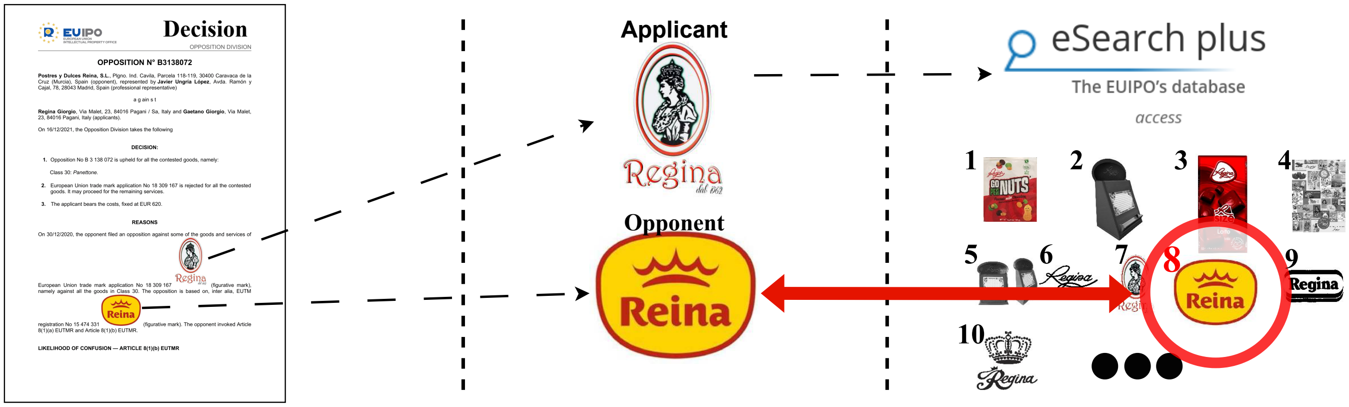
A Quantitative Evaluation of Trademark Search Engines' Performances through Large-Scale Statistical Analysis

Thomas Vandamme
thomas.vandamme@ulb.be
Université libre de Bruxelles
Brussels, Belgium

Julien Cabay
julien.cabay@ulb.be
Université libre de Bruxelles
Brussels, Belgium

Olivier Debeir
olivier.debeir@ulb.be
Université libre de Bruxelles
Brussels, Belgium



EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Decision

OPPOSITION DIVISION

OPPOSITION N° B3138072

Postres y Dulces Reina, S.L., Plgno. Ind. Cavila, Parcela 118-119, 30400 Caravaca de la Cruz (Murcia), Spain (opponent), represented by **Javier Ungría López**, Avda. Ramón y Cajal, 78, 28043 Madrid, Spain (professional representative)

against

Regina Giorgio, Via Malet, 23, 84016 Pagani / Sa, Italy and **Gaetano Giorgio**, Via Malet, 23, 84016 Pagani, Italy (applicants).

On 16/12/2021, the Opposition Division takes the following

DECISION:

1. Opposition No B 3 138 072 is upheld for all the contested goods, namely:

   Class 30: *Panettone.*

2. European Union trade mark application No 18 309 167 is rejected for all the contested goods. It may proceed for the remaining services.
3. The applicant bears the costs, fixed at EUR 620.

REASONS

On 30/12/2020, the opponent filed an opposition against some of the goods and services of European Union trade mark application No 18 309 167 (figurative mark), namely against all the goods in Class 30. The opposition is based on, inter alia, EUTM registration No 15 474 331 (figurative mark). The opponent invoked Article 8(1)(a) EUTMR and Article 8(1)(b) EUTMR.

LIKELIHOOD OF CONFUSION — ARTICLE 8(1)(b) EUTMR

Figure 1: Our method's principle.

ABSTRACT

Intellectual Property Offices now offer their users trademark search engines to help them identify earlier trademarks in their register. Such tools have proven to be extremely useful given the growing number of trademarks registered but have never been subjected to thorough evaluation, despite the necessity for openness and accountability in justice systems. Additionally, their performance is unknown, in particular the reliability of their results pertaining to applicable legal rules. In fact, their "black box nature" makes automatic and at-scale evaluation hard to perform directly, which is why we propose a novel method for evaluating their performance using settled case-law for ground truth, and at-scale analysis. Based on this methodology, we evidence the performance for two such systems, the Benelux Office of Intellectual Property (**BOIP**) and European Union Intellectual Property Office (**EUIPO**), using 8 126 opposition division decisions from the EUIPO. We show important disparities between the two systems, along with surprisingly good results for EUIPO's system.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; *Image search.*

KEYWORDS

Intellectual Property, Trademark, EUIPO, BOIP, Search Engine, Likelihood of Confusion

ACM Reference Format:
Thomas Vandamme, Julien Cabay, and Olivier Debeir. 2023. A Quantitative Evaluation of Trademark Search Engines' Performances through Large-Scale Statistical Analysis. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3594536.3595149

1 INTRODUCTION

According to settled case law at the European Union (EU) level, "the essential function of the trade mark is to guarantee the identity of the origin of the marked product to the consumer or end user by enabling him, without any possibility of confusion, to distinguish the product or service from others which have another origin"[1].

The proprietor of a registered Trademark (**TM**) in the EU is therefore entitled to prevent uses in the course of trade[2], as well as to oppose to the registration as a TM[3], of a sign that triggers a Likelihood of Confusion (LoC) with its own earlier TM.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06.
https://doi.org/10.1145/3594536.3595149

[1] CJEU, 28 September 1998, *Canon Kabushiki Kaisha v Metro-Goldwyn-Mayer Inc.*, Case C-39/97, EU:C:1998:442, point 28

[2] Art. 9(2)(b) of Regulation (EU) 2017/1001 of the European Parliament and of the Council of 14 June 2017 on the European Union trade mark (codification), OJ L 154/1 (hereafter EUTM Regulation) ; Art. 10(2)(b) of Directive (EU) 2015/2436 of the European Parliament and of the Council of 16 December 2015 to approximate the laws of the Member States relating to trade marks (Recast), OJ L 336/1 (hereafter EUTM Directive)

[3] Art. 8(1)(b) of EUTM Regulation ; Art. 5(1)(b) of EUTM Directive

The test for confusion factors in several criteria (similarity of marks, similarity of goods and services, distinctiveness of the marks) that are appreciated globally, from the point of view of the average consumer[4]. As to the assessment of the similarities between the marks, in the seminal case SABEL v Puma the Court of Justice of the European Union (CJEU) stated that a "global appreciation of the visual, aural or conceptual similarity of the marks in question, must be based on the overall impression given by the marks, bearing in mind, in particular, their distinctive and dominant components"[5]. This test has changed very little since then.

In the EU legal system, those principles govern both EU TMs and National TMs, pursuant respectively to the EUTM Regulation and the EUTM Directive implemented in all Member States. Though autonomous and independent, an EU TM and a National TM can coexist for the same territory. In the particular case of Belgium, Luxembourg and the Netherlands, national laws have been unified in one regional law covering the three territories, namely the Benelux Convention on Intellectual Property (CBPI). The Benelux TM has accordingly replaced the national TM in those three countries. Yet it is governed by the exact same principle as those mentioned above.

Registrations of TM have been growing over the years, especially at the EU level. Indeed in 2021, about 198 000 TM applications were made at the EUIPO, a volume increasing year on year (in comparison, "only" 106 000 applications were made in 2011; 49 000 in 2001)[10]. This high volume, and the expectation of even more in the coming years, stressed the importance to consider automated solutions for helping human beings, be it IP Offices' staff members or customers of their services to process and access to relevant information in relation to those registrations. In its Strategic Plan 2025, the EUIPO decided to develop AI based solution, including image search tools[9]. In this framework, it has released its first in-house image search tool, that was made available on the 29th of November 2021[8]. This tool can be accessed on the EUIPO website, via eSearch Plus[6].

Similarly, on the 2nd of October 2020 the BOIP announced that a private company had licensed to them a "new trademark image recognition technology for use within the trademark register available to the public on the BOIP website, as well as integrated in the internal tools to be used by examiners"[7]. This tool can be accessed on the BOIP website[7].

Both the EUIPO and BOIP image search interface provides the user with the possibility to upload an image, and search for earlier TM valid in their respective territories (namely European Union and Benelux). As a matter of law, EU TM are valid in both territories and can therefore be retrieved with either tools. In addition, the BOIP offers the possibility to identify Benelux TMs (that are valid in this territory only).

Following their purpose, those tools will give anyone with interest the possibility to assess risks of conflicts between existing TMs and new applications in the relevant territory. Those interested parties are mainly TM users who are self-applicants (businesspeople), IP professionals (TM attorneys, agents), as well as IP Offices staff members (examiners or lawyers). Business cases are numerous, and attracted recently the interest of IP scholars.

So far, legal literature is however rather scarce as to the potential implications of relying on such tools. In the USA, the most relevant study by Katyal and Kesari [4] was dedicated to a thorough assessment of some of those tools (private and public). The authors attempted an interdisciplinary study into search engines ability to identifying potential conflicts under Section 2(d) of the US Trademark Act, 15 U.S.C. § 1052(d), which forbids the registration of a TM that is confusingly similar to an existing registered TM. They however limited the scope of their study to word TMs, excluding (semi)figurative TM, and considering US TM Law only, which differs to some extent of EU TM Law. More recently, Lim [5] stressed some of the issues associated with deploying AI-enabled likelihood of confusion analysis, and suggested to use empirical studies as training data. In the EU, Gangjee in particular has mapped the initiatives of IP Offices and potential legal issues associated with the use of such tools [2, 3]. Only Moerland and Freitas [6] ran limited tests (seven types of search) on the IP Offices image search tools.

Hence, those tools have so far not undergone thorough scrutiny, making elusive the examination of potential legal issues associated with their use. We tried to fill this gap through performing an at-scale analysis of EUIPO's and BOIP's search engines, which are only two amongst several[8].

From the outset, those tools have helped offices automate parts of their procedures and provide the public, applicants and law firms with TM search engines able to search for image TMs based on specific queries.

While these solutions are welcome and might even be deemed necessary with an overload of TM registrations, they currently exhibit major drawbacks. First of all, the models used for the similarity assessment are not public, preventing the evaluation and criticism of these by outside actors. Secondly, the systems are based on Deep-Learning techniques, probably similar to the solutions described in the literature. Perez et al. [11] use two fine-tuned VGG16 neural networks, and report enhanced performances from previously used classical methods. Tursun et al. [13] propose a solution based on a pre-trained neural network as a feature extractor, which they improve by removing the text from the images. Notably, Trappey et al. [12] use deep learning methods to assess similarity of TMs under different aspects (figurative, spelling and phonetic), and report similarity estimations on real-life cases. Finally, deep learning methods are heavily criticised for their susceptibility to biases and absence of explainable behaviour. Moreover, given the purpose of identifying relevant TM likely to trigger a LoC, features of these

[4] See in general Fhima and Gangjee [1]

[5] CJEU, 11 November 1997, *SABEL v Puma*, Case C-251/95, EU:C:1997:528, point 23

[6] https://euipo.europa.eu/eSearch

[7] https://boip.int/en/trademarks-register

[8] Amongst the publicly available tools, see also the one developed by the World Intellectual Property Office (WIPO) : https://branddb.wipo.int/en/similarlogo

different techniques are subject to criticism from a legal standpoint.

In relation to those tools, we addressed and compared the performances of the EUIPO's and BOIP's tools, using 8 126 opposition division decisions from the EUIPO for evaluation purposes. Our results substantiate some of the criticisms, and emphasize the need for caution when using such tools.

2 METHODOLOGY

As a ground-truth for our comparison, we used the administrative decisions of the EUIPO's Opposition Division, from 23/03/2016[9] to 31/05/2022. All those decisions addressed a LoC as to (semi)figurative TMs, pursuant to Art. 8(1)(b) and 46 of EUTM Regulation.

As a matter of law, since the substantive TM Law principles governing EU and Benelux TMs are identical, such a dataset can be used for the purpose of testing and evaluating both EUIPO's and BOIP's tools. As a matter of fact, independent of the outcome as to the LoC, we make the fair assumption that the marks in conflicts discussed in each decision of this dataset were similar to some extent, at least from the standpoint of the earlier TM proprietor (the Opponent) who considered appropriate bearing the costs of opposing to the registration of the subsequent TM (by the Applicant).

We consider that those tools prove efficient if they are able, following the upload of the Applicant's TM under discussion in this case, to display amongst the results the Opponent's TM (hereafter a Match). From a practical perspective, in each business case one would expect a search engine to retrieve the most relevant results earlier. Given this, for each decision, we use the Applicant's TM as a query to the systems, and retrieve the position (if any) of the corresponding Opponent's TM. Performances can be assessed along two axes, the first one being the proportion of queries that resulted in a match (hereafter the **Match Ratio**), the second is based on the position of the match in the results (hereafter the **Rank**), when there indeed was a match. Given that the BOIP's tool only display 50 results, for the purpose of the comparison with the EUIPO's, we consider a virtual system where all results above the first 50 of the EUIPO's are discarded (hereafter **EUIPO-50**).

Finally, we evaluate the potential reasons for the differences in performance for these two systems, using statistical analysis. We demonstrate that such an approach can be used to both evaluate TM search engines from a practical standpoint, as well as investigate their strengths and weaknesses.

2.1 Data

For our analysis, we used the administrative decisions of the EUIPO's Opposition Division, from 23/03/2016 to 31/05/2022. The dataset was built through compiling informations from legal databases. It contains all decisions where both parties had a figurative mark, leaving us with 8 126 decisions to test for. For each of these, we acquired from the EUIPO public database eSearch Case Law[10] the corresponding decision document, from which we extracted the high-resolution TM images, as well as the TMs' unique ID numbers (**TM ID**). We paired this TM IDs with the unique identification number of their holders (**Holder IDs**), whose information is publicly available via eSearch Plus.

Because of the nature of this experiment, and the dependability on external databases, a few issues arose. Firstly, we were not able to retrieve the Holder ID for some decisions. Secondly, in some decisions, the high-resolution images could not be acquired, due to corrupted images. Lastly, and most importantly, we could not acquire some decision documents (and therefore no high-resolution images), accounting for 12% of the relevant cases we identified. The numbers of decisions concerned by each case is given in Table 1.

All decisions and ID number we use were lawfully acquired from publicly available databases. We provide an Excel file containing the list of decisions used[11]. It also includes the search engines' performance for each decision. We do not, however, disclose the specific search engines' outputs as well as the decisions, as it would fall beyond the scope of authorized uses under the Text and Data Mining exception for the purposes of scientific research[12].

2.2 Query principle

From the 8 126 decisions, we isolated decisions where the Applicant is presented one single image, which is compared to an Opponent with a single image. We query the search engine with the applicant's image and report the position of the first image invoked by the opponent for the procedure in the search engine output. This procedure is represented in Fig.1. In this example[13], the upper TM is the one applying for registration (the Applicant's TM), and the bottom TM is used as a basis for the opposition claiming there is a LoC (the Opponent's TM). We enter the upper TM into a search engine, and retrieve the results as a list of TMs, in a decreasing order of similarity. If a Match is found, we consider the position of the match as the Rank of the search. In order to assure that the search engines only use the image as a basis for similarity assessment, all the query images are stripped of metadata and are renamed a generic name prior to upload.

This approach was chosen in order to assess the quality of the search results, since we expect similar TMs (as deemed by the decisions) to be found with minimal rank. This constitutes a practical proxy for the performance of the system, as it relies on real-life business cases, and to the direct purposes of these systems. This is, in effect, the first time this approach has been taken for this problem, with this volume of real-life data.

[9] Date of the entry into force of the EUTM Regulation (EU) 2015/2424, subsequently codified in EUTM Regulation (EU) 2017/1001

[10] https://euipo.europa.eu/eSearchCLW/

[11] Available here (https://github.com/thovdamm/ICAIL23-Evaluation-Trademark-Search-Engines), or upon request at thomas.vandamme@ulb.be

[12] Art. 3 of 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, OJ L 130/92

[13] EUIPO Opposition Division, 16/12/2021, *Regina dal 1962 v. REINA*, Opposition N° 3138072

This process might seem flawed, in the sense that we cannot verify that the images appearing before the Opponent's TM are not relevant from a legal perspective. Indeed, some of those might also trigger a LoC, according to TM Law principles. However, the impact of this flaw on the relevance of our analysis is limited to some extent, as TM proprietors are eager to oppose registration of conflicting TMs[14]. One can also assume they have some information as to their market and competitors, and will oppose the TMs which conflict the most with their own. It is true, however, that they might disregard marks that are similar yet not likely to raise LoC concerns, especially when they are concerned with goods and services different enough from the ones of their own TM[15]. Hence, arguably similar TMs might appear in the results with a lower rank. This is a limitation of our method, due to the nature of the problem: we of course do not have the similarity information for every possible pair of TMs.

The applicants' trademarks used for the queries originated from the EUIPO registry, where the images' quality is best, whereas the images present in the decisions are of reduced quality and resolution. This fact ensures the optimal behaviour of the systems. We could have degraded the images in various different ways (compression artefacts, noise, colourspace manipulations, ...), as a mean to assess the robustness of the systems. However, while this information would be of interest, doing any manipulation would, in our opinion, invalidate the legal conclusion upon which we base our approach. Moreover, the time needed to perform such research without having access to the underlying model constitutes a significant undertaking, outside the scope of this research.

2.3 Match criteria

Our process implies assessing the identity between the Opponent's TM referenced in the decision and the TMs retrieved by the search engine. For this, we consider two options: the first is by considering the exact match between the TMs, using a unique identifier (by TM ID), while the second considers any TM owned by the same owner as a match (by Holder ID).

Considering that the opposition by the Opponent was based on one TM, but that this proprietor could own other very similar TMs (i.e. with slight variations), considering a perfect match is going to present potentially underwhelming results. Indeed, it could happen that these similar TMs appear before the specific TM from the opposition, which would result in a diminished performance assessment. Conversely, some owners are in possession of a massive amount of marks, and considering a match with any one of those could be optimistic. Both situations are illustrated on Fig.2, where all three TMs are owned by the same holder. Assuming the TM 2b is the one used as a basis for an opposition, the presence of the TM 2a would not constitute a match under the TM ID criterion, while the presence of the TM 2c would, under the Holder ID criterion. There does not exist any realistic solution to this problem, since both of those criteria could only be enhanced with real-life expert evaluation of those, ideally from boards of opposition. Furthermore, the number of evaluations to be conducted would far exceed the number of applications.

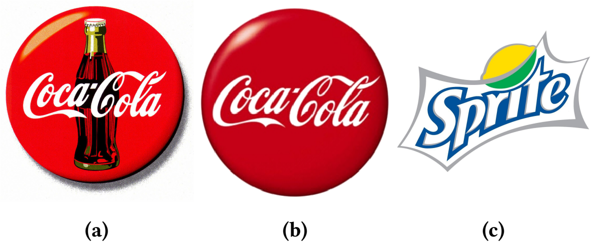


Figure 2: Three TMs owned by the same organization. a and b would not match under the TM ID criterion, while b and c would match under the holder ID criterion.

Hopefully, these two criteria yield respectively lower and upper bounds on the performances of these systems. We conducted the experiments for both of them, and evaluated the tightness of the bounds. These results will be discussed in Section 3.3.

2.4 Search Engines Specifics

We conducted the experiment for two different search engines; the BOIP's and the EUIPO's[16]. As explained, both were chosen because of their relevance with regards to EUIPO's opposition decisions. Furthermore, these solutions are both publicly available and offered by official TM registries.

There are however major differences between the two. Firstly, whereas the EUIPO's tool has been developed in-house by this public body, the BOIP's tool is licensed through a private company. Secondly, the output that can be displayed in both tools largely overlap (TMs valid in the whole EU) but do not coincide (only the BOIP's tool will display TMs valid in the Benelux only). Thirdly, the number of output TMs displayed differ greatly: the BOIP's yields a maximum of fifty TMs, while the EUIPO's yields tens of thousands. In our study, we only consider the first thousand results of the EUIPO, for a practical reason: the output of the search can be downloaded in Excel format up to 1000 results.

Since the decisions used originated from the EUIPO, we know the applicant applied for registration as an EUTM and that the opponent has based its opposition on either a registered (or application of an) EUTM, or a TM (or an application for a TM) with effect in the EU or one of its Member States[17]. We were not able either to determine automatically which decisions were issued, after an opposition

[14] For the past ten years, the number of oppositions filed each year ranges from 15 658 (in 2014) to 20 125 (in 2021). From the starting of the Opposition Division's work in 1997, almost 400 000 opposition have been filed[10].

[15] According to settled case law, "A global assessment of the likelihood of confusion implies some interdependence between the relevant factors, and in particular a similarity between the trade marks and between these goods or services. Accordingly, a lesser degree of similarity between these goods or services may be offset by a greater degree of similarity between the marks, and vice versa", see **CJEU, 28 September 1998, *Canon Kabushiki Kaisha v Metro-Goldwyn-Mayer Inc.*, Case C-39/97, EU:C:1998:442, point 17**

[16] The queries were performed from 30/11/2022 to 06/01/2023 for the EUIPO and from 30/11/2022 to 06/12/2022 for the BOIP. Unfortunately, the systems do not provide a software version number.

[17] Art. 8(2) EUTM Regulation

Issue	Number of cases	%
More than one Fig. TM	1 121	13.80%
No Holder Data	96	1.18%
No decision document	963	11.85%
No High Resolution Image	89	1.10%
No issues	5 857	72.08%
Total	**8126**	**100%**

Table 1: Issues in decision acquisition and the number of decisions concerned by each kind of issue.

System	Potential Queries	Effective Queries	Issues
BOIP	5 857	5 761	96
EUIPO	5 857	5 852	5

Table 2: Queries malfunctioning for either systems.

based on an EUTM or other earlier TM. Furthermore, even if we could determine this, the results would still present a bias since holders can have a similar TM be registered both nationally and at the European level, and the choice of one or the other to oppose a registration is dependant on the situation. This will be further discussed below.

3 RESULTS

3.1 Query Numbers and Issues

Out of the 8 126 decisions considered, because of the nature of the processing/data acquisition pipeline (i.e. we have to query external databases from their user interface), certain issues arise. Our method requires different bits of information, in order to evaluate which marks are concerned by the decision, who are the holders and what are the TM IDs, and extract the TMs from the decision file.

Each of these steps could be incomplete for a given decision, as the databases might refer to a corrupted decision document, the images in them could be themselves corrupted, or the databases might not yield results for the owners information (such as the TMs they own). These issues are common for both search engines, since they take place before querying. They are further detailed, along with the number of decisions concerned by each of those, in Table 1.

It should be noted that in those issues, the absence of a decision document, or of a high-resolution image leads to the impossibility to perform the search, since we do not have any information about the decision.

On top of these, the search engines themselves would occasionally not run for specific images, for reasons unknown to us. Table 2 gives the number of decisions ran for each system, as well as number of queries that failed to yield results.

System	Criterion	Decisions	Matches	Ratio
BOIP	Holder ID	4 784	441	9.22%
	TM ID	5 761	445	7.72%
EUIPO	Holder ID	5 852	4 023	68.75%
	TM ID	5 852	3 647	62.32%
EUIPO-50	Holder ID	5 852	3 626	61.96%
	TM ID	5 852	3 626	61.96%

Table 3: Performances in terms of matches for both systems and for both match criteria. The difference in considered decisions for BOIP is due to issues when retrieving holder data (not included in the Holder Data bugs of Table 1).

System	Criterion	Decisions	Matches	Ratio
BOIP	Holder ID	4 762	419	8.80%
	TM ID	5 757	424	7.36%
EUIPO-50	Holder ID	5 757	3 566	61.94%
	TM ID	5 757	3 570	62.01%

Table 4: Performances for both systems on the same set (decisions ran for both BOIP and EUIPO). The issues for holder comparison at BOIP are deducted from the decisions possible.

3.2 Match Ratio

The performance of the systems is comprised of two metrics: the number of decisions for which the system correctly found the opponent, i.e. the **Match Ratio**, and the average position of the opponent in the output, when it was found, the **Rank**.

Table 3 gives the different match ratios for both systems, and for the two match criteria considered. The difference in decisions considered between both criteria for BOIP's system is due to issues when retrieving holder information. This is a necessity because of the absence of holder information in the query output, we have to try and match all possible TMs owned by the owner.

As expected, both criteria exhibit slightly different performances in terms of Match Ratio, with the Holder ID one being superior to the TM ID one.

Given that the opposition could be based on a TM that is absent from the BOIP's or the EUIPO's registries (e.g. in the case the opposition is based on a TM registered in a EU member state), the Match Ratio of either system should be computed solely using the decisions based on a TM present in its associated registry. Unfortunately, we could not determine automatically which decisions were concerned. However, we can deduce lower and upper bounds on the match ratio for each system.

Fig. 3 illustrates the different types of decisions present in our dataset. The main separation lies in the presence or not of the TM in the EU registry (lower right part of the diagram). By that definition, all EUIPO matches (Red hatching) lie in the upper part,

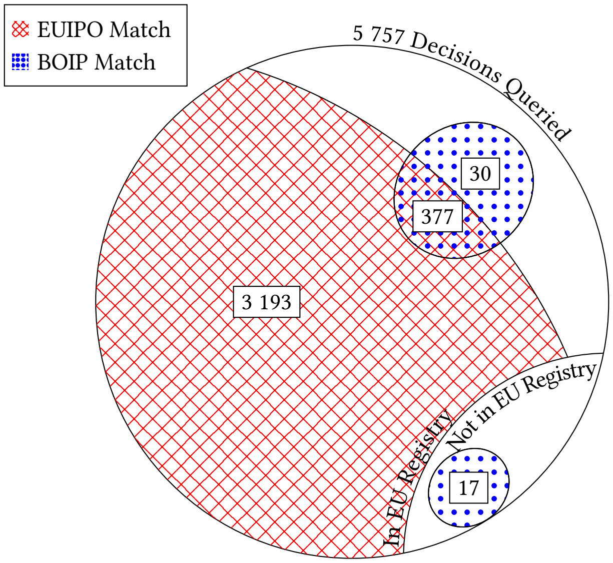
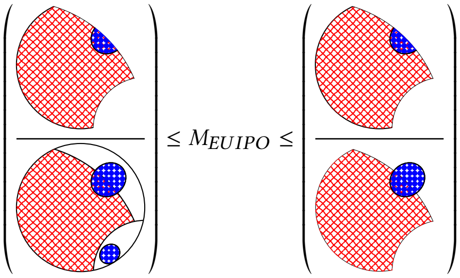


Figure 3: Representation of the different sets present in the dataset.

and the TMs in the lower part should not be considered for the EUIPO computation. BOIP's registry, in addition to containing the EUIPO's registry, also contains its own TMs (those registered in the Benelux). Therefore, some could be matched at BOIP's level only, and some could not be match in any case. After examination, we found that 17 matches were made on decisions based on Benelux opposition.

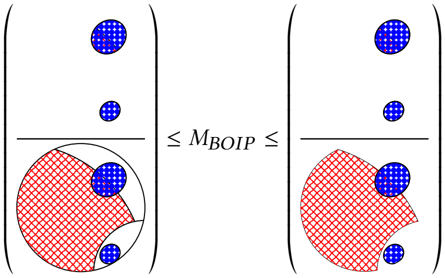
The bounds for the EUIPO's system are as follows, where M_{EUIPO} denotes the Match Ratio of the EUIPO:



$$\left(\frac{3570}{5757}\right) \leq M_{EUIPO} \leq \left(\frac{3570}{3600}\right)$$

$$62.01\% \leq M_{EUIPO} \leq 99.16\%$$

Similarly for BOIP:



$$\left(\frac{424}{5757}\right) \leq M_{BOIP} \leq \left(\frac{424}{3617}\right)$$

$$7.36\% \leq M_{BOIP} \leq 11.72\%$$

The BOIP's system performance is at least multiple times lower than that of the EUIPO's (truncated at 50), in terms of Match Ratio. Moreover, the EUIPO's match ratio is at least of 62%, which seems very good, given the complexity of the problem.

3.3 Ranks

Rank distributions are shown in Fig. 4, for BOIP and EUIPO-50. Note that ranks are 1-indexed, such that a rank of 1 means that the very first result was a match. Since the query itself is most likely present in the database, a rank of 2 is probably the best achievable result (at least for EUIPO, since the database contains the applicant TMs, which is not necessarily the case for BOIP). With this in mind, we observe an almost perfect performance of the EUIPO's system, while BOIP's seems more natural. Indeed, given the complexity of the task, we question how the EUIPO's system achieves to retrieve more than 60% of the queries, and most often in the second position (which is usually the best achievable).

To assess the consistency between the two rank criteria, we compute the difference of rank for each query for both systems. Fig. 5 shows the difference in ranks, where the rank by TM ID is subtracted from the rank by TM holder. This subtraction is done for every query, and the figure gives the number of queries for each rank difference. We are considering queries which gave a match for both methods, and this concerns 291 queries for BOIP and 3 327 queries for EUIPO.

The graphs only show rank differences between -5 and 5. Considering this interval, the EUIPO graph includes 3 240 of the 3 327 (97%) queries, while the BOIP graph includes 264 of the 291 (91%) queries. For the vast majority of queries, performing the analysis with one criterion or the other does not affect the performance.

Moreover, we perform a two-sample Kolmogorov-Smirnov test between the distributions for either criteria, for each system. The null hypothesis is that both distributions are identical. The test on BOIP yields a p-value of 0.099, while the test on EUIPO yields a p-value of 1.0. Since both are superior to our threshold of 0.05 (for a confidence of 95%), we accept the null hypothesis that both

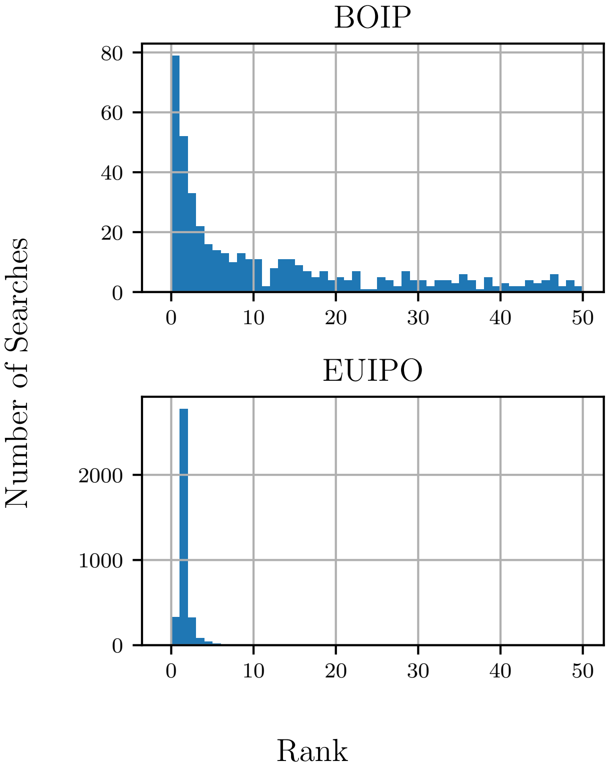


Figure 4: Rank distributions for BOIP and EUIPO. Match using TM ID.

distributions are statistically equivalent.

3.4 Differentiated Analysis

Using the decisions' information, we can try to derive useful knowledge about the search engines' methods. One important aspect of these is the presence or not of a LoC, as determined by the examining body. There are three kinds of outcomes for such decisions, either there is a complete refusal of the registration (LoC +), or a partial acceptance for only a portion of the goods and services (LoC +/-), or the dismissal of the opposition (LoC -). Performing the previous analyses by differentiating for this LoC, we expect to gain insights into the inner workings of the search engines.

The match ratios for this differentiated analysis are presented in table 5, along with the mean ranks. We can observe that the difference in performances is very limited with regards to the outcome of the decision used for the query. This is surprising, given that the search engines are supposed to retrieve relevant TMs, and TMs that presented a LoC of course are more relevant than those that did not.

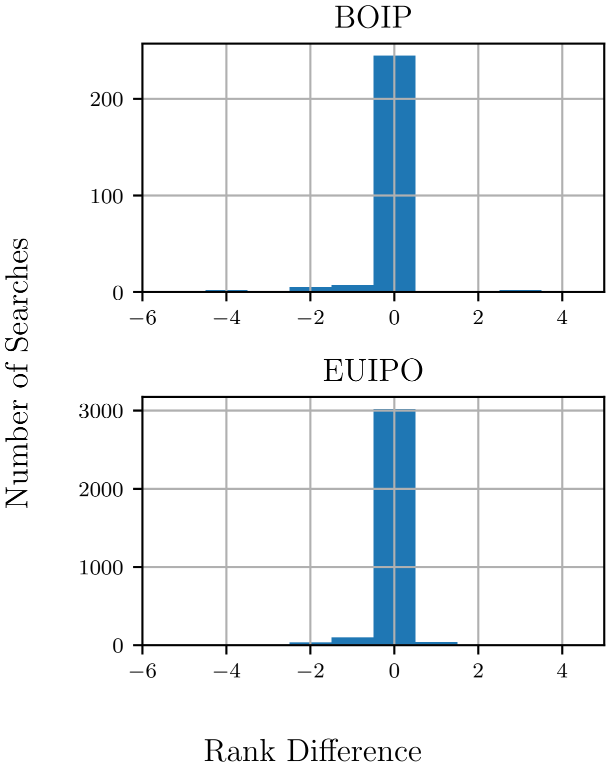


Figure 5: Distributions of rank differences. The difference (Rank by Holder - Rank by TM) is performed for each query where a match is found for both criteria.

System	LOC	Decisions	Matches	Ratio	$\overline{Rank}$
BOIP	+	2 597	312	12.01%	13.56
	+/-	1 065	65	6.10%	15.95
	-	1 122	64	5.70%	15.61
EUIPO	+	3 213	2 182	67.91%	46.06
	+/-	1 294	910	70.32%	50.11
	-	1 345	931	69.22%	60.81
EUIPO-50	+	3 213	1 989	61.90%	2.76
	+/-	1 294	828	63.99%	2.85
	-	1 345	809	60.15%	3.32

Table 5: Match ratios and mean ranks differentiated for the three possible decision outcomes.

4 CONCLUSION

We evaluated, for the first time, TM Search Engines using real-life decisions, at scale. The two search engines that we tested, the EUIPO's and the BOIP's, presented very different performances. The BOIP's system found the opponent TM a maximum of 11.72% of the possible queries, while the EUIPO found it for a minimum of 62.01% of its possible queries. Moreover, on those cases when the BOIP's system had a match, the mean rank of those matches was only 14.21. In the case of the EUIPO, the same mean rank was of

2.91. As a conclusion, the EUIPO's system is significantly more performant than BOIP's, for both metrics. However, we question the validity of the EUIPO's numbers, since the almost systematic presence of the match in second position. Our main hypothesis is that EUIPO's system was indeed trained on the very same set (or at least a major subset) that we used. Further research will need to verify this, for example by using another set of decisions.

We can consider, in light of the BOIP's performances, that the business case this system aims at solving is not met. According to our findings, in almost 88% (3 193 out of 3 617) of the business cases concerning the Benelux, if the applicant were to use the image search tool accessible on the BOIP website in order to identify conflicting earlier TMs, this applicant would not encounter at least one earlier TM giving rise to a LoC concern. This is even much of a concern when one considers that the earlier TM at hand was the most relevant, or amongst the most relevant ones, since in real life its proprietor actually opposed on that basis to the registration of the subsequent TM.

In general, both systems make errors. There is therefore no guarantee that they work at the level of requirements associated with the application. We claim that users of these tools should exercise caution when relying on them. This research highlights the necessity for transparency of the systems, along with independent evaluation. In a broader scope, we believe that these systems should be open, accessible and explainable, in order to be used in such contexts.

ACKNOWLEDGMENTS

The authors are grateful to the Fédération Wallonie-Bruxelles (Belgium) and the Université libre de Bruxelles who funded the research in the framework of the 'Actions de recherche concertées (2020-2025)', with a grant awarded to the IPSAM research project (Promotor: Julien Cabay; more information on the project : https://droit-prive.ulb.be/ipsam-adressing-intellectual-property-relevant-similarities-in-images-through-algorithmic-decision-systems/)

They are thankful to the Benelux Office for Intellectual Property (BOIP) for its support in this research. The authors affirm that they had complete freedom to conduct this study without any interference from the BOIP.

They also wish to thank all persons who shared comments on the preliminary findings of the research during the Artificial Intelligence Technology Policy Talk (organized by the Digital Law Center at University of Geneva on 4 November 2021), and the 4th Waseda Brussels Conference – New Technologies and Regulation in Japan and Europe (organized by the ULB and the Waseda University in Brussels on 13 September 2022), in particular: Jacques de Werra, Jean-Marc Deltorn, Dev S. Gangjee, Stéphane Marchand-Maillet, Anke Moerland, Alina Trapova, Tatsuhiro Ueno and Lonneke Van der Plas.

Thanks also to the students who helped us in different manners on this project, and to the JurisLab (Centre de droit privé, FabLab ULB), where the research was conducted.

REFERENCES

[1] Ilanah Fhima and Dev Gangjee. 2019. *The confusion test in European trade mark law* (first edition ed.). Oxford University Press, Oxford.

[2] Dev S. Gangjee. 2021. Eye, Robot: Artificial Intelligence and Trade Mark Registers. In *Transition and Coherence in Intellectual Property Law* (1 ed.), Niklas Bruun, Graeme B. Dinwoodie, Marianne Levin, and Ansgar Ohly (Eds.). Cambridge University Press, 174–190. https://doi.org/10.1017/9781108688529.020

[3] Dev S. Gangjee. 2022. A Quotidian Revolution: Artificial Intelligence and Trade Mark Law. *SSRN Electronic Journal* (2022). https://doi.org/10.2139/ssrn.4081317

[4] Sonia K. Katyal and Aniket Kesari. 2020. Trademark Search, Artificial Intelligence, and the Role of the Private Sector. (2020). https://doi.org/10.15779/Z380V89H87 Publisher: Berkeley Technology Law Journal.

[5] Daryl Lim. 2022. Trademark Confusion Revealed: An Empirical Analysis. *American University Law Review* 71 (2022), 1285–1365.

[6] Anke Moerland and Conrado Freitas. 2021. Artificial Intelligence and Trade Mark Assessment. In *Artificial Intelligence and Intellectual Property.* Oxford University Press, 266–291.

[7] Benelux Office of Intellectual Property. [n. d.]. BOIP and Darts-IP. https://www.boip.int/en/darts-ip

[8] European Union Intellectual Property Office. [n. d.]. New AI solution for image search (the AI solution is on the tool image search). https://euipo.europa.eu/ohimportal/en/web/guest/-/news/new-ai-solution-for-images-search

[9] European Union Intellectual Property Office. [n. d.]. Strategic Plan 2025. https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/contentPdfs/about_euipo/strategic_plan/SP2025_en.pdf

[10] European Union Intellectual Property Office. 2023. EUIPO Statistics for European Union Trade Marks, 1996-01 to 2022-12 Evolution. https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/contentPdfs/about_euipo/the_office/statistics-of-european-union-trade-marks_en.pdf

[11] Claudio A. Perez, Pablo A Estevez, Francisco J. Galdames, Daniel A. Schulz, Juan P. Perez, Diego Bastias, and Daniel R. Vilar. 2018. Trademark Image Retrieval Using a Combination of Deep Convolutional Neural Networks. In *2018 International Joint Conference on Neural Networks (IJCNN).* IEEE, Rio de Janeiro, 1–7. https://doi.org/10.1109/IJCNN.2018.8489045

[12] Charles V. Trappey, Amy J.C. Trappey, and Sam C.-C. Lin. 2020. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Advanced Engineering Informatics* 45 (Aug. 2020), 101120. https://doi.org/10.1016/j.aei.2020.101120

[13] Osman Tursun, Simon Denman, Sabesan Sivapalan, Sridha Sridharan, Clinton Fookes, and Sandra Mau. 2020. Component-based Attention for Large-scale Trademark Retrieval. *IEEE Transactions on Information Forensics and Security* (2020), 1–1. https://doi.org/10.1109/TIFS.2019.2959921 arXiv: 1811.02746.