Nathalie Lefever

THE CHALLENGES OF THE METADATA INFRASTRUCTURE FOR DIGITAL WORKS AND THE ROLE OF CULTURAL POLICY

While digitalization has opened up new opportunities for disseminating creative contents, issues related to the management of metadata (data to identify works and authors) impact the capacity of creators to obtain fair copyright-based remuneration. The reasons for incomplete or incorrect metadata are the lack of properly assigned metadata, of interoperability, and of authoritative sources. This creates difficulties in the distribution of revenue to creators, editors and other rightholders, missed licensing opportunities and recognition for authors, as well as increased risks of unauthorized use. As problems of metadata directly impact digital creation, digital distribution and access to creative works, cultural policy has a role to play in their resolution. Any long-term and efficient solution will require the support of the artistic community. For these reasons, the issues related to metadata and the data infrastructure around creative works should also be discussed in the context of cultural policy.

Keywords: metadata, copyright, digitalization, remuneration, artistic work

Digitalization has opened up an unprecedented opportunity in disseminating creative contents, paving new ways of accessing cultural products and compensating artists (for a definition of digitalization and its effects in creative and cultural field, see Prokůpek, 2020). Whilst this change initially created disruptions in the markets, new platforms of exchange of works have now proven their capacity to benefit creators and users. For instance, digitalization can automate registration and distribution of content, lift up geographical hurdles, lower the cost in clearing rights and facilitate monitoring the global use of digital contents on a real-time basis. (Concerning the music industry, see Lyons et al., 2019.) However, there seems to be a disconnect between the increased digital creative consumption and how creators are paid. Whilst many factors may explain this gap, one of the major problems owes to the issues in poor management of metadata.[1]

Metadata most commonly refers to data to identify works and authors in the digital environment. This includes metadata concerning the identification of the work, the author and other rightholders on the work, the terms and conditions of use of the work, and numbers or codes that represent such information. This is the information that is considered necessary to render easier the management of rights attached to the works and other subject-matter for the purpose of distributing them on networks.[2] Metadata useful in management and remuneration of creative works also includes usage data, provided by platforms or users, which is necessary for collective management organizations (CMOs) to collect and distribute remunerations to authors[3].

The identification of works and rightholders in the digital environment is crucial for the dissemination of digital works. As problems of metadata directly impact digital creation, digital

distribution and access to creative works, cultural policy has a role to play in their resolution.

Reasons for incomplete or incorrect metadata

The first problem for the identification of works and rightholders is the **lack of properly assigned metadata**. In order for metadata to be useful for the licensing process and to ensure the recognition of the authors and other rightholders, metadata should be correctly and permanently assigned to as many works as possible. A lack of, or incomplete metadata can be the result of many possible factors, such as shortcomings in the digitization of analogue content, lack of information (for example in case of orphan works or unsettled copyright heritage cases), and lack of awareness on the importance of metadata (Council of the European Union, 2019). Correct entry of metadata, and ensuring that it is updated throughout the lifecycle of the work, is crucial to ensure that works and rightholders are properly identified and compensated in the digital environment.

Metadata can also be removed from a subject matter, in which case the work and its use can easily get out of the rightholder's control. This removal can happen accidentally, for example when missing standards for data exchange cause submitted metadata to disappear. It can also happen on purpose, a phenomenon commonly called 'data-stripping', which can be extensive in some area. For example, it was reported that 97 % of images on news websites are stripped of their credit metadata (Imatag study, 2019).

Another issue hampering the efficiency of metadata use is its **lack of interoperability and uniform formats**. Uniform formats are important to promote the availability of accurate and comprehensive information on works, in particular for collective management of rights; however, in practice different actors often apply different metadata systems. Several industry-specific identifiers have been developed to identify copyrighted works. The most successful is the ISBN – International Standard Book Number, widely used internationally[4]. More recently,

some identifiers were created with the purpose of offering a cross-domain identification method, such as the ISCC (International Standard Content Code) for digital media content or the International Standard Name Identifier (ISNI) for individuals or characters to connect artists with their credentials, including different names or pseudonyms, and provide links to other systems where information on their works and performances is held. (Lowe & Koskinen-Olsson, 2014.) However, the standards most often used are usually specific to particular industries; for example, science publishing has national standards[5], and the music industry has several descriptive tools and numerous identifiers (see Lyons et al., 2019). Their multiplicity may result in data duplication and data conflict. Additionally, the lack of uniform standards creates a risk of losing metadata when moving works between different platforms or services.

An additional challenge is the **lack of authoritative metadata sources**. Metadata must not only be kept but also indexed in a searchable way. Prospective users and rightholders, as well as collective management organisations, should have access to the information they need in order to identify the repertoire that CMOs are representing. However, in the current situation, information necessary for licensing a particular work can be difficult to trace because such information is stored in multiple heterogeneous databases. Past attempts to create centralised platforms comprehensively covering certain industries or certain areas of rights management have not been successful. For example, in the music industry, a project to create a Global Repertoire Database (GRD) failed in 2014 because no agreement could be reached on the standardization of the data to be submitted to the database. (See Lyons et al., 2019.)

Moreover, existing databases may not be accessible to all those who need them. Stakeholders who collect, update and maintain metadata do it at their own costs and for their own commercial interests. They are rarely willing to share the data collected without additional incentive, a situation which impairs the transparency of licensing processes. For example, in the music

field, the Interested Party Information (IPI) database is a global database on rightholders, their CMO affiliation and the territorial scope of entrusted rights. Access to this database is encrypted and allowed only to authorized employees from participating organizations. This system has the advantage of allowing confirmation of the data provided by rightholders who subscribe to a given CMO from other organizations, but limits the availability of identification data outside of this scope.

The fact that relevant databases are set up and maintained by private parties, often CMOs, also creates issues related to their comprehensiveness. Membership agreements with rightholders may enable CMOs to require that rightholders update and confirm the information available on their works, which improves the quality of the data. However, the use of CMOs' services is optional, so that CMOs' databases are necessarily incomplete.

Finally, various **practical and technical problems** complicate metadata management. One of them is the use of different publishing channels and changes in the ownership of rights; for example, in scientific publishing, there can be different versions of the same text or content, with different sharing policies and licensing conditions (de Waard & Kircz, 2003). Another difficulty is to keep data on licensing up to date when there is a change in the ownership of rights. Legacy content, or non-digital content that is being digitized, presents the challenge of adding metadata to large catalogues of old content; identifying publishers and rights requires important resources.

In the digital era, the rapid growth in the volume of data exchange and the emergence of diverse digital distribution outlets have resulted in unprecedented level of complexity in data management[6]. A lack of cooperation between all players has resulted in an increasing fragmentation of datasets, along with rising administration costs for data management and a duplication of data solutions built by individual organisations. This has resulted in multiple silos, fragmented datasets and the development of similar solutions operating in isolation. (Concerning

the music industry, see Lyons et al., 2019.)

The effects of inadequate metadata on creators, rightholders and users

The lack or poor quality of metadata on digital creative works causes difficulties in the distribution of revenue to creators, editors and other rightholders. Standardized metadata entries and registrations with CMOs enable automated processes for distribution of revenue streams. Poor identification of content due to missing or erroneous metadata will cause missing royalty payments or other revenue streams for rightholders that remain unidentified. In the field of music, a recent survey commissioned by the EU Commission found that metadata is missing in 10 to 50% of tracks, resulting in additional administrative costs of at least 50 million euros per year for the EU recorded music industry and possibly a licensing volume decrease of 10-50%. (Berger & Radauer, 2021.) The problem is particularly acute in the redistribution of revenue from streaming platforms, where the need for data exchange is greatest: it is estimated that 20 to 25% of music streaming revenue owed to songwriters cannot be correctly allocated (ibid.). Inaccurate data is also likely to cause delays in payments and additional transaction costs to users trying to locate rightholders. When information on the amounts of the use of works is missing, the valuation of works as economic assets can be challenging.

Poor quality of metadata also results in missed licensing, innovation and business opportunities. For example, adaptation of existing content might be slowed down due to uncertainty or fear of litigation. The lack of transparency on who has contributed to the creation of a work might also decrease job opportunities for creators. At the moment, digitalization causes concerns for a large number of artists and can affect their livelihoods both positively and negatively. A barometer survey conducted in Finland concerning the livelihood of artists showed that most of them consider clear data about copyrighted works and their authors important when working in a digital environment. Half of

the respondents were willing to make an effort to facilitate the exercise of their copyrights, e.g. by registering their works or adding rightholder data to their works. (Hirvi-Ijäs et al., 2020.)

Another consequence when a protected work is not accompanied by accurate metadata is the lack of recognition of authors and performers. Authors and performers are not likely to get proper attribution if data to identify works and authors is not accessible. For many authors and creators, the focus is less on obtaining revenue and more on sharing ideas or works and obtaining recognition. This cannot be achieved without proper and enduring identification.

Accurate and up-to-date metadata is crucial in the monitoring of the use of works. Artists will be more willing to publish their works in a digital form if they are confident that they will be able to enforce their rights thanks to metadata. From the point of view of users, a lack of metadata and the resulting difficulties in finding the rightholders of works can prevent access to or sharing of works and cause consumers to turn to unauthorised use. In a recent survey among the Finnish population, 88 % of respondents believed the author should be identified when sharing works on social media. Among those who had sometimes used unauthorised content, 11% were motivated by their opinion that the remuneration for digital works does not go to the creator. (Kautio et al., 2020.)

The role of cultural policy in securing long-term solutions

Metadata is protected under international and European law. International legal instruments[7] safeguard "electronic rights management information", or identification metadata: there is no obligation for rightholders to add rights management information to electronic works, but wherever this information is present, it has to be preserved. However, even though metadata stripping is a significant concern in particular in certain creative industries, there is little subsequent caselaw throughout Europe enforcing the metadata protection rules. This suggests that in the absence of other supporting measures, even stronger punishments would not prevent data-stripping to take place.

In the absence of efficient legal solutions, past and ongoing initiatives have emerged in different creative industries to try and solve problems related to copyright data management. These include different market-based, technological and regulatory solutions, as well as attempts to create new cross-domain identification methods. For example, the World Intellectual Property Organisation (WIPO) has developed and promoted WIPO Connect, an IT solution to facilitate the management of documentation on licensed works.[8] Artificial intelligence and blockchain solutions have recently received great attention as a means to enhance data efficiency, in particular in the field of digital music but also in other areas such as the film industry. (See for instance Council of the European Union, 2019; Tanjala et al., 2021.) At the moment, it is not clear whether any or a combination of these approaches will prevail. It also remains to be seen whether the needs of different industries, producing different types of content by creators with different needs and priorities, can be unified within a common set of metadata standards and solutions.

What is certain however is that creating a suitable infrastructure for managing metadata on creative works is crucial to ensure that creators' rights are safeguarded in the digital environment. The full potential of the digital creation and digital distribution of creative works cannot be attained if long term solutions are not found. Any effort in this direction will require collaboration between all actors involved in the value-chains of creative products and services to gain sufficient traction and make a real impact. Lack of rights awareness by both creators and consumers, lack of understanding of metadata flows and purposes, as well as inadequate cooperation between actors in creative industries have been identified as root causes for the issues around rights metadata. (Rixhon, 2021.) Current developments are driven by market needs, but market forces alone will not be sufficient. Any effective solution must be voluntary, and its governance endorsed by a

majority of stakeholders, which will require the support of the artistic community. The copyright data framework will only be sustainable if it is built on a culture of data sharing, of trust, and of collaboration for the benefit of artists and rightholders.

For these reasons, the issues related to metadata and the data infrastructure around creative works should also be discussed in the context of cultural policy. Creators must be informed of the impact of a lack of metadata on their capacity to monetize their works as well as on their right to recognition. Their taking an active role in requiring more transparency and better data management is another push towards efficient solutions. Some cultural institutions also face metadata-related difficulties which limit access to creative works by the public. For example, cultural heritage institutions may choose not to digitize or disseminate works when they are unable or find it too costly to identify their authors for licensing. (For an overview of the importance of this problem, see Martinez & Terras, 2019.) When legal solutions have been proved inefficient and technical initiatives are too fragmented, it is time for soft regulation supported by both public authorities and field actors. At its core, creating an efficient metadata infrastructure is an issue of governance and coordination. Cultural policy at the national and international level must embrace this question.

References

Berger F. & Radauer A. (2021). *The impacts of poor rights metadata on the creative industries*, Presentation for "Study on New Technologies: Copyright Data Management and Artificial Intelligence – Stakeholder Workshop on behalf of DG Connect", 24 June 2021.

Council of the European Union. (2019). *Presidency note, Developing the Copyright Infrastructure – Stocktaking of work and progress under the Finnish Presidency* (20 December 2019), https://data.consilium.europa.eu/doc/document/ST-15016-2019-INIT/en/pdf

Hirvi-Ijäs, M., Sokka, S., Rensujeff, K., Kautio, T. & Kurlin, A. (2020). *Taiteen ja kulttuurin barometri 2019. Taiteilijoiden työ ja toimeentulon muodot*. Cuporen verkkojulkaisuja 57. Kulttuuripolitiikan tutkimuskeskus Cupore ja Taiteen edistämiskeskus. (Arts and Culture Barometer 2019, Summary in English on p. 19). URL: https://www.cupore.fi/fi/julkaisut/cuporen-julkaisut/taiteen-ja-kulttuurin-barometri-2019-taiteilijoiden-tyo-ja-toimeentulo

Imatag study. (2019). *State of image metadata in news sites – 2019 update*. URL: https://blog.imatag.com/state-of-image-metadata-in-news-sites-2019-update

Kautio, T., Oksanen-Särelä, K. & Kurlin Niiniaho, A. (2020). *Suomalaisten näkemykset tekijänoikeusjärjestelmästä*. Cuporen verkkojulkaisuja 61. URL: https://www.cupore.fi/fi/julkaisut/cuporen-julkaisut/suomalaisten-naekemykset-tekijaenoikeusjaerjestelmaestae.

Lowe N. & Koskinen-Olsson T. (2014). *Management of Copyright and Related Rights in the Audiovisual Field (Educational Material on Collective Management of Copyright and Related Rights – Module 3)*. WIPO, p.42.

Lyons F., Sun Y., Collopy D., Curran K. & Ohagan P. (2019). *Music 2025 – The Music Data Dilemma: Issues Facing the Music Industry in Improving Data Management*. Intellectual Property Office Research Paper. DOI: 10.2139/ssrn.3437670

Martinez, M. & Terras, M. (2019) 'Not Adopted': The UK Orphan Works Licensing Scheme and How the Crisis of Copyright in the Cultural Heritage Sector Restricts Access to Digital Content", *Open Library of Humanities* 5(1), p.36. DOI: 10.16995/olh.335

Prokůpek, M. (2020). Digitalization of Cultural and Creative Industries and Its Economic and Social Impact. In R. Brunet-Thornton (Eds.), *Examining Cultural Perspectives in a Globalized World* (pp. 117-140). IGI Global. DOI: 10.4018/978-1-7998-0214-3.ch006

Rixhon P. (2021). *Copyright Data and New Technologies – Towards a performing copyright*

data framework. Presentation for "Study on New Technologies: Copyright Data Management and Artificial Intelligence - Stakeholder Workshop on behalf of DG Connect", 24 June 2021.

Tanjala M., Arpa S. & Saluveer S.-K. (2021). *Copyright Data and New Technologies – AI and blockchain, two keys to a level playing field and fair remuneration, perspectives from the film industry*, Presentation for "Study on New Technologies: Copyright Data Management and Artificial Intelligence – Stakeholder Workshop on behalf of DG Connect", 24 June 2021.

de Waard A. & Kircz J. (2003). "Metadata in Science Publishing", in *Conferentie Informatiewetenschap 2003 – Proceedings*, Technische Universiteit Eindhoven, p.80.

Endnotes

1. See for example Call for tenders CNE-CT/2020/OP/0009 SMART 2019/0038 (Study on Copyright and New technologies: copyright data management and Artificial Intelligence)

2. Data or metadata are protected as 'rights management information' and defined as such in Article 7 of Directive 2001/29/EC.

3. This is why Directive 2014/26/EU provides, in its article 17, that users shall provide collective management organizations with "relevant information at their disposal on the use of the rights represented by the collective management organisation as is necessary for the collection of rights revenue and for the distribution and payment of amounts due to rightholders".

4. Other identifiers include ISAN – International Standard Audiovisual Number; ISRC – the International Standard Recording Code, used ubiquitously to identify recordings; DDEX – used to convey recordings and their metadata to services globally; ISSN – International Standard Serial Number; ISTC – International Standard Text Code; ISWC – International Standard Musical Work Code; ISBN – International Standard Book Number; and ISLI – International Standard Link Identifier.

5. Science publishing is characterized by a trend towards open access to metadata gathered in large scale databases such as CrossRef or PubMed. See European Commission, Open Metadata of Scholarly Publications, July 2019, available at https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_metadata_of_scholarly_publications_0.pdf

6. For instance, PRS for Music, a UK-based copyright society, processed 18.8 trillion music performances across multiple channels in 2019, a 67,8% increase on 2018, with further, dramatic growth predicted in the future. Source: PRS for Music, Press release, 14.5.2020, available at https://www.prsformusic.com/press/2020/songwriters-and-composers-generate-record-810-million-in-music-royalties

7. WIPO Copyright Treaty (1996), EU Copyright Directive (Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC)

8. See https://www.wipo.int/global_ip/en/activities/wipo_connect/