BEYOND the BOOK
A Copyright Clearance Center Podcast

Get Fit for Licensing: Healthy Metadata and the EU Copyright Directive

with

- **Roy Kaufman, Copyright Clearance Center**
- **Duncan Campbell, John Wiley & Sons**

For podcast release

Monday, March 30, 2020

KENNEALLY: Welcome to Copyright Clearance Center's podcast series. I'm Christopher Kenneally for Beyond the Book.

Earlier this month, London Book Fair organizers announced cancellation of the 2020 program schedule for March 10-12. The news was disappointing, of course, though not unexpected at a time when the world is confronting the pandemic spread of the COVID-19 corona virus.

At CCC, we recognize the difficulty in making the decision not to go ahead with this year's London Book Fair given global health concerns. We also believe in the strength of the content we had prepared to present as well as the importance of information sharing for the publishing community.

Throughout March, CCC is delivering a series of virtual programming planned for London Book Fair presentations. For a complete schedule, please visit copyright.com/lbf2020

This is a podcast edition for "Get Fit for Licensing: Healthy Metadata and the EU Copyright Directive," originally scheduled for the second day of London Book Fair 2020.

Over the next two years, EU member states are required to adopt The Directive on Copyright in the Digital Single Market, which passed in 2019. Importantly for scholarly publishers – whether based in the EU or not – this directive provides a clear and explicit formulation of the legal status of copying materials for text and data mining (TDM) and other types of information extraction.

While a narrow, non-commercial exception for scientific research does exist, the Directive leaves in place critical protections around licensing. To capitalize on any

opportunities, publishers must maintain clean, reliable metadata for their content, including about authors, institution, license types, and citations.

Roy Kaufman is Managing Director of both Business Development and Government Relations for Copyright Clearance Center. Prior to CCC, he served as Legal Director, John Wiley and Sons, Inc. Kaufman also advises the US Government on international trade matters through membership in the International Trade Advisory Committee (ITAC) 13 – Intellectual Property.

Duncan Campbell is Senior Director, Global Sales Partnerships at John Wiley & Sons, where he is responsible for licensing, agent relations and copyright & permissions for Wiley's academic journal and database content. In addition, he is also engaged in developing Wiley's strategies and policies in areas such as government affairs, content sharing/syndication and text & data mining.

+++++

KENNEALLY: Roy Kaufman is managing director of business development and government relations for Copyright Clearance Center. Roy Kaufman, welcome to the program.

KAUFMAN: Thanks for having me, Chris.

KENNEALLY: It's important to discuss these issues for scholarly publishers, because the directive provides a very explicit formulation of the legal status of copying materials for text and data mining. And I'd like to ask you, Roy, to explain what that new status is.

KAUFMAN: Sure, Chris. The Digital Single Market Directive actually answers a lot of questions about the copyright status of text and data mining, while, like any piece of legislation, also opening up a brand-new set of questions.

So when you think about what the DSM, which is what I'll call it for short, actually says, there's actually two main copyright exceptions. One that I'll call the scientific research exception really is about subscription content in scientific and academic journals. This is the peer-reviewed content. And the question that the EU was trying to wrestle with was do I create an exception for people who otherwise have lawfully acquired the material? So again, this is not like you can get a copy of the material from a pirate site and then text and data-mine it. It's very much if you have a subscription to this content, do you need the publisher's further consent for text and data mining?

For that content, what the EU decided was if it's sort of an academic, non-commercial use, you don't need further permission from the publisher, but if it's a commercial use, you do. This is very much in line with the position that the publishers had actually taken with respect to where they thought was a fair compromise in the DSM.

The other exception came a lot later in the game and was introduced – and I refer to it, and this is just my nomenclature, it's not really what it says – as the open web exception. Basically, what this is meant to deal with is if you find content on the open web and you want to mine it, you can, unless the rightsholder has expressly retained the rights to mine. If they've expressly retained those rights, it doesn't mean you can't mine it. It just means you have to go to the rightsholder and get permission. So it's up to the rightsholder whether they want to allow text-mining and under what terms – in other words, whether they want to issue you a license or not.

KENNEALLY: I appreciate that explanation, Roy. So the question then is an obvious one. How can publishers, how could rightsholders, reserve their rights?

KAUFMAN: It's kind of funny when you look at the way the law was written. For example, the directive seems to state that if you want to reserve your rights, at least for material online, you have to do it in machine-readable format. Why I think that's ironic and somewhat circular is text and data mining is essentially – and I'll say this in air-quotes – an exception to allow machine reading. So it has to be a pretty interesting machine that can read text but can't read a copyright notice that says I reserve my rights. Essentially the way I interpret this, since machines can read anything – they can read faces, they can read minds, they can read photographs – they can certainly read some text that says I retain my rights for text mining.

Now, do I expect people on the other side of this issue to accept that? No. But again, the EU could have said it has to be language put into robot text files. It could have said it needed to be machine-actionable. But it didn't. It said machine-readable. So notwithstanding that I think all content is essentially machine-readable if it's online, the advice I tend to give publishers is vote early, vote often, sort of. What you do is not only put it in human-readable so that humans can see it, human-readable is machine-readable, so a machine can read it, and also if you have robot text or crawler information or metadata, put it in your metadata, too. It doesn't cost anything extra to reserve your rights in many places.

KENNEALLY: I want to ask you about an additional wrinkle to all of this, which is the EU copyright directive was adopted – the discussions even begun much longer ago

than Brexit, but it was adopted the same year as Brexit, and we are now in the transition period for the UK as it leaves the European Union. I wonder if you can comment on how the UK situation is going to change, possibly complicate this.

KAUFMAN: Well, I mean, the UK will not pass the DSM directive, and that's the decision the UK has made. I've spoken to people within the UK government – not for attribution on this – but basically the position of the UK government is you have two years to implement the directive. We will be out. Therefore, we're not going to bother to implement the directive.

So there's a lot of people who, when they heard the UK has decided not to implement the directive are making policy conclusions – oh, the UK doesn't believe in this, or it doesn't believe in that. I think the UK just doesn't believe it has to pass any EU directives that are not due until it's left. So there's really not a big policy decision there.

KENNEALLY: Certainly we can rely on the experience of Copyright Clearance Center to draw some inferences about the future activity that publishers should undertake regarding these rights and the reservation of these rights. As a licensing organization, Copyright Clearance Center relies on metadata – that's the data about data, the information about a particular work. Roy, connect the dots for us. How does this opening, this opportunity, arise? How can they use metadata to ensure that those rights are reserved properly but that are also licensable?

KAUFMAN: Once you understand what metadata is, you understand that you can do almost nothing well if you don't have good metadata. Think of it at the highest level – if you were doing a one-to-one negotiation and you want to get, let's say, movie rights for a book, what's the metadata? The name of the book, the title of the book. Right? That's how you define things. That's a very, very simple example.

But then once you go online and you want to start licensing – reserving rights and licensing at scale, that's when metadata gets really important. Can you look at the name of an author and know who controls those rights? The answer is probably not. So you need further metadata. Well, who has the rights to republish this? Who has the rights in this country? Who has the rights to make translations?

So the metadata issues go well beyond the text and data mining exception, where it's probably less of a question than some of the other exceptions and obligations that have been added by the DSM that we're not really talking about today. For example, there's an entire section that's called the value gap which really says if you're a large platform – think someone like YouTube – and you want to use someone's content, and it's professional content, you need a license. How are you

going to find out who to get a license from? The answer is it's got to be in the metadata.

KENNEALLY: And this metadata will also help establish an important fact for the future, which is that a licensing market exists. Why is that important?

KAUFMAN: So the United States, we have a doctrine called fair use, and we also have what's called a common law legal system. A common law legal system is the courts look at other court cases to decide what the law means. Fair use, which is a fact-determinant way of looking at things – you look at a case and you look at what someone's done, and you look at, among other things, what's the market for this? Is there a licensing market that exists for this content? If there is a licensing market that exists for that content, that doesn't end the fair use analysis, but a court is much more likely to say if there's a license market and it's easy to get a license, and it's available, then you should get a license.

Now, that's not supposed to be how it works in a civil law country, which is less about how other courts have interpreted it and where copyright exceptions tend to be very, very specific and drafted in the law. But a similar thing happens. For example, one of the reasons we believe the DSM doesn't have a commercial text mining exception for scientific content is because the science publishers had a readily available market and were already actively licensing commercial text and data mining rights to corporate users. That's been done both by individual publishers and by CCC acting on behalf of collectives of publishers.

KENNEALLY: Roy Kaufman is managing director of business development and government relations for Copyright Clearance Center. Thanks for joining me on the program today, Roy, and for participating in CCC's virtual book fair.

KAUFMAN: It's my pleasure, Chris.

++++++++++++++++

KENNEALLY: Duncan Campbell is Senior Director, Global Sales Partnerships at John Wiley & Sons, where he is responsible for licensing agent relations and copyright and permissions for Wiley's academic journal and database content. In addition, Duncan Campbell is also engaged in developing Wiley's strategies and policies in areas such as government affairs, content-sharing syndication and text and data mining. Welcome to the program, Duncan.

CAMPBELL: Great. Thanks very much, Chris.

KENNEALLY: Tell us a bit about the work you do at Wiley, engaged in following the development of copyright legislation across Europe and across the world. In Brussels, in particular, over many years, there was an effort to prepare and eventually pass this so-called Copyright Directive, which did finally pass earlier last year and is now in the process of being adopted by all the member states. So that the audience can appreciate better the work that you do for Wiley on this particular topic, Duncan, tell us about your engagement with Brussels, with the EU and with other organizations such as publisher trade associations.

CAMPBELL: Well, my role at Wiley is really focused on our third-party business, so that's essentially, you know, digital licensing, copyright and permissions and agent relations. And it's really about working with third-party partners to drive Wiley's revenue and reach globally.

Now, as part of that, of course, one of the biggest pieces of my business is digital licensing. And I'm incredibly, very interested in how our content can be accessed and used by our customers and how we can extend – as I said, extend – our reach through various services. And a crucial piece of that, of course, is around how IP and copyright legislation globally both protects and enables licensing in the digital marketplace, so I'm very involved in developing our policies around – in areas such as content sharing and syndication – sort of the government affairs framework that that sits within – and also specific policies, such as text and data mining, which has had huge amounts of interest over the last few years. Though I will say that the kind of interest and focus from the legislative level isn't always matched by the actual interest and uptake from the user base.

In Wiley's case, we actually a government affairs team who are based in – we have people in the UK and in the US, primarily, and a couple of people in Asia who are involved specifically in managing our relationships with governments and with legislative bodies.

My role, really, is to act as sort of a subject matter expert, so, as I said, my areas to interest around content sharing and syndication, text and data mining, how copyright and IP is translated into the – or happens in the digital space. And so my role is to support our government affairs team when we need to reach out to policymakers and to legislators to help them better understand Wiley's position in relation to legislation that's being proposed. And, also, just to support our trade associations as well, in the work that they do so obviously being a scientific, technical, medical publisher, we do a huge amount of work with the STM association and with their government affairs team as well.

KENNEALLY: Well, an important issue, an important part, of the European Copyright Directive, Duncan, does address those areas you were just describing, particularly text and data mining. That raises the question of how exactly publishers are going to assert and reserve their rights. It's a real challenge.

CAMPBELL: Yeah, so I think the whole situation is quite quite interesting, because the exception for academic text and data mining is very similar to the one that was adopted in the UK a couple of years ago, which was after the Hargreaves Review of Intellectual Property. And essentially, that allows – I can't remember the exact text off the top of my head, but it allows computation analysis or allows copying of subscribed content or lawfully accessed content for the purpose of noncommercial scientific research. So essentially what the EU has implemented is very, very similar to the UK exception.

Neither copyright directives have a specific language around commercial licensing, so while the – and I'm sure Roy – Roy has explained this very clearly – while the EU Copyright Directive has a second clause that allows access and text and data mining of content that's available mainly on the open Web but is essentially not covered by the academic exception. But there isn't – in either legislation, there isn't an explicit permission for commercial text and data mining.

So for publishers such as Wiley, we find that there's obviously a huge amount of interest from large companies, so big tech, for example, such as Google, Microsoft, Amazon. Or sort of small and more focused artificial intelligence companies, who are wanting to license and access content either for themselves or, for example, on behalf of pharmaceutical companies they're working for, who are really interested in – so let's call it – electronic analysis of published literature to identify interesting patterns, new entities to try and essentially use that data to discover some new knowledge or potentially, in the case of pharmaceutical companies, interesting new objects for research and development.

And I think the key thing with this – the European directive, as you point out, Chris, is that it's asking publishers to make a very explicit reservation of rights to make clear that, if a piece of content is freely available either on the open Web or, let's say, I think, in terms of an article that is accessed off a publisher platform, how to make clear that that piece of content is not available for text and data mining if the publisher has explicitly said that it isn't.

And so this is one of the things we – it's clear from the language in the European legislation that the commission will not propose standards around this. We think that there have been some discussions, but the commission is more focused on

Article 17, which is around content sharing rather than on this exception, which is a relatively small one.

So the question for publishers is how do we make a declaration that we wish to reserve our rights in this case? Should that declaration be sort of horizontal, as in applying to everyone, or should it be more targeted vertically, so what's valuable in the – let's say – the sports arena is very different to what would be applied in the scientific, technical, medical publishing arena.

So the question is should it be horizontal or vertical? And also, do we need to have a reservation of rights at the article level, so it's machine readable if an article is found in the wild? Could that be at the platform level? So for example, already, there's a file called robots.text, which every Web page will have, which makes clear whether Google is allowed to crawl that page or not. Or Google and other search engines.

Should it be at that level or could it be reserved at another level, such as a subscription agreement between a publisher and a third party? So for example, in the case of a pharmaceutical company, Wiley would probably license quite a lot of content to that company. And it would be very simple, in that sense, to reserve our rights and to say that, you know, unless otherwise permitted, text and data mining of the content that the company was subscribing to could not take place.

KENNEALLY: What do you think is the best way for publishers to address all of this? Would you rather see an individual approach or an industry-wide tackling of this issue?

CAMPBELL: I think I'm always in favor of a broader approach at the industry level. I don't think one horizontal approach across all sectors, necessarily, will work. But I certainly think, at the level of, say, the STM community, I think that some form of standard would be best. I'm a great believer in sort of sort of development of common infrastructure and common standards to help us, as a community, deliver content to our users and make that content as useful and usable to them as possible.

I think, in this context, it would be good to have some form of understanding as to what a machine-readable declaration might need to look like. But realistically, given the timeframe of implementation and the fact that, depending on the current situation, of course, the aim is to have the copyright directive in the European Union transposed to the member states by the end of 2021. There may not be a huge amount of time to go through the work of actually developing a common standard.

I mean I think, to be frank, none of this is that difficult. It just actually needs to get done and get put into practice. And that may mean that individual publishers do decide to just do their own thing in order to be able to make it happen quickly. For example, in the case of Wiley, if we were to need to add a statement at the article level, we would need to probably reprocess seven or eight million content items, which is not a trivial task, but it's not huge.

KENNEALLY: Give us some vision into post-2020 for the UK and its relationship with the European Union as far as copyright goes. Now, I understand that, for example, the UK will not implement this EU Copyright Directive. What do you expect to see in the copyright and IP environment in the UK post-Brexit?

CAMPBELL: Well, I think that's a great question. And again, given current circumstances, that timeframe may stretch out a bit. I think it's interesting to see that the UK won't implement the directive. I think, in some ways, that potentially means that the UK might lean more towards some US ideas of what – of copyright jurisdiction – what that should look like. I think it's also, you know, it's very early days. We still don't really know that much about the approach to intellectual property that the current government is going to take. And I'm assuming, again, that intellectual property is rather lower on their radar right not than it might otherwise have been. But I think, certainly, if the UK does not impose the – transpose the Copyright Directive, by the end of 2021, we'll certainly start to see less regulatory alignment with the EU.

There haven't been any specific policy proposals as yet. But I think one of the aims of the current government is to be very clear – to very clearly differentiate itself from the EU in terms of some aspects of legislation. And again, that may see more of a movement towards US style – perhaps free use or those kind of ideas, which are not currently in play in European copyright legislation.

KENNEALLY: And in such a fluid environment, Duncan, it really highlights the importance of metadata and, in particular, that the metadata associated with content be as clear and specific as possible. And I wonder if you can give a picture of the relationship of metadata for licensing and all of the downstream content that comes from Wiley's journals and other publications. If again we return to the early point about the opportunity that it presents for publishers to license content for text and data mining – nonacademic uses – that metadata needs to be as crisp and precise as possible, and it will just have a real importance for the entire workflow.

CAMPBELL: Absolutely. I think it's really – again, I think I said before that I'm a real believer in having infrastructure and standards that kind of enable better and quicker, more effective uses and reuses of content. And I think that metadata is a crucial component of that. I think what we can see with – to take another example from the European Copyright Directive – Article 17, which is about sharing of content on online platforms. One of the crucial aspects of that is how to identify articles, how to identify versions of articles and the kind of – the sharing rules that apply to those articles.

So if, for example, I'm a researcher and I've downloaded some PDF files of articles to my desktop and I'd like to upload them to a platform, at what point do I know whether I have the rights to upload that? Do I know which article version I'm uploading? And so it's really important for us, as an industry, to be tagging articles with things like which article version of it, is it an author manuscript, it is accepted, is it the version of record. And also what license applies to that piece of content, so it can be understood whether it can be used and reused. And that also applies in the sort of downstream licensing for things like text and data mining.

We have a very rigorous DTD for our articles, so you obviously have information associated with them around, obviously, not just the very basic metadata around the author and the title and the journal that the article's published in, but the digital object identifier, which allows – which basically links that piece of content back to Wiley as a publisher, links it back to its journal, and then also, as I said, other information that's associated with the article that can allow a user to say what can be done with that piece of content. Is there a link in that that says there is a TDM licensing associated with this article? Is there a link that says this article version can be shared?

All of that stuff is really important, because, as more and more content is sort of shared off platform, let's say, we need to have ways of just understanding for users who are accessing that material to understand what exactly they are able to do with a piece of content that they have in front of them.

KENNEALLY: And what's very interesting to me, Duncan Campbell, is this rather unexpected – at least unexpected for me – the unexpected relationship of IP law and the technology of publishing here in 2020. They really are intertwined. And the lesson for publishers would seem to be that know your rights and make sure you get the metadata right from the start.

CAMPBELL: Absolutely. I think that's a really great point to make, Chris, that it's – as systems become more sophisticated, it's really crucial that the IP, the licensing

information is embedded in the content. That you have kind of rights at the point of content. You understand, when you're using or sharing a piece of material, what can be done with that content.

And I think, also, it's really important in the wider world, as we're thinking about open access and as we're thinking about implementing new workflows in terms of production of articles, how we make sure we're getting that right all the way down. From the author submission through all those kind of systems all the way through to our output so we know what a piece of content is, who it's been written by, who funded the article, which is incredibly important in the open access world, and then what can be done with that article. Is it published under a Creative Commons license, which allows completely free reuse and recirculation? Is it published under a slightly more restrictive right, which means you can probably do many more things with it, but you maybe need to query the publisher or you maybe need to check with a Crossref database to say have I got the right to do something with this piece of content?

So the closer the actual licensing information is embedded in the content, the more efficient and effective the – both our licensing and our technology frameworks will be.

KENNEALLY: Duncan Campbell, Senior Director of Global Sales Partnerships at John Wiley & Sons, thanks for joining me on the program and for participating in Copyright Clearance Center's virtual book fair.

CAMPBELL: Thanks very much, Chris. It's a pleasure.

KENNEALLY: For a complete schedule of virtual programming from Copyright Clearance Center originally intended for London Book Fair presentations. please visit copyright.com/lbf2020.

Beyond the Book is produced by Copyright Clearance Center. Our co-producer and recording engineer is Jeremy Brieske of Burst Marketing.

Subscribe to the program wherever you go for podcasts and follow us on Twitter and Facebook. The complete Beyond the Book podcast archive is available at beyondthebook.com.

I'm Christopher Kenneally. Thanks for listening and join us again soon on CCC's Beyond the Book.

END OF FILE