WORKSHOP 1: Copyright Infrastructure - Digiskills and metadata
1 December, 2020

Jussi Mäkinen, Technology Industries of Finland: Data governance – meta-data as the basis for data value
(transcript translated to English)

> *"…what I want to stress that, when it comes to data sharing, there are three golden letters: they're A, P, and I. So it's advisable to aim for structures and arrangements in which the data is transferred across interfaces."*

Good morning to everyone and, especially, good December – I just ate my first "joulutorttu" this morning. Thank you, Maria, for the previous presentation, and it leads us to the next subject, … it's by the way worth it to check out the book, because the technology industry believes in it too and has sent it to all student counsellors in Finnish schools. So I guess that's the best recommendation I can give you. To say a few words about myself, I have a background in copyright, like Anna mentioned in her own introduction: I have worked in copyright infringement monitoring as well as in copyright organisations, and in some way and my work experience must be almost 20 years already as to how the data infrastructures were back then. But if we think about the use of artworks, in particular, as well as the copyright business, it is largely data business – especially in the music field, where there are many performance events and data coming from various sources, in which case managing the data is key to efficient and functioning copyright economy. Let's move on to the next page. My background is, like I said, in copyright. I'm a graduated lawyer, and I have now worked more than two years in the technology industry, as the head of legal affairs in charge of digital regulation. I guess my focus has gradually shifted from the IPR side to the use of data in general, and I'm trying to combine the two expertise areas to the best of my ability. This workshop had great timing, because last week, the EU Commission gave the first legislative proposal executing the data strategy – and it's called Data Governance Act. The purpose of that is to create three different things related to sensitive public information and making that availability, and I guess the content actually closest to the data governance is the data-sharing service providers, i.e. regulation of data operators. At the end, there are also separate regulations related to data altruism. In a nutshell, the European data strategy is based on nine data spaces. The Commission wants to advance the availability of data and the use of data, especially in European industry. The way they see it is kind of that personal data was already dealt with – I don't agree with that, I mean, that GDPR was enough to deal with that – I didn't agree on that either. But the industrial data is where Europe has to succeed in a way. The European economy is, after all, based on quite a high-level industry, the productivity of which can be improved with the aid of data. I guess the most recent observation is that making use of data is a tool with which we can achieve our carbon-neutrality goals as well. As for how this data should be governed, the Commission's idea has been kind of vague, but maybe the data spaces will become more conceptual, *i.e.* not concrete data storages, rather conceptual spaces which will be based on certain standards and interoperability. This

interoperability or interconnectivity, in fact, was the idea that we would have liked to see in the Data Governance Act by the Commission. (So if in the telco circles they have identified the operators) kind of similar regulation for the operators as in the telco side, meaning that they have to remain neutral about the data; you cannot use the data transmitted for your own benefit or the data that you're processing in other ways. But the important thing that makes telco operators a universally applicable network is that the operators are interconnected. And that the data is transmitted in a certain order (--) need to modify it when moving between operators. So interconnectivity and standards. In actuality, this Act will not reinforce neither of these things, but that's the kind of work – especially when it comes to the standards – that is done within European data spaces. So it is preferable to have a certain degree of interoperability between operators, and that happens via APIs which Maria mentioned, *i.e*. via programming interfaces, via standardised interfaces, you have to get the data moving so there is no need to interpret each batch of data separately, rather that the data moves around in an automated fashion. But what's also related to data governance is that, out of the functioning interfaces and standards, a basis for networks is born, but as for the networks in which the data is transmitted, the networks also need agreements where certain practices are set as to how data moves around. And when it comes to, say, industrial data or why not personal data as well, there needs to be clarity as to what one is allowed to do with the data, what one is perhaps not allowed to do with it, who it can be given to, what other features are related to the data that play a role in how it can be used. When it comes to data in general, few rules are set by the legislation. GDPR legislation is probably the closest, regarding personal data regulation – of course, it sets certain rules. But when it comes to industrial data or data on works – which, however, is not part of the work itself, but is rather related to its authors or something like that – in that case, the rules related to the use of data are set through agreements, in case of bilateral relationships but for instance when a bunch of, a couple of operators, or, say, ten operators, so if they all make bilateral agreements, I think you have 49 of them already – so it gets tricky. And then changing the rules by changing each dyadic agreements is a logistically difficult process, so that's why the rulebooks are perhaps a handier way. What's more, in the data networks it would be good… when you try to build a functioning data market, it would be good that the identity of things, companies, and people could be verified in way that is independent from the network – if we're aiming for a multi-operator market that is the European idea, it is particularly important to strengthen these identities, or say, to verify the correctness of certain data. And I guess this is a layer, a layer of soft infrastructure that is either missing from the current data economy or is underdeveloped. And for example, in the big platforms, it has been replaced by the fact that the correctness of certain things or the runner of the platform are beyond the people's reach. And if we can build this layer of identities and trust services so it becomes functioning and strong in Europe, it is a good prerequisite for the birth of quite a flexible multi-operator competed market. Such services are already available: in Finland, a Findy cooperative is about to launch, the background of which is in expertise in decentralised systems, and that's when we're talking about universally applicable

operator who can consolidate the identity of pretty much whatever. And if we consider, say, automatic agreement or the like… in the copyright discussion, there were these micro-licences, so that can help to confirm certain transactions or certain performance events, or royalty shares can be attached to the work data, in which case it can automatically distribute the remuneration between all authors. So this is a crucial basis for a functioning data economy, that we have certain… or at least a place in the architecture, in the data models mentioned by Maria I have been taken into account.

Real-time economy is one good example as to how the standards, the interfaces, and the network are needed for everything to work smoothly for us. So it's a bit like an e-bill that goes from the biller to the payer so that there are different sorts of mediators in between that the rest don't know about. And another thing that we're supposed to make happen is that the same data would be sent to the other direction, too, in the form of a receipt. But when we reach this level of automation and reliability, we can replace these physical-world paper-based process with much smarter and automated procedures.

Shall we move on to the next page? Well, the title here is the value of industrial data. And that is actually largely based on how the organisation's maturity regarding data governance is. In that case, the meta-data are very central regarding how the value of data is defined. In the industrial environment we encounter lots of discussion where organisations think they have lots of precious and valuable data, but when you look into it more closely, you may observe certain deficiencies, or like Maria mentioned in her presentation, biases, so that it's not necessarily usable for teaching AI as it is. So when you only have a data dump and when you add certain metadata to it, such as, say, where it comes from, whether it's annotated – which means its correctness has been confirmed –, whether it has been combined with some other data, and stuff like that, combine data about the data with the data, which describes the qualities of the data. In that case, the value and the usability and the reliability reaches a whole new level. I can imagine that, in the copyright world, there are quite… say, certain things are related to artwork data, they're a bit unclear, for there isn't necessarily a uniform standard but certain practices, and what's more, data practices between different operators may vary a bit, which makes data processing quite challenging. Instead if you are able to standardise, so that the data would arrive in a certain order, for example, and if we can attach certain metadata; for example, one significant piece of metadata may be, say, year of the author's death – that could be one example, then you would know how long it's protected by copyright, it's related to how the value of the copyright is defined, it's also related to how a civil society operates, because the work becomes, for instance, free to use – things like that. And of course, when you apply that to performance data, there are even more possibilities. In the industry side, this standardising is done by this IDSA – International Data Spaces Association. That's a part of the French-German GAIA-X project, the purpose of which is to build quite an extensive chart of governance of European data, starting from the physical infrastructure, ending with

certain, say, data governance reference architectures. In one sense, it's a promising European project that seems to have a lot of traction in the European discussion. But as we've discussed it here within experts, on the other hand, it has all the characteristics of a public project that seeks to involve business operators and the third sector, but no one ever thinks of asking what it is the users actually need. But I guess the work data could be one niche that could find its home in GAIA-X. And it's a particularly good idea for people operating in culture or in the copyright field to keep tracking it and try to stay active, because it's still forming as to what they embark on – for example, could they advance, for example, certain data governance standards and the interoperability between different registers on the European level.

Shall we move on to the next slide, so we'll more or less stay in schedule. What can metadata be then? And how have they been taken into account in our existing judicial standards or tools? The idea of metadata actually comes from… after all, it's existing regulation: GDPR requires registrars to govern certain metadata related to personal data – where the data comes from, what the source of the data is, what the purpose of the processing is, to whom it can be disclosed – things like that. So the same idea has been introduced in the data-sharing model terms of our technology industry – which may inspire one when it comes to author data governance. And that's precisely where things like source, processing purpose, permission to use, life cycle are. And if you look at the other complete judicial tools for making use of and sharing data – the Rulebook for a fair data economy by Sitra – it includes similar things: you have to govern the data regarding its source, its further sharing, the restrictions related to it, permission to use, life cycle. What's more, the rulebook includes a functional section that is also helpful when creating these meta-data structures. But the Sitra rulebook is CC-licenced. It's available on the Sitra website: if you search by the word rulebook, you will find it. That may inspire you. And the model terms of our technology industry are somewhat simpler, meant to be the first step in the path of data sharing. And it is another potential source for inspiration as to how data models should be built in a way that the data would be as useful as possible.

I guess that concludes my presentation. I will gladly answer any questions, and I apologise to Maria that I referred to you incorrectly, but I hope this was a somewhat logical continuation from the first presentation. Perhaps one thing I'd still like to add is that I want to stress that, when it comes to data sharing, there are three golden letters: they're A, P, and I. So it's advisable to aim for structures and arrangements in which the data is transferred across interfaces. That's the direction of our model terms as well.