COPYRIGHT ISSUES OF IOT CRAWLERS AND SEARCH ENGINES – IS THERE ANY ANALOGY TO WEB SEARCH ENGINES?

Michal Koščík

Researcher, Institute of Law and Technology, Faculty of Law, Masaryk University
Veveří 70, 61180 Brno, CZ
michalkoscik@gmail.com; http://cyber.law.muni.cz

The publication of this paper is supported by the Czech Scientific Foundation – project ID no. GA17-22474S – «Adapting Exceptions and Limitations to Copyright, Neighbouring Rights and Sui Generis Database Rights to Digital Network Environment».

Keywords: ***Internet of things, IoT, Search engine, Copyright, Database***

Abstract: *The paper discusses the current and potential legal issues that arise with the development of tools that crawl, aggregate, index and enable search within the IoT platforms. The web crawlers and search engines opened copyright questions that had never been encountered before. Most copyright issues revolved around the appropriateness of use someone's work and its display to the «new public». The paper analyses, which lessons we learned from the development of web engines can apply to the environment of the IoT and discuss potential problems of the future.*

1. Introduction

The principal function of a search engine is to link the providers of the content with the potential users of the content. The emergence of the Internet of things (IoT) opens the door for new projects and companies that are striving to become «the next Google»[1] within the IoT world. The computer scientists are now aware that the amount of active IoT devices is increasing exponentially[2] and that humankind needs a context-based search engine for industrial IoT which will contain a mechanism that would be able to search within different devices from different manufacturers[3].

This paper analyses the copyright issues of the «IoT crawlers», «IoT search engines» or «contextual finders», which are different names for similar service that can find specific device connected to the internet based on preselected criteria. The importance of search engines in IoT increases together with an exponentially growing number of devices and sensors connected to the internet. First, the article describes the technical function of existing IoT crawlers, then it proceeds to the analysis of the rights to the raw data generated by the sensors, rights to the data and databases generated by middleware and possible rights (or lack of thereof) of a search engine to access and exploit these data.

2. The function of IoT crawlers and search engines

While there may be analogies between search engines in the content of IoT and World Wide Web («WWW»), their functionality differs in several characteristics. The backbone of the WWW is a website, the backbone of

[1] «Hayward, Martin. Industry Perspectives: Why Pay-As-You-Go Internet Pricing Calls for Careful Consideration - Streaming Media Magazine.» n.d. Accessed January 14, 2019. http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Industry-Perspectives-Why-Pay-As-You-Go-Internet-Pricing-Calls-for-Careful-Consideration–69736.aspx.

[2] Lunardi, Willian T., et al. Context-based search engine for industrial IoT: Discovery, search, selection, and usage of devices. In: Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on. IEEE, 2015. p. 1.

[3] Ibid. p. 2

the IoT is a sensor.[4] The data from the sensors are connected with an application that uses it by software layers that are generally referred to as «middleware». Middleware is defined as «*a software layer interposed between the infrastructure of devices and applications, and is responsible for providing services according to devices functionality»*[5]. As opposed to the WWW search engine, where an algorithm crawls HTML codes of websites and indexes the content, the IoT crawlers gather the information on the existence and nature of individual devices connected to the internet and index them by the nature of the data they can generate and share. The customers and end users of both types of products may be different. The end user of the IoT search engine may be another machine or computer programme. The end user of the IoT engine may be the manufacturer of the device, who wishes to learn about how his customers actually use the device. The customer may be the software developer, who will use the data as input for other application.

The most prominent search engines in IoT of today are *Shodan*, *ZoomEye*, *Fofa* and *Thingful*. Some IoT crawlers are designed to identify all components of a certain manufacturer or serving a certain purpose, such as how many «Dell 2330dn laser printers» or «CISCO switches using SSHv2» are currently online within a particular geographical area. These search engines are designed to help with the understanding of the component coverage and eventual of vulnerabilities in their networks[6].

The more advanced crawlers do not focus on mere localisation of devices and their settings but give information on the data that can be acquired from these devices in order to achieve interoperability with other devices[7]. As opposed to WWW crawlers, the advanced IoT search engines index content that would often not be consumed by a human, but rather another IoT device in order to make the device function better. The ultimate future goal of technological development is to advance indexing and crawling mechanisms to the level, where it would be possible to search and discover massive data streams in real-time[8]. The critical task is to discover, search, select, and interact with devices[9]. Similar objectives have been outlined by a three-year EU funded project named IoT crawler, which started in 2018[10].

The desired functions of an IoT search engine were summarised by Perreira et al. in 2012. An ideal search engine should be able to perform these tasks: 1) Ability to connect sensors to the IoT middleware easily, 2) Ability to understand and maintain context information (what, when, who, how, why) about sensors 3) Ability to understand the user requirement / request / problem 4) Ability to fill the gap between high-level user requirements and low-level sensors capabilities 5) Ability to extract high-level context information using low-level raw sensor data 6) Ability to manage users[11]. In other words, the IoT search engine will be able to answer questions, such as, «find me the sensors that measure temperature in this city» or «find me the data on the position of all public transport vehicles in the city of XY».

[4] Perera, Charith, et al. Ca4iot: Context awareness for internet of things. In: Green Computing and Communications (GreenCom), 2012 IEEE International Conference on. IEEE, 2012. p. 775.

[5] Ibid. p. 4, see also: See also Atzori, Luigi; Iera, Antonio; Morabito, Giacomo. The internet of things: A survey. Computer networks, 2010, 54.15: 2787-2805.

[6] See https://www.zoomeye.org/about.

[7] A good example of such service is an engine operated by «thingy».

[8] Middleware systems solutions for IoT have been developed in both research and industrial environments to supply this need. However, discover, search, select, and interact with devices remain a critical challenge.

[9] Lunardi, Willian T., et al. Context-based search engine for industrial IoT: Discovery, search, selection, and usage of devices. In: Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on. IEEE, 2015. p. 1.

[10] See the website of the project: https://iotcrawler.eu/.

[11] Perera, Charith, et al. Ca4iot: Context awareness for internet of things. In: Green Computing and Communications (GreenCom), 2012 IEEE International Conference on. IEEE, 2012. p. 775-782.

3. Copyright aspects of the IoT data

3.1. Method of the analysis

A search engine cannot function properly without aggregating content and information that is created by third parties. The search engines faced several prominent complaints for copyright violations, the most notable were the lasting dispute between the Authors guild Ind. and Google Inc regarding the indexing of printed books[12] or prominent CJEU case *Infopaq* where CJEU ruled, that a media monitoring and analysis business that aggregated and selected articles based on data capture process constituted a reproduction of copyrighted work, even if as little as 11 words were reproduced. As this paper focuses on the copyright issues of the IoT crawlers, the first point of analysis is, whether the IoT search engines aggregate the content that is actually protected by copyright. The second step of the analysis is whether such aggregation can be considered as an act of reproduction of copyrighted work or its communication to the public. The third step of an analysis is to explore, whether the search engine can enjoy any form of copyright exception under current legislation or, eventually, whether a new exception is needed in order to enable future expansions of IoT technologies.

3.2. Copyright and neighbouring rights to device-generated data

As we concluded above, the significant difference between search engines in WWW and IoT is that WWW search engines aggregate content from websites, IoT search engines aggregate data (mostly) from sensors. Most of the content of the WWW stems from human creativity, most of the content in IoT systems stems from routine operation of. The streams of data generated by sensors and devices are not literary or artistic works and they are also not intellectual creations, or any manifestation of human creativity. The data are manifestations of objective reality and not subjected to creative choice of a natural person. The data generated by sensor do not have an author within the meaning of Berne convention for the Protection of Literary and Artistic Works or WIPO copyright treaty («WCT»). Thus they cannot be protected by copyright. However, they may still enjoy protection under some of the neighbouring rights.

The streams of data generated by sensors cannot be protected as compilations of data under the Art. 10 of the Agreement on Trade-Related Aspects of Intellectual Property Rights («TRIPS»)[13] as these data are not *intellectual creations* as requested by the first sentence of the second paragraph of the Art. 10 TRIPS[14]. The machine-generated data could be however protected as undisclosed information under the Article 39 of the TRIPS, if the operator of the device takes adequate steps from making the data generated by the particular device available to the public.

The law of European union recognises unique sui generis right of the database maker, to databases that are not necessarily an intellectual creation of a natural person. The directive on the legal protection of databases[15] (hereinafter «DatD» or «Database directive») defines a database as data or other materials which are systematically or methodically arranged and can be individually accessed[16]. The database may be protected as an intellectual property even if it does not fulfil the condition of being author's own intellectual creation[17], This sui generis right of the maker of the database was introduced in order to protect «*any investment in obtaining,*

[12] See Travis, Hannibal. Google Book Search and Fair Use: iTunes for Authors, or Napster for Books. U. Miami L. Rev., 2006, 61: 87.; Opderbeck, David W. Implications of the Google Books Project Settlement for the Global Library Community. International Information & Library Review, 2016, 48.3: 190-195.

[13] The Agreement on Trade-Related Aspects of Intellectual Property Rights, adopted in Geneva on January 1, 1995

[14] Wording of the article: Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.

[15] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77, 27.3.1996, p. 20–28.

[16] Recital (17) of the Database Directive.

[17] Art. 3. of the Database Directive.

verifying or presenting the contents of a database for the limited duration of the right; whereas such investment may consist in the deployment of financial resources and/or the expending of time, effort and energy»[18]. The holder of the sui generis right to a database is granted the exclusive rights to control the extraction and re-utilisation database or it substantial part by Art. 7 of the DatD. A dataset generated by one or several sensors or devices can be considered as a database, as long as it is systematically or methodically arranged and can be individually accessed, but not every such database is eligible for sui generis protection. In order to claim protection, the person who created or holds the dataset must demonstrate that «*there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents*[19]*»*.

It is important to note that the mere purchase of a sensor or device that collects and generates data is not enough to prove substantial investment, no matter how expensive the device is. CJEU has made it clear, that one has to distinguish between the investment into the creation of the data and creation of the database in Fixtures Marketing Case[20]. In this case, the CJEU ruled that the investment in obtaining the database refers «*to the resources used to seek out existing independent materials and collect them in the database, and not to the resources used for the creation as such of independent materials»*[21]. The reasoning behind this conclusion is, that the *«purpose of the protection by the sui generis right provided for by the directive is to promote the establishment of storage and processing systems for existing information and not the creation of materials capable of being collected subsequently in a database»*[22,23]. Anybody who wishes to claim sui generis rights to any IoT data has to prove that he undertook significant effort or incurred significant costs to the collection of data generated by the devices and not to the purchase of the devices.

3.3. IP rights of middleware operators and search engine operators

In the previous subchapter we concluded, that sui generis database rights are granted to those, who invest in obtaining, verifying and presenting the contents of the database. The person who establishes gateway or operates middleware software that would identify and pool data from several devices and present them to third parties would most likely meet such criterion. If the investment is substantial, the generated data can be accessed, but cannot be extracted[24] and re-utilized[25]. This means, that the holder of these rights can transfer his rights to the database or license the rights to extract and re-utilize database rights for monetary compensation.

There is a notable difference between the IP protection of the datasets generated by individual devices and datasets generated by gateways or devices that pool several devices and index them. The IoT search engine operator has to be able to distinguish, whether he indexes raw data that are not protected by sui generis right or data that have already been filtered or processed by middleware and are protected. While the search engine is free to use and exploit data generated by devices, it needs a license to extract data from more complex datasets.

The finding that there are very few legal tools that would award any IP rights to the data generated by the IoT device might not be acceptable for operators of respective devices, as they bear costs but earn little rewards for sharing data and making their devices visible within the IoT. It could be even less acceptable to find out, that his data are commercially exploited by the third party without any right to compensation. If the

[18] Recital (40) of the Database Directive).

[19] Art. 7.(1) of the Database directive.

[20] Judgment of the Court (Grand Chamber) of 9 November 2004. Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP), C-444/02.

[21] Fixtures marketing, p. 40.

[22] Fixtures marketing, p. 40.

[23] For critical analysis of Fixtures marketing see: Obhi, Harjinder; Nettleton, Ewan. Legal update: Database right—Place your bets. Journal of Database Marketing & Customer Strategy Management, 2004, 11.4: 373-378.

[24] under the Art. 7 of the DAtD, the extraction' means «the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form».

[25] under the Art. 7 of the Database Directive, the «re-utilization» means «any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies..».

operators of devices have an interest in monetising the data, they may take the option of protecting their data via contractual limitations on the use of data. CJEU supported this approach in a *Ryanair*[26] case, but was criticized by commentators for opening ways to enclose the open data[27] and to monopolize the rights to synthetic data to the extent, where one could argue that sui generis protection is no longer necessary[28]. The second approach to gain grasp on the non-protected datasets would be to claim sui generis database rights by pooling resources among operators of similar devices, index them jointly, and claim co-ownership of database rights. Joint ownership and joint exercise of sui generis database rights brings its own problems especially in the international context[29], but mainly leads to further closure of the data from being indexed without license agreement. We can conclude that European regulatory environment may lead to the fragmentation, privatisation and closure of data sources. This can have a negative long term effects, where all protected IoT data could be bought and controlled by larger entities and consortia, which could prevent new companies to enter this market with innovative sollutions.

3.4. Use of exceptions under current and proposed directives

Building a search engine on the basis of bilateral contracts with individual points in the network is arguably very impractical if not impossible. If the regulatory environment leads to privatisation of certain datasets, it might be feasible to analyse all the possible exceptions offered by the legal framework. The copyright exceptions under the InfoSoc directive[30] are not relevant for this analysis, because the IoT search engines do not deal with copyrighted works and the InfoSoc directive explicitly excludes its applicability on the legal protection of databases[31]. The exceptions under Art. 9 of the DatD are rather limited when it comes to their practicality for search engine providers. The first of three exceptions allowed by DatD relates to the exception for extraction for private purposes of the contents of a non-electronic database; the second exception enables member states to introduce statutory exceptions for extraction of databases for research purposes and third enables exceptions for extraction and re-utlilization of database for the purposes of public security or an administrative or judicial procedure[32].

The European Comission is aware, that the current legal framework does not provide satisfactory solutions for all current problems in digital economy. This is manifested by its new Digital Single Market strategy. The Commission proposed for a new Directive on Copyright in the Digital Single Market[33] («DSM») and is preparing new regulation on the free flow of non-personal data. The proposed wording of the DSM contains explicit exception for text and data mining for research purposes in the Art. 3 which may address the problem of data privatisation, however, its scope is restricted only to the research institutions. The more thoughtful approach to the legal nature of data is visible in the working documents on the free flow of data and emerging issues of the European data economy[34] which contemplate the possible introduction to rights of the data and access rights to them.

[26] Judgment of the Court (Second Chamber) of 15 January 2015, Ryanair Ltd v PR Aviation BV, Case C-30/14.

[27] Bottis, Maria. How Open Data Become Proprietary in the Court of Justice of the European Union. In: International Conference on e-Democracy. Springer, Cham, 2015. p. 169-174.

[28] Myska, Matej; Harasta, Jakub. Less is More: Protecting Databases in the EU after Ryanair. Masaryk UJL & Tech., 2016, 10: 170.

[29] As to the disadvantages of joint wonership database right see previous works of the author: Myska, Matej; Koscik, Michal. Controlling Data in Networked Research. Internationales Rechtsinformatik Symposion: Tagungsband des 19. Internationales Rechtsinformatik Symposions, 2016, 537-544.; Koscik, Michal; Myska, Matej. Database authorship and ownership of sui generis database rights in data-driven research. International Review of Law, Computers & Technology, 2017, 31.1: 43-67.

[30] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[31] See Art. 1.2.(e) of the InfoSoc directive.

[32] DatD, Article 9.

[33] Proposal for a Directive Of The European Parliament And Of The Council on copyright in the Digital Single Market 2016/0280(COD).

[34] Most notably the working document on the free flow of data and emerging issues of the European data economy SWD(2017).

4. Conclusions

The proportion of copyrighted content that is aggregated by IoT search engines will arguably be much lower on the IoT platforms. Nevertheless, it may be still present. The paper concluded that IoT search engines will most likely not encounter with copyrighted artistic works but will aggregate quantities of both raw and processed data. The owners or operators of devices (hardware) that create the data have very few intangible rights guaranteed by law but may opt for their prioritisation via contractual methods. The data can be purposefully aggregated and indexed in order to receive the sui generis protection under the database directive. We identified a strong disbalance in current European law, where data can be privatised without adequate exceptions for search and contextual engines, which can create either significant obstacles to the development of further search engines or monopolisation of the rights to IoT data by resource-rich entities.

We analysed the impact and potential applicability of proposed Directive of the European Parliament and of the Council on copyright in the Digital Single Market and the prepared regulation on the free flow of non-personal data. We did not find direct impact of the DSM for the IoT applications, but future regulatory developments in the area of free flow of non-personal data may bring systematic changes that would be beneficial to this field of electronic industry.

5. References

Atzori, Luigi; Iera, Antonio; Morabito, Giacomo. The internet of things: A survey. Computer networks, 2010, 54.15: 2787-2805.

Bottis, Maria. How Open Data Become Proprietary in the Court of Justice of the European Union. In: International Conference on e-Democracy. Springer, Cham, 2015. p. 169-174.

Hayward, Martin. Industry Perspectives: Why Pay-As-You-Go Internet Pricing Calls for Careful Consideration - Streaming Media Magazine.» n.d. Accessed January 14, 2019. http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Industry-Perspectives-Why-Pay-As-You-Go-Internet-Pricing-Calls-for-Careful-Consideration–69736.aspx.

Koscik, Michal; Myska, Matej. Database authorship and ownership of sui generis database rights in data-driven research. International Review of Law, Computers & Technology, 2017, 31.1: 43-67.

Lunardi, Willian T., et al. Context-based search engine for industrial IoT: Discovery, search, selection, and usage of devices. In: Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on. IEEE, 2015. p. 1-8.

Myska, Matej; Harasta, Jakub. Less is More: Protecting Databases in the EU after Ryanair. Masaryk UJL & Tech., 2016, 10: 170.

Myska, Matej; Koscik, Michal. Controlling Data in Networked Research. Internationales Rechtsinformatik Symposion: Tagungsband des 19. Internationales Rechtsinformatik Symposions, 2016, 537-544.

Obhi, Harjinder; Nettleton, Ewan. Legal update: Database right—Place your bets. Journal of Database Marketing & Customer Strategy Management, 2004, 11.4: 373-378.

Perera, Charith, et al. Ca4iot: Context awareness for internet of things. In: Green Computing and Communications (GreenCom), 2012 IEEE International Conference on. IEEE, 2012. p. 775-782.

Opderbeck, David W. Implications of the Google Books Project Settlement for the Global Library Community. International Information & Library Review, 2016, 48.3: 190-195.