researchers that were gathered through a series of interviews conducted throughout 2023 as part of the Right to Research in International Copyright Law project. The aim of the initiative is to better understand the needs and challenges faced by European and non-European researchers, particularly those interested in joint research activities and cross-border research collaborations. The interviews discuss issues of access, use and reuse of knowledge in unilateral and multilateral projects and the interviewees' perception of the current limitations related to copyright having an impact on their work.

The presented views do not constitute a representative research sample. Nevertheless, they serve as strong evidence supporting the need for a **global copyright reform** in the field of research addressing the essentials outlined by the researchers.

Interviews

Introducing the researchers

Anamaria Dutceac Segesten

I am a Senior Lecturer in European Studies at the Center for Language and Literature at Lund University (Lund, Sweden). I am also a reader at the department of strategic communication at the same university. Right now, I am doing research on social media and politics, or on a more abstract level about the intersection between technology and democracy.

Angelo Mario Del Grosso

I am a researcher (level III) at CNR-ILC (Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale) (Pisa, Italy). My research interests span from designing and implementing applications for Digital Humanities to build digital textual corpora for philological studies.

Deborah Grabc

I am a librarian at the Università Cattolica del Sacro Cuore (Milano, Italy). I hold a Ph.D. in Public Law from the Université Toulouse 1 Capitole, France. I work as a librarian and accessorily I do my research activity. My current main area of research is Text and Data Mining Literacy for librarians, teachers, and students.

Dominique Santana

I am a postdoctoral researcher at the Luxembourg Center for Contemporary and Digital History, where I am currently leading a transmedia project on the history of Radio Luxembourg, the biggest commercial radio station. I also finished my PhD there last year, based on my transmedia documentary project – A Colônia Luxemburguesa – on the history of steel-framed migration trajectories of Luxembourgers and other Europeans to Brazil.

Jonas Ingvarsson

I am a Senior Lecturer in Comparative Literature at the University of Gothenburg (Gothenburg, Sweden). I am also responsible for the master's program in digital humanities at the University of Gothenburg. I did my dissertation on literature and technology / cybernetics in the Swedish 1960s and have since been working on media history and literature.

Maciej Maryl

I am the Director of the Digital Humanities Center at the Institute of Literary Research of the Polish Academy of Sciences (Warsaw, Poland). As a researcher, I am involved in analysing bibliographic data, data-driven sociology of literature and research infrastructures.

Måns Magnusson

I am an Assistant Professor in Statistics at Uppsala University (Uppsala, Sweden). I am active in three to four different research fields, where we try to draw statistical conclusions from large amounts of textual data. I do this in collaboration with researchers in history, political science, law, sociology, and so on.

Marcin Jacoby

I am an associate professor at SWPS University in Poland. I am a Sinologist, which means a person who does research on China. And my main area of research is literary studies. What I do is research on ancient Chinese literature, but there is also another leg to my research, which is contemporary Chinese culture.

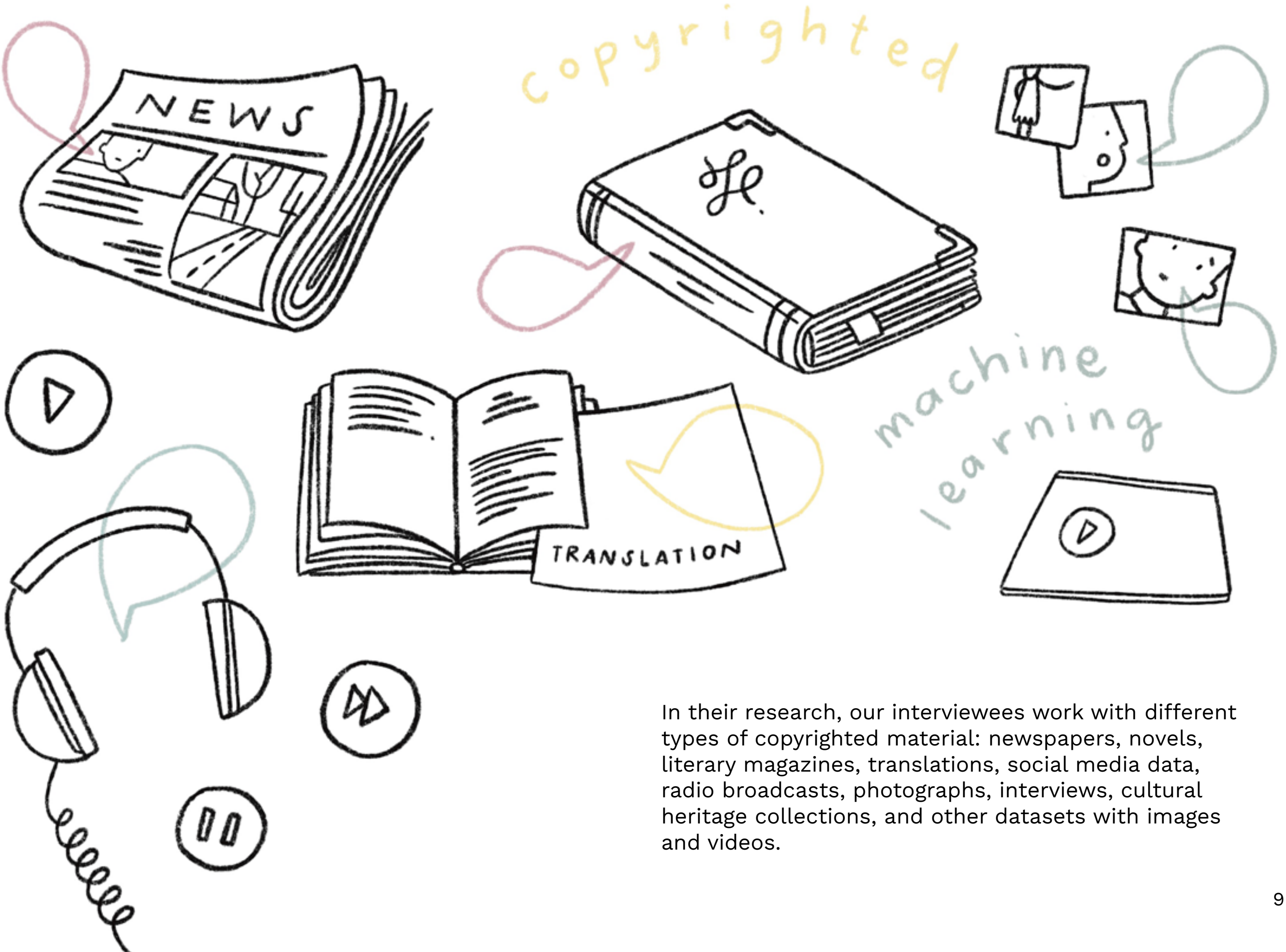
Mitar Milutinović

I am a researcher at Layer8 Institute (Slovenia). I was an AI researcher at UC Berkeley. My background is in computer science and my main current area of research is artificial intelligence, more precisely automated machine learning.

Rasa Bočytė

My official title is Senior Advisor for Research Collaborations at the Netherlands Institute for Sound and Vision, but within that role I take on different hats. One of them is being a researcher on a number of European projects that we work on. I did my bachelor's in History of Art at the University of Manchester in the UK and then I moved to the Netherlands and here I did an information and archival sciences master at the University of Amsterdam.

Which materials do they use in their research?



In their research, our interviewees work with different types of copyrighted material: newspapers, novels, literary magazines, translations, social media data, radio broadcasts, photographs, interviews, cultural heritage collections, and other datasets with images and videos.

How are they using those materials?

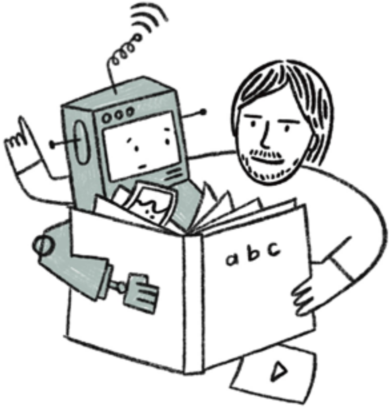
Måns: We research public discourse and analyse daily newspapers, radio broadcasts, etc. We really get into copyright issues and there it has been very messy. Much of it is based on the kind of research that is becoming commonplace right now in Europe and the US, ie. text as data – where text is treated as data for scientific research.

Anamaria: I am investigating the media discourse around refugee flows over a five-year period, to examine the evolution of the migration discourse over time (including the Syrian refugee crisis).

Maciej: In one of my projects, we have created a corpus of literary discourse for the past 200 years in Poland. We are trying to get copyright permissions to get the data from publishers.

Jonas: We are studying book reviews in Swedish newspapers from 1906, 1956 and 2006. We want to train the computers to understand different expressions in their context. We also have a dream that feels more and more likely, insane at first but now maybe real? That is, to train a text corpus to identify what is a book review!

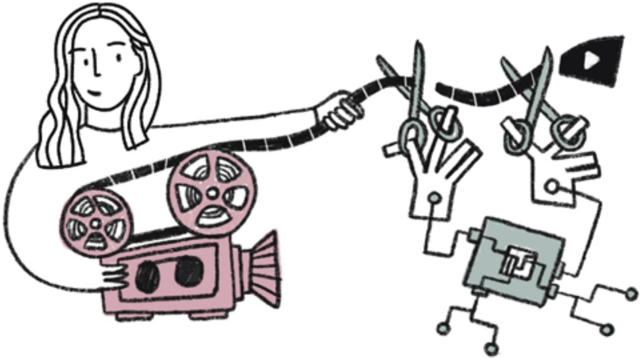
Angelo: I use copyrighted materials to extract info and recognize patterns and build indices to perform textual retrieval tasks. In my research project, we need to collect bilingual scientific publications (papers, monographs, metadata) in order to create bilingual corpora for machine translation training, terminology extraction and translation memory creation.



Mitar: We have made AI models which were trained on code and data to be able to process that data automatically.

Marcin: I do research on very old Chinese texts from 3rd century BC – this is what I specialise in. And since November 2022 I am part of a Horizon Europe consortium, where I'm responsible for leading a team of six researchers doing research on contemporary culture in Mainland China. Among other things, we use films or theatre productions that we analyse as data.

Dominique: For my transmedia project we recorded oral history interviews and, very often, people came along with their treasure boxes full of photo albums, documents, passports, everything you can possibly imagine. We digitised all that material that was given to us, and a considerable part of it is also available online. I was very lucky to receive an incredibly complete family archive containing letters, which I transformed into a character in the film. I also spent months in the corporate archives of the steel industry. It's a very large scope of multimedia archives. And then film archives – that was also quite an adventure: to bring a large amount of material to Luxembourg to be digitised and preserved.



Rasa: The main project that I'm currently focusing on is called AI for Media. It's a large scale project, there are many many research partners focusing on different aspects of AI development, and they need, let's say, really scalable data sets. So, for instance, [one research partner] is developing video summarization technologies, where you can actually take a long video of – I don't know – 40 minutes for instance and make kind of a representative summary of that, and we are interested in using that in our collection because that can improve access to long-form content that we have in our [media] archive.

Deborah: I use copyrighted material in my research work, mainly books and journal articles, they can be both published on paper and online. I always cite authors in my writings, and I have access to those materials thanks to our library subscriptions and acquisitions. To study text analysis and do some applications to improve and test my knowledge on the matter, I also built an authors' corpus of text belonging to [an Italian philosopher].

What copyright-related challenges do they face?

Our interviewees encounter a variety of copyright-related issues in their research. Måns and Jonas are unable to remotely access the collection of the National Library of Sweden. Deborah explains how a restriction to web scrape the contents of the United Nations Digital Library made her quit a project. Geoblocking is an issue that affects Marcin's team of researchers. Rasa struggles with the fact that copyright and cultural heritage laws prevent the sharing of collections from cultural heritage institutions with researchers. Marcin, Dominique and Angelo discuss how burdensome and time-consuming is the rights clearance process. Anamaria is worried about the legal and contractual limitations that are imposed on the sharing of research resources. Mitar and Maciej have similar concerns.

Måns: Our big problem at the moment is that because this type of material that we research is protected by copyright, it can only be analysed at the National Library [in Stockholm]. The research community is really annoyed at this, it's difficult for everyone outside of Stockholm.

Many projects that I would like to do, I have not even considered, I have just assumed that it is not possible. If I can't analyse the Swedish daily press in Umeå, I can't do it in Finland.

If I want to research French material, I must collaborate with French people and assume that they are aware of their copyright. I would never have dared to do research on Le Monde or anything like that without collaborating with someone in France.

Jonas: To access material from 1956, we have to go to the National Library Lab in Stockholm. It is a small glass cage with three data terminals. You sit in the lab, annotate. Access to it costs SEK 70,000 the first year, and 35,000 in the following years. You are not allowed to take data in or out, all labs must be done in the cage.

The transparency is non-existent. If someone wants to verify the results, they also have to buy the licence for a lot of money. An incredible anxiety!

Deborah: In the Spring of 2021 I carried out a project aiming at retrieving some bibliographic data to reuse them in our library records. I recovered a collection of the United Nations official records (1969 –1994) that was no longer visible online to users because cataloguing was lacking. Subsequently, I used the service offered by the United Nations Digital Library to retrieve the data, but I found out that the retrieved data was not reusable directly for my purposes. This was because the service was designed to enable users to search and discover and not to reuse data for different purposes, as the one I intended to. In addition to this there was a formal interdiction to do "web scraping" on the contents of the United Nations Digital Library without previous authorization.

At that point, redoing the original cataloguing would have been too long. So, after some manual and time-consuming cataloguing attempts, and because of the lack of personal knowledge on how to automate some data processing, I decided to quit the project. Now that my text analysis knowledge has improved, I should retrieve the project, but the copyright issues still represent a possible obstacle.

Marcin: A considerable part of the work is thinking about what I can do and what I can't do, what is legal, what is illegal. I can give you an example. Many of the films that we need to watch in China are available on Chinese portals. But these films are not available in Europe. To watch these things from Europe, we would have to pretend that we are Chinese, we would need to use VPN and we would need to use Chinese accounts. But this is illegal in China and, according to European Horizon Europe grant rules, we can't do things that are illegal in the country we are doing research on, so we can't do it. My colleague, who was doing research on film, actually went to China to go to a film festival and to sit in his hotel and watch things that he can't watch legally from Europe. This was very costly and time consuming, but we had no other option to deal with this problem.

Rasa: In many cases people come to us expecting that, as an archive, we can just provide our data, our collections, to them, just send it to them as, you know, via V-transfer, and we cannot. Most of our content is very recent, because it's broadcasting material, so pretty much everything is protected by copyright. And they really do not have that understanding that we cannot share it with them, which often comes to a surprise to them because they're just used to working with data sets that they can get openly from YouTube and they're often quite disappointed that we cannot do that. So that's also often an interesting conversation, when you have to explain to them this: we would love to do this, we would love to share our collections with you, but we simply cannot.

If we wanted to share thousands of videos, that would not be possible. It's usually under a hundred items, which in the case of developing AI is quite problematic because you do need really scalable data sets.

For instance, if an organisation has developed an algorithm and they would like to retrain it for our purposes, for our use case, we need to provide them with enough data and we cannot do that. Or maybe the only thing that we can provide them with are materials that are currently in public domain or published under more open licences, which in most cases are much older materials, like from the 40s, 50s, 60s. Old television materials that are not exactly to the quality and relevance that is needed. So that really becomes an issue that we cannot really share, let's say, the most recent collections, the most recent data.

This definitely affects us and usually affects us at the early stage of collaboration, as in the selection of partners who we are able to work with. That often applies to other cultural heritage partners that we reach out to. Usually because of copyright restrictions, they may not have digitised data or may not be able to provide data for joint purposes.

So if we're building, let's say, a joint platform or we're planning to reuse data for whatever purposes, for an audience engagement campaign, we need to be sure that those partners are going to be able to provide data. It's quite unfortunate that, then, you usually go to the usual partners.

And we're quite lucky in the Netherlands. When you look at Italy and the copyright regulations related to cultural heritage access, where it's much more difficult, especially with the latest articles about the minimum fees that you have to assign even to public domain content, that becomes quite problematic. So it just really restricts our ability to collaborate with partners who might want to collaborate with us, but because of copyright regulations they're just not going to be able to fully engage in the activities that we would want to work with them on.

Angelo: In my experience, a lot of constraints have been raised from copyright laws both in using and in publishing digital resources. Mainly I have to deal with publishing companies and personal successors of cultural material.

The main copyright-related obstacles concern awareness about licence topics when new Digital Humanities projects start. Moreover, copyright owners are unaware of the constraints and consequences of the licences to which they have subscribed. In general, there is a lack of law awareness.

Marcin: We have a big problem because in cultural research we can't really use the films or theatre productions that we analyse as data that we can freely give [access] to other people and [include in] data repositories. So what we can do, we are doing this now, we are creating a dataset, which is basically just a list of what we have been using, but we can't put anything more than this.

Part of the research is on Chinese modern literature and we did ask some of the publishers for the right to use longer fragments of texts, but it's very time consuming because if you are analysing, you know, 30 stories, then you have to reach to 30 publishers or authors and ask them for permission.
And the same with films. So it is possible, but you know it's too difficult for researchers to actually go through the process one by one with each and every partner, also remembering that many of them will not consider it their priority to answer your email, since they will not be earning any money and it's not anything that is attractive to them.

Dominique: My transmedia documentary tells the story of the massive movement of hundreds of Luxembourgish migrants to Brazil in order to erect a colossal steel plant and its surrounding industrial city, giving birth to the cradle of the Brazilian steel industry.

After going there on site and seeing and experiencing and feeling the very visible heritage of this migration history, I really felt this need to show it to the world and to make people feel it as well. This industrial town now has 85,000 inhabitants and only a very small minority knew about the history of that region, for a simple reason that they didn't have access to their history. So it was important for me to democratise this heritage, to make it accessible to everyone. And that's why I decided to put everything online, the film and all the interactive maps and everything [family archives, private archives, letters] are accessible to everyone worldwide. And yeah, so bringing the history to the public sphere, physically with Laço kiosks and online with this website, this platform, was very important to me also to strengthen these ties and for these two communities, both sides, to be able to be part of this heritage preservation work and to also continue this heritage work without me.

However, to have it online and in theory accessible to anyone the material, of course, we needed to clear the rights for every single, like, for the different elements on the website. And that takes a lot of, like, a crazy amount of time. Sometimes there were moments where I thought, okay, what have I done? Why did I have to invent this project?

Anamaria: I purchased data from Twitter, Reddit and Youtube. The agreements oblige me not to share data except within the project. The data must be on my server. Only researchers in the consortium have access to my data. We are not allowed to share the data for peer review, which is a huge problem.

Perhaps there are situations where, for example, Swedish researchers cannot share data from, for example, an experiment, but Norwegian researchers can, and then either the Swedish researchers break the law or they cannot publish in the journals that require open data.

In the future, we will see that social media (web data, chat apps, etc.) will be used less in research because there are so many requirements and it is so cumbersome. People will say it's not worth it.

Mitar: The main limitation we encountered was that we wanted to train our AI models on academic papers describing novel AI models (our project uses AI to build new AI models themselves, so training it on human AI models helps), but many papers are behind paywalls so we were not able to access them.

Another issue was collecting various datasets which exist on the web, that often lack clear information about licensing/copyright of the data so it is unclear if you can use them, even if you can access them.

And finally, after we have collected all this data, we wanted to make it public, but for many datasets we were unable to confirm that we have redistribution rights so we were unable to do so.

Maciej: The first obstacle would be access to texts and data. The second one would be the shady border between copyright and fair use, so having regulations that are understandable by non-legals. The third one is on how to licence your output and share it legally. When it comes to literary corpora, if you collect texts to do machine learning to support your research, it is really difficult to share them later or even to access them. The fourth obstacle would be the lack of proper legal guidance – even if you have an institutional lawyer and legal department, they are not always trained in this particular issue. And they are conservative by default as they are trying to protect the interest of the institution.

Are they involved in any cross-border research projects?

Marcin: I am part of a Horizon Europe consortium where I'm responsible for leading a team of 6 researchers doing research on contemporary culture in Mainland China. We are working with European partners, so we have partners from Germany, Belgium, France, Spain, Italy and Denmark.

As far as my ancient Chinese literature research is concerned, I have contacts with researchers in mainland China, in Hong Kong, in Taiwan (to a lesser extent). There's one researcher in Korea and Singapore, so basically some Asian countries, and mostly the US, US in my area of research is a very, very important partner.

Anamaria: Right now, I am involved in two Horizon 2020 projects. There are two parallel projects and in both I am the work package leader, which means that I am responsible for a segment in each project. I therefore coordinate research and implementation with many different partners. In one project I work a lot with civil society; we have 4 partners from civil society in 4 countries: Greece, Italy, Spain and Poland. Four cultural institutions are also involved, as well as three technological partners dealing with engineering in Greece, Italy and Spain. A university in Spain is also involved in the project. It's a very diverse group!

The second project where I am work package leader includes the universities of Ljubljana, Bergen and Stuttgart. There we collaborate on a larger scale with nineteen different universities, including in Canada and South Africa. That project only includes universities, which makes it easier. It is more difficult if you involve actors other than universities. Other types of organisations think differently, have a completely different everyday life. It is easier to predict what will happen if you work with universities.

Rasa: The main project that I'm currently focusing on is called AI for Media. I think it was still funded under the Horizon 2020 program. It's a large scale excellence centre that's been funded by the European Commission, one kind of the large scale excellence centres in AI focusing in different domains, and this one is specifically on the media sector. It's quite a large initiative. If I remember correctly, it's around 30 partners, there are at least 15 universities or research centres.

So for instance one of the partners that is coordinating the project and it's also our long-term collaborator is a research institute in Thessaloniki. There's kind of another pocket of partners that come from industry, so for example broadcasters who are based in Germany or right in Italy. We also have some communication partners: a design company based in Portugal if I remember correctly. And also a few universities who bring in humanities expertise focusing on kind of the ethical and societal questions, and also policy, which is also a domain that we closely collaborate with them. So this includes the University of Amsterdam and also KU Leuven University in Belgium.

Do they value collaborating across borders?

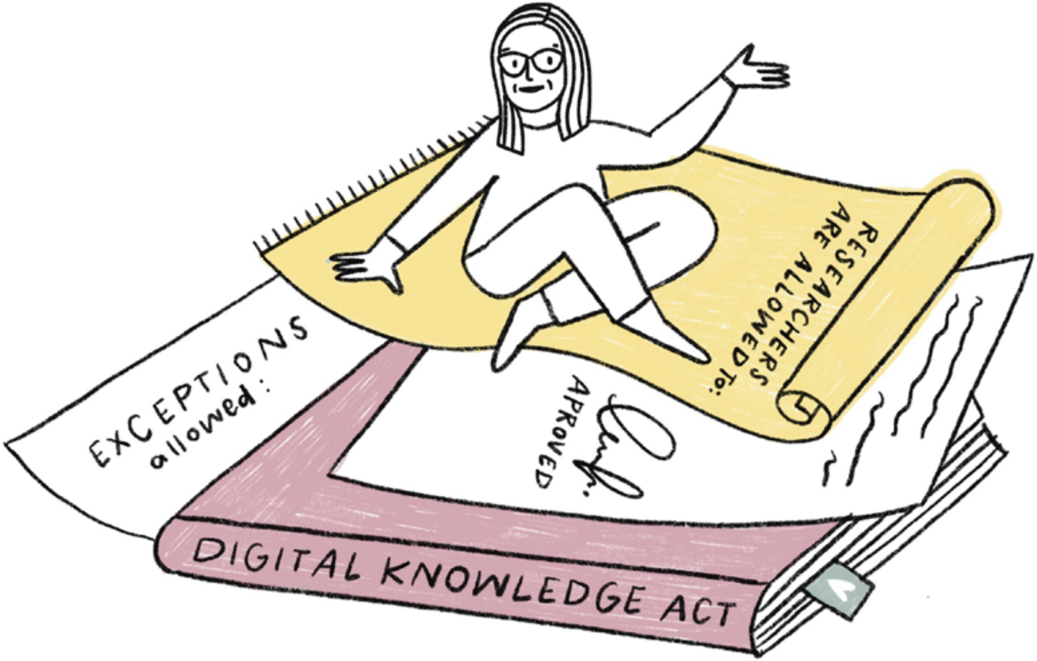
Angelo: The main value of supra-national joint research initiatives is exchanging knowledge, best practices, and opportunities to open new research questions from different perspectives and experiences. Cross-national networking is crucial for our research field. Colleagues and resources are spread around the world. European and pan-European research infrastructures are particularly important to guarantee sustainability of DH initiatives.

Anamaria: What gives me the greatest joy is being able to collaborate with people who are doing similar things elsewhere. In my department and at Lund University, there are not many others who do the same thing as me, or who can relate their research to my research – even though it is a large university. I imagine that the problem becomes even more urgent at smaller universities. Complementary to that is finding people who do other things that I can't do here locally. They can bring their perspectives and their methodological skills, creating a complementary project. I have many such collaborations, which result in publications but also in other initiatives supporting fellow researchers, students, new networks and so on.

Måns: I collaborate with researchers in many different countries, such as the UK, Germany and so on. Many of the scientific questions we have are not national. Such as in my research project: what effect do different events have on the migration discourse. If you can only analyse it in Sweden, you get a small piece of the puzzle, the more countries you can look at, the more puzzle pieces you can find, and thus add more puzzle pieces to the big knowledge puzzle.

Dominique: If you expand your research scope beyond national borders to analyse everything from a more transnational and fluid approach, it's always much more enriching for your research, especially if you're dealing with migration or transfer of knowledge, like any kind of transfer of people, like mobility, but also of skills and cultural heritage and so on. It's important to forget the national borders for a while. Well, consider them, of course, because you have to cross them, but not to limit yourself. So that's very enriching for this multi-perspective approach. And to understand the processes, like really the processes that happen organically. And enriching also not only for comparative research, I wouldn't say that I did comparative research, but really to nuance this history.

What would make their research work easier?



Rasa: I think for us specifically what would be beneficial is exceptions that would support research, education and creative reuse activities within these international collaborations especially when it comes to more recent data. In our case, let's say, the collections that we have in our archive are extremely valuable to researchers but they cannot get access to them because of copyright reasons. So having some sort of exceptions that we can rely on and more easily set up these kinds of collaboration agreements, saying that we're going to share this data specifically for the research purposes because this is going to benefit XYZ, would be extremely beneficial.

This is especially relevant in the area of AI development where you do need to have large access to data and high quality data. In most cases very strict copyright restrictions apply to high quality data. If we're talking about excellence in research, this is something that needs to be addressed, especially talking about the whole context of social injustices and biases that are really affecting the everyday citizens and causing harm to citizens and the distorted worldviews and patriarchal ideas that we enforce. It seems that this is quite a priority to make sure that the training of the AI systems should be supported by copyright regulations that actually allow us to make use of data that can improve those systems for the societal benefit.

Marcin: It would be great if I could use fragments of cultural productions in my research, for example giving readers open access to 10% of a film or a literary work, without needing to apply to copyright owners. That would be great. But, you know, I can't do it now.

Deborah: I recognize that sharing research data may not be sufficient for partners to verify results, protocols and techniques used in order to acquire them, and be able to reproduce them. Being able to share copyrighted resources, from which data have been extracted, for scientific purposes and between partners in research would be useful and would ease cooperation and truthfulness between researchers.

Anamaria: There must be a better balance, in legislation and structures, that makes exceptions and establishes means and methods for those who have legitimate purposes to study something that is problematic.

Recommendations

Protect the right to research at the national level

National policy makers should amend their copyright laws to permit research uses. Researchers should be allowed to conduct research on all kinds of copyrighted materials, and to access research resources remotely. They should be able to reproduce protected materials as well as to share them, in order to facilitate collaborative research and research transparency. It is equally important to ensure that research uses are protected from contractual and technological overrides.

Protect the right to research at the EU level

Text and data mining is already mandatory in the European Union. The EU now needs other research rights, to set up a level playing field for researchers in the region and offer legal certainty to cross-border initiatives. This requires, among other legal safeguards, an EU-wide mandatory research exception to copyright and other exclusive rights for scientific research. This exception should have a cross-border effect and permit different acts of reproduction and sharing of protected materials.

Protect the right to research at the international level

Policy makers should develop non-binding instruments on research uses, to support countries when reforming copyright laws to create an enabling environment for scientific research. In order to ensure that research is permitted in all countries and to foster research collaboration across borders, policy makers should also reach an international agreement on a set of minimum standards for research.

Editors:

Teresa Nobre, COMMUNIA
Maria Drabczyk, Centrum Cyfrowe
Eric Luth, Wikimedia Sverige
Deborah De Angelis, DDA Law Firm
Maja Bogataj Jančič, Intellectual Property Institute

Designer:

Magda Arażny, magdarysuje.pl

Acknowledgements:

We would like to express our gratitude to all the individuals who accepted our invitation to participate in the interviews that supported this publication.

COMMUNIA, Centrum Cyfrowe, Wikimedia Sverige and the Intellectual Property Institute are members of the Access to Knowledge Coalition – a group of civil society organisations from all around the globe that promotes access to, and use of, knowledge. Learn more about the Right to Research at **a2k-coalition.org.**

The publication is issued as part of the Right to Research in International Copyright Law project funded by Arcadia – a charitable fund of Lisbet Rausing and Peter Baldwin.

March 2024

The QR code will take you
to the online version
of the publication.