2022-10-31

Harmonizing Open Licenses among Online Databases of Enthusiast Communities: Challenges for the Legal Integration of Databases in the Japanese Visual Media Graph Project

Schroff, S

10.54590/pop.2022.005
POP!
Canadian Institute for Studies in Publishing

Harmonizing Open Licenses among Online Databases of Enthusiast Communities: Challenges for the Legal Integration of Databases in the Japanese Visual Media Graph Project

Magnus Pfeffer, Zoltan Kacsuk, Simone Schroff and Martin Roth

Abstract

The Japanese Visual Media Graph project has created a knowledge graph for researchers working on popular Japanese visual media by combining data compiled by various enthusiast online communities. In order to open up the knowledge graph to researchers around the globe, the databases needed to be integrated legally. This article discusses the hurdles we encountered in this integration process and the solution we settled on to overcome these problems. We provide a brief look at the complexity of the legal protection afforded to databases, which we found to be an important source of problems even for communities that attempted to apply appropriate open licenses to their data. Finally, we detail how using the CC BY-NC-SA 4.0 license as a smallest common denominator and asking communities to provide us with a separate tailor-made licensing agreement helped address both the concerns of the communities and the long-term needs of the project.

Introduction

New directions in media research have long started to explore the potentials of large-scale data. In the field of Japanese visual media, such data exists thanks to the efforts of enthusiast communities and researchers alike – however, due to the complexity of the field and the diverse perspectives applied to it, these data are scattered across multiple, heterogeneous databases. One way of drawing out the potential of such data for research is to combine them into a single knowledge graph. Using RDF (Resource Description Framework[1]) technology for linking heterogeneous data is becoming more and more common in a range of fields (e.g. Adamou et al. 2019, Heath and Bizer 2011, Marshall et al. 2012). Combining databases, however, is not only a technical and ontological problem, but also a legal one.

In the present article we discuss our experience with the Japanese Visual Media Graph (JVMG) project[2] as an example of extending the logic of creating clearly licensed databases from sources outside the regular scope of research communities or libraries. Integrating these alternative data sources into the academic discourse, we feel, is a

1 RDF is a World Wide Web Consortium standard used for storing data in subject-predicate-object triples, for more details see: https://www.w3.org/TR/rdf11-concepts/.

2 The project is funded by the German Research Foundation's (Deutsche Forschungsgemeinschaft, DFG) e-Research Technologies program. The project website can be found at https://jvmg.iuk.hdm-stuttgart.de/, and the knowledge graph is available at https://mediagraph.link/.

crucial step towards more diverse and nuanced research in these fields, as enthusiast communities have created rich resources on various cultural subfields.[3] By working with these communities towards integrating their descriptive metadata resources into a single knowledge graph[4] for a specific domain – in this case Japanese visual media such as anime, manga, video games and so on – the project aims to open up new avenues of quantitative analysis for researchers in the field, and at the same time provide a template for building similar resources in other areas of inquiry. Although the creation of open knowledge graphs in the digital humanities and the cultural heritage field specifically is becoming increasingly common (see for example Bikakis et al. 2021, Haslhofer et al. 2018), integrating data compiled by online enthusiast communities is still quite novel and enables a rich range of new possibilities for research. At the same time, this approach also elicits a specific kind of legal challenge.

In the following we explain the background for our adopted solution of using the CC BY-NC-SA 4.0 license[5] as well as only incorporating parts of the source databases in our project; and how this enabled us to both address all the concerns of the participating communities and meet the licensing needs of the project. First, we briefly describe the problems related to license clarity and license compatibility in the context of database licensing. Then, we address one of the most important obstacles we came up against in trying to solve our licensing issues, namely the complexity of the legal protection afforded to databases. Last, we discuss both the licensing practices of the communities, their concerns in relation to opening up their data, and the way we managed to find a solution that could not only address these practices and concerns, but also satisfy all the licensing needs of the JVMG project itself.

Our methodology for coming up with the adopted licensing solution for the project involved a series of iterative steps. The very first step was the discussions we had with the community representatives at our first project workshop in Leipzig in the summer of 2019. This event allowed us to engage in long discussions with a larger number of online

3 In this regard the JVMG project follows in the footsteps of the Databased Infrastructure for Global Games Culture Research (diggr) project (https://diggr.link/).

4 For a detailed introduction to knowledge graphs see Hogan et al. (2021). It is important to note that a knowledge graph is still a database from a legal perspective.

5 CC stands for Creative Commons, an organization dedicated to offering alternatives to copyright with a range of available licenses (see https://creativecommons.org/). These licenses are often identified by the types of obligations and prohibitions they entail. In the above example BY stands for Attribution, meaning the need to clearly identify the original author(s); NC is the abbreviation for Non-Commercial, pointing to the fact that rights are not granted for commercial uses of the work being licensed; and finally SA means Share-Alike, adding the obligation that any derivative works have to be licensed under the same or a compatible license.

enthusiast communities[6] about their various data collection, ontology development and community management practices among other topics. Next, we studied the licenses that the communities we were working with employed, as well as the most common open license solutions, namely Creative Commons and the Open Database License (ODbL).[7] This was followed by a series of proposed licensing solutions for the JVMG knowledge graph and their detailed discussion among the co-authors of the present article. Once we reached the version that was eventually adopted for the project, we engaged in email discussions with the community representatives, where this solution was not compatible by default. These email exchanges then led to us receiving individual license agreements from the concerned communities.

The six databases from which data have been incorporated into the JVMG knowledge graph are as follows. (1) Anime Characters Database[8] (ACDB) is dedicated to collecting information on characters; while the majority are from anime, the community also catalogues characters from other media, even beyond Japanese visual media. ACDB features more than 107.000 characters from over 10.000 different works.[9] (2) AnimeClick[10] is an Italian website focusing on anime, manga and Japan related further interests. They list around 9.500 animation and more than 11.500 manga/comics titles as well as, for example, almost 40.000 creators. (3) The Visual Novel Database[11] specializes in cataloguing visual novel games only, of which they have recorded more than 71.000 releases for over 28.000 titles. (4) Media-Arts Database[12] is an initiative of the Japanese government's Agency for Cultural Affairs, which collects data on manga, animation, games and media art published in Japan. Their database lists, for example, more than 12.000 anime titles and over 170.000 manga magazine issues. (5) Wikidata[13] is a knowledge graph owned by the Wikimedia Foundation that features open data usable by anyone, and listing, among other things, almost 4.500 anime titles, close to 14.000 manga series and over 47.000 video games. Finally, (6) AniDB[14] is a fan database recording data

6 For a full list of the communities that participated in the workshop see: https://jvmg.iuk.hdm-stuttgart.de/2019/07/17/workshop-report/.

7 The Open Database License is a copyleft license for databases created by the Open Data Commons of the Open Knowledge Foundation, for more information see: https://opendatacommons.org/licenses/odbl/.

8 https://www.animecharactersdatabase.com/

9 All of the listed data sources are still actively maintained and expanded, and thus all numbers provided here are most likely obsolete by the time this article is being read, and should be taken only as indicators of the scale of the data that are being integrated in the JVMG project.

10 https://www.animeclick.it/

11 https://vndb.org/

12 https://mediaarts-db.bunka.go.jp/

13 https://www.wikidata.org/

14 https://anidb.net/

on Chinese, Japanese and Korean animation, they have close to 14.000 anime titles in their database. For the employed licenses of each of these databases see Table 2. below.

Questions of license clarity and compatibility for combining databases

There are two main problems with combining databases from a legal perspective, one is related to license clarity, and the other concerns license compatibility. Regarding the problem of license clarity, several authors have stressed how important it is to include clear licensing information for datasets that are made available online (Carbon et al. 2019, Heath and Bizer 2011). Including RDF format license information in the databases themselves,[15] especially in the case of RDF data, is also encouraged (Heath and Bizer 2011, Marshall et al. 2012). In this way the license information travels with the data and is always retrievable. Furthermore, providing license information in machine readable – not necessarily RDF – format allows for automatic license filtering and composition (Governatori et al. 2013, Villata and Gandon 2012, Wilke et al. 2021). To enable the interoperability of different license description languages in this context some research groups have suggested approaching licenses as bundles of permissions, prohibitions and obligations and creating general vocabularies and frameworks for describing them in machine readable formats (Governatori et al. 2013, Rodríguez-Doncel et al. 2013), and even taking advantage of RDF knowledge graphs to do so (Wilke et al. 2021).

The second issue of license compatibility is much more difficult to solve and requires a more hands-on approach, as we will demonstrate below. While many researchers agree that truly open data, preferably with a public domain dedication, is the most beneficial for the scientific community (Carbon et al. 2019, Marshall et al. 2012), even databases that employ open licenses often have some form of restriction in place for various reasons. In some cases, requirements that appear in different licenses of the same "family", such as "share-alike" (meaning that derivative databases need to be licensed under the same or compatible licenses) and "non-commercial" (excluding the use of the database from for-profit endeavours) used in the Creative Commons (CC) licenses, can render multiple databases incompatible on the legal level.[16] For example, a database that is licensed under the requirements of attribution and share-alike cannot be combined with one that is made available with the non-commercial and share-alike clauses, as one share-alike requirement would need the derivative database to enjoy the same openness that the original did (thus allowing for commercial use), while the other would need commercial

15 There are different Rights Expression Languages (REL) for defining licenses that are machine readable. For example, CC REL, the Rights Expression Language by Creative Commons (https://wiki.creativecommons.org/wiki/CcREL) has an RDF version among other implementations. For an overview and genealogy of RELs see Pellegrini et al. (2018).

16 See https://wiki.creativecommons.org/wiki/Wiki/cc_license_compatibility for a detailed chart of all possible compatible and incompatible pairings among CC license variations.

use to be excluded to satisfy its share-alike clause. Further problems of incompatibility arise between certain CC licenses and the Open Database License (ODbL).[17]

In our project, we encountered an additional difficulty that significantly contributes to the problem of licensing composite databases: the application of copyright (not least in the context of these licenses) and other forms of legal protection for databases is far from straightforward.

The complexity of the legal protection afforded to databases

Drawing on several detailed analyses of the problems in relation to the legal protection of databases, copyright and the available different open licenses (see for example Derclaye 2014, Giannopoulou 2018), the following section provides an overview of the various parts and aspects of databases that can fall under some form of legal protection.

Databases feature four elements that can fall under some form of legal protection:

1) The contents of the database
2) The field labels of the database
3) The structure of the database
4) The work of compiling the database

First of all, and most importantly, the content elements of a database can and do enjoy copyright protection on their own merit if they would have copyright protection when considering them by themselves individually. In this way images, audio, videos, texts (such as summaries, reviews, comments or descriptions in the types of databases we are dealing with) all fall under individual copyright protection, and their creators are the sole copyright holders unless they have somehow shared or bestowed these rights on other parties (e.g. through open licenses, signing agreements, etc.). Facts and information, however, do not fall under copyright protection, and this also includes titles and even trademarked names, which can be freely included in a database.[18] Based on these criteria certain content elements of the databases we are working with fall under copyright protection, while other content elements do not.

Second, although not necessarily often considered in relation to the legal protection of databases, should the field labels of a database be sufficiently original they can also fall under copyright protection in a large number of jurisdictions (see for example Wilson 2017).

17 For a detailed examination of these problems see Giannopoulou (2018).

18 There is some variation in the national legislation concerning the scope and implementation. See also WIPO Copyright Treaty, article 2.

Third, the structure of the database, the way that its contents are arranged can also fall under copyright protection if it is deemed original enough to be afforded copyright protection under the collection of works clause of the most widely adopted international copyright agreements such as the Bern Convention.[19] Should the structure of the database not be deemed original (e.g. alphabetical ordering of names in a telephone registry) its structure in itself will not fall under copyright protection.

However, and this brings us to the fourth element in our list, the work of compiling the database can still benefit from legal protection in this case. Under European Union law the work of compiling facts – that in themselves do not fall under copyright protection – into an unoriginal database structure is protected under what is called *sui generis database rights* (for a more in depth discussion see Derclaye 2014, Database Directive 1996).[20]

This is, however, further complicated by the fact that a database can be within the scope of contractual legal protection. This is the case for the Open Database License (ODbL), which is a license that manages the copyright and additional database rights mentioned above. As a result, the default law does still apply but the database right holder has decided on certain aspects he does not wish to enforce or at least not in a particular way. In other words, these licenses dictate the shape the law takes in this particular instance. Table 1. summarizes the different forms of non-contractual legal protection enjoyed by various features of databases with a focus on the jurisdictions most relevant for the JVMG project.

Database element	**Non-contractual type of legal protection**			
	EU	**Japan**	**US/Canada**	**Other**
Database content elements	Copyright except for facts and information	Copyright except for facts and information	Copyright except for facts and information	Copyright except for facts and information
Database field labels[21]	Copyright	Copyright	Copyright	Copyright
Original database	Copyright	Copyright	Copyright	Copyright

19 WIPO Copyright Treaty article 5, which is an addition to the Berne Convention.

20 Not all databases are afforded this type of legal protection in the EU however, as “courts from some Member States have ruled against the possibility of public bodies asserting sui generis database rights.” (Giannopoulou 2018, 5)

21 Can be copyrighted but only if they meet the national/regional originality standard.

structure				
Work of compilation if non-original database structure	Sui generis database rights	No protection	No protection	No protection

Table 1.: Non-contractual types of legal protection afforded to various elements of databases

Licensing practices of online enthusiast communities

Turning now to the actual licensing practices of online enthusiast communities, we encountered two main approaches: either the lack of any clear license information, or the adoption of one of the common open license variants from Creative Commons or Open Data Commons. However, even databases that have clear license information often suffer from various problems caused by the complex nature and interplay between copyright and database rights, as explained above. The two most obvious examples we encountered were the following.

First, most often there are no set user agreements in place that would grant the copyright of individual contributions to the community or the database. This in effect means that all individual contributions that go beyond a simple recording of factual data, such as synopses, reviews, comments, etc. all fall under the individual copyright of their respective authors and as such are not covered by the license under which the database is made available. Therefore it is impossible to assimilate any such data from these databases into a composite database licensed under an open license. Our solution to this problem was to exclude all such data elements from our consideration.

Second, the compatibility of licenses was not necessarily fully understood by the communities that built on other openly licensed resources. And even though they obviously used the concerned data in good faith, clearly displaying their origins and pertaining licenses, they were in fact violating their terms – most importantly the share-alike clause – by including them in databases with incompatible other forms of open licenses. This was not a problem for our endeavour, but rather another testament to just how difficult it is to make sense of database licensing requirements in practice.

Finding the right licensing solution for the JVMG project

In order to find a licensing solution for our project, we had to take into consideration both the fears and interests of the communities in relation to opening up their data, and the long-term needs of the project itself. The most important elements that came up in the discussions with the communities were a) the need for the acknowledgement of their work, b) the fear of having their databases copied wholesale and c) of traffic being

subverted from their sites. On the project side, we knew that we required an open license to allow researchers to freely work with the data. We also needed the license to cover most jurisdictions as both our data sources and our end-users come from a range of different countries. Finally, for the long-term extensibility of the knowledge graph, we needed a license that could act as a lowest common denominator not only for the databases we are working with currently, but for all possible future databases as well.

We settled on the CC BY-NC-SA (attribution, non-commercial, share-alike) 4.0 license, as it explicitly covers databases (including sui generis database rights), is ported internationally, and because it was our expectation that it is easier to ask communities with more permissive licenses for a separate non-commercial license for our project than the other way around. Since our aim is to make the data available for research purposes, we are not concerned about the non-commercial restriction, on the other hand some communities clearly do not want to enable for-profit entities with their efforts.

We obtained this license in separate agreements for the project from the communities who had no license in place or whose licenses would not have been compatible with this CC license. Luckily both CC and ODbL licenses permit the dual licensing of databases, thus no legal problem arose as a result of asking for a separate license agreement for the project. Importantly, these separate license agreements only cover the parts of the community databases used in our project database, thereby mitigating some of the concerns that communities had about the potential wholesale copying of their data. This was further accentuated by the fact that we also omitted all database elements that would fall under the individual copyright claims of their respective authors, such as reviews, summaries, etc. Taking into account that the JVMG web-interface also lacks any image data and many of the custom functions of the various community websites (such as personalized user accounts, comments and ratings, as well as various forms of interactivity, like the mini-games found on Anime Characters Database), the fear of subverting traffic was also found to be adequately addressed. Furthermore, thanks to the BY and SA clauses of the license, all present and future versions of the knowledge graph will have information on the source databases – thereby satisfying the need for acknowledgement in relation to the concerned communities' work – and can as a result also serve as a potential, albeit rather limited, source of reverse traffic.

In summary, adopting the CC BY-NC-SA 4.0 license along with the way the JVMG knowledge graph only builds on parts of the source databases enabled us to address all the concerns of the participating communities while also matching the needs of the project. Table 2. offers an overview of the ways license compatibility was achieved for the various databases that the knowledge graph currently builds on. For two databases with no available license information and for one with a non-compatible open license individual

license agreements were signed by the community representatives granting the use of the selected database parts under a CC BY-NC-SA 4.0 license. Two data sources had less restrictive CC licenses in place, which allow for the integration of their data into our chosen CC license. Finally, one data source has an identical license to the one we adopted and thus is also fully compatible with our knowledge graph.

Data source	**License**	**Compatibility with the CC BY-NC-SA 4.0 license**
Anime Characters Database	None	CC BY-NC-SA 4.0 license provided for the JVMG project by **individual agreement** for the parts used in each case
AnimeClick	None	
The Visual Novel Database	ODbL	
Media-Arts Database	CC BY 4.0	**yes**
Wikidata	CC0	**yes**
AniDB (publicly available anime titles only)	CC BY-NC-SA 4.0	**identical**

Table 2.: License compatibility with the databases the JVMG knowledge graph builds on

In closing we would also like to highlight that following the already cited best practice recommendations (Heath and Bizer 2011, Marshall et al. 2012) we have also included the license information for the JVMG knowledge graph in all our subgraph descriptions using the Creative Commons Rights Expression Language (CC REL).[22]

Closing remarks

In our attempt at addressing the licensing issues we found that the complexity of the legal protection afforded to databases contributes significantly to the difficulties when trying to harmonize licenses for composite databases. Our efforts to find a compatible license that mitigates community concerns and addresses the project needs drew attention to the fact that tailor-made solutions can be necessary when integrating heterogeneous data sources legally. Although the automatic composability of licenses would be a great solution for licensing composite databases in an ideal world, contacting database owners and discussing data reuse and licensing terms with them directly can be the key to

22 See for example https://mediagraph.link/graph/acdb. For more information on CC REL see: https://wiki.creativecommons.org/wiki/CcREL.

achieving true legal integration and open licensing for even larger composite database projects.

References

Adamou, Alessandro, Simon Brown, Helen Barlow, Carlo Allocca, and Mathieu d'Aquin. 2019. "Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data." *International Journal on Digital Libraries* 20 (1): 61–79.

Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). https://wipolex.wipo.int/en/text/283698.

Bikakis, Antonis, Eero Hyvönen, Stéphane Jean, Beatrice Markhoff, and Alessandro Mosca, eds. 2021. "Special Issue on Semantic Web for Cultural Heritage." *Semantic Web* 12 (2).

Carbon, Seth, Robin Champieux, Julie A. McMurry, Lilly Winfree, Letisha R. Wyatt, and Melissa A. Haendel. 2019. "An analysis and metric of reusable data licensing practices for biomedical resources." *PLoS ONE* 14 (3): e0213090.

Derclaye, Estelle. 2014. "The Database Directive." In *EU Copyright Law: A Commentary*, edited by Irini Stamatoudi and Paul Torremans, 298–354. Cheltenham, UK: Edward Elgar.

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive). *OJ L* 77 (27.3.1996): 20–28.

Giannopoulou, Alexandra. 2018. "Understanding Open Data Regulation: An Analysis of the Licensing Landscape." In *Open Data Exposed*, edited by Bastiaan van Loenen, Glenn Vancauwenberghe, and Joep Crompvoets, 1–21 (page numbers for UvA-DARE version: https://dare.uva.nl/search?identifier=5666ac49-fff1-4400-ab72-6fbb382dccb8). The Hague: TMC Asser Press.

Governatori, Guido, Antonino Rotolo, Serena Villata, and Fabien Gandon. 2013. "One license to compose them all: A Deontic Logic Approach to Data Licensing on the Web of Data." In *ISWC - 12th International Semantic Web Conference - 2013*, edited by Harith Alani et al., 151–166. Berlin and Heidelberg: Springer.

Haslhofer, Bernhard, Antoine Isaac, and Rainer Simon. 2018. "Knowledge Graphs in the Libraries and Digital Humanities Domain." In *Encyclopedia of Big Data Technologies*, edited by Sherif Sakr and Albert Zomaya, 1–8. Cham: Springer, doi: 10.1007/978-3-319-63962-8_291-1.

Heath, Tom, and Christian Bizer. 2011. *Linked data: Evolving the web into a global data space*. (Synthesis lectures on the semantic web: theory and technology) San Rafael, CA: Morgan & Claypool Publishers.

Hogan, Aidan, et al. 2021. "Knowledge graphs." *ACM Computing Surveys* 54 (4) article 71: 1–37.

Marshall, M. Scott, et al. 2012. “Emerging practices for mapping and linking life sciences data using RDF: A case series.” *Journal of Web Semantics* 14: 2–13.

Pellegrini, Tassilo, et al. 2018. A genealogy and classification of rights expression languages – preliminary results. In *Data Protection/LegalTech-Proceedings of the 21st International Legal Informatics Symposium IRIS*, 243–250, Editions Weblaw.

Rodríguez-Doncel, Víctor, Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and María Poveda-Villalón. 2013. “License Linked Data Resources Pattern.” In *Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns co-located with 12th International Semantic Web Conference (ISWC 2013)*, edited by Aldo Gangemi et al., 1–4. CEUR-WS.org.

Villata, Serena, and Fabien Gandon. 2012. "Licenses compatibility and composition in the web of data." In *Third International Workshop on Consuming Linked Data (COLD2012)*, 1–12, https://hal.inria.fr/hal-01171125.

Wilke, Adrian, Arwa Bannoura and Axel-Cyrille Ngonga Ngomo. 2021. "Relicensing Combined Datasets." *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 241–247, doi: 10.1109/ICSC50631.2021.00050.

Wilson, Neil. 2017. “The British Library experience of open metadata licensing.” In *Open Licensing for Cultural Heritage*, edited by Gill Hamilton and Fred Saunderson, 111–118. London: Facet Publishing, doi:10.29085/9781783302505.007.

WIPO Copyright Treaty. 1996. https://wipolex.wipo.int/en/text/295166