The Role of Key Words and the Use of the Multilingual Eurovoc Thesaurus when Searching for Legal Regulations of the Republic of Croatia - Research Results

T. Didak Prekpalaj*
* Central State Office for the Development of the Digital Society, Zagreb, Croatia
tanja.didak-prekpalaj@rdd.gov.hr

Abstract - As legislation today is one of the largest collections of documents in electronic form, access to such documents should be publicly available by making legal regulations accessible to both legal experts and all interested citizens, in a transparent way. The most comprehensive, free, and permanent access to legislation is provided by the Central State Office for the Development of the Digital Society through the CADIAL search engine. The collection of legal regulations in electronic form contains legal regulations of the Republic of Croatia published in the electronic edition of the Official Gazette. All regulations can be searched by full text, and the collection is organized as a system of interrelated texts of regulations and information about them. Unique content analysis of texts and carefully selected key words that describe the document allow the user to quickly search for documents. The Eurovoc Thesaurus enables the content of documents to be unambiguously displayed in a documentation system, and their metadata to be created in high quality. It also ensures the unambiguity of user queries addressed to the documentation and information system.

The paper describes the results of a three-month user queries analysed through five case studies, with detailed processing of semantic queries by examining their correspondence with the words in the title, the words in the text of the regulation and the correspondence with Eurovoc descriptors. As part of the research, a survey of CADIAL search engine users' satisfaction was conducted.

Key words - search for legal regulations, Eurovoc Thesaurus, subject indexing, CADIAL search engine.

I. INTRODUCTION

In today's age of digital technologies, most developed countries publish their legislation in electronic form, and a very important part of the process of making legislation available is document retrieval. Searching for legal regulations, although more accessible than ever, also brings certain problems in communication with citizens searching for information from legal information systems. With the advent of the Semantic Web, legal information is processed according to its content, not just as plain text. It is the semantic gap that is one of the biggest obstacles in understanding between layman and legal experts - layman use different semantics and phrases when searching, while the computer uses the syntax and semantics of programming languages [1]. Semantic search is considered the most accurate since it gives the opportunity to contextualize online search. Also, given that there is an increasing amount of publicly available legislation, a good content analysis and carefully selected key words describing the document are necessary.

The most comprehensive, free, and permanent access to legislation in the last 20 years has been available through a collection of legal regulations on the website of the former Croatian Information and Documentation Referral Agency (HIDRA), and today on the website of the Central State Office for the Development of the Digital Society (hereinafter: SDURDD).

II. CREATING A COLLECTION OF LEGAL REGULATIONS

The Official Gazette is the main source for collecting legal regulations of the Republic of Croatia. It is the official newspaper of the Republic of Croatia where all legal regulations are published before they come into force. Documents published in the Official Gazette are followed and checked daily so that all new laws and other regulations are covered by the documentation processing.

The collection includes texts of legal regulations of the Republic of Croatia in electronic form, while complete texts of legal regulations with all amendments are available on the SDURDD website through the CADIAL search engine. In this way, online access to the entire valid legislation of the Republic of Croatia has been completed and enabled for the first time, including the possibility of reviewing the still valid regulations that were adopted before 1990 [2].

The collection of legal regulations is part of the Central Catalogue of Official Documents of the Republic of Croatia, which has a legal basis in the “Law on the Right to Access Information” from 2013. It is defined as “a publicly available tool on the Internet that provides users permanent access to documents stored in an electronic document database and/or physical collections, through full text and/or edited set of metadata”. Its purpose is to provide users permanent access to official
documents for information and reuse. Because it is available online, its content can be used regardless of the time and location of the user, which greatly facilitates the use of the right to access to information. According to the “Ordinance on the Central Catalogue of Official Documents of the Republic of Croatia” [3], documents published in the Official Gazette of the Republic of Croatia are considered duly submitted to the Central Catalogue.

All documents in the Central Catalogue are for information purposes only, including legal regulations, given that they are downloaded from the electronic version of the Official Gazette which is not considered as an official document in the Republic of Croatia. Only the printed edition of the Official Gazette is the official version of Croatian legal regulations.

This part of the Central Catalogue, based on the “Decision of the Government of the Republic of Croatia on the connection of the database of legal regulations of the Republic of Croatia to the N-Lex portal”, represents the legislation of the Republic of Croatia on the EU’s common gateway to National Law. [4].

III. DOCUMENTATION PROCESSING OF LEGAL REGULATIONS AND THE USE OF EUROVOC THESAURUS

Documentation processing can be described as the process of summarizing the topic of a document into words or expressions listed in a controlled dictionary (glossary, thesaurus), while descriptors are generated from the content of the document without going into details. The topic of the document is thus briefly presented, and the user finds the information either through a search engine based on a controlled dictionary or by navigating through the hierarchical structure of a controlled dictionary. By indexing, the documents are summarized into informative expressions contained in them, thus ensuring the retrieval of the requested document [5]. Without proper indexing, both search and retrieval are practically impossible.

The controlled multilingual Eurovoc Thesaurus was published in 1984. It was developed by the Publications Office of the European Union for the purpose of precise documentation processing of official documentation of the European Union. So far it has been published in 23 official languages of the European Union, and in three languages of the EU candidate countries.

During the translation and adaptation of the original thesaurus in Croatian, the selection, control, harmonization, and adaptation of terms was considered. It was agreed to use Croatian terminology in the translation, and that terms and nouns in the nominative are singular as in the original. For precise processing of official documentation of the Republic of Croatia, a National appendix to Eurovoc was developed, which consist of a structured directory of Croatian state bodies and bodies of local and regional self-governments, the Republic of Croatia diplomatic missions as well as foreign diplomatic missions in the Republic of Croatia, the list of political parties and geographic features. All records in the Croatian appendix are marked as Crovoc. Eurovoc originates from the need for uniform and efficient content processing, and over the years has developed into a documentation tool that helps to clearly define the content of documentation from all areas of EU activity. The thesaurus consists of 21 fields and 127 microthesauri in the form of a structured and controlled list of names used for all important terms that appear in different official documents [6]. It is important to mention that Eurovoc is used in the search of European Union legislation on the EUR-Lex portal.

Eurovoc is a strictly controlled thesaurus, hierarchically structured with explicit use of terms. It allows users and documentation experts to be precise in marking and finding information by looking at the semantic connections between systematic terms. It is a “living” thesaurus that is used daily for the processing of public official documents and is regularly updated, especially in the non-descriptor part. Also, it is revised in accordance with new texts and terminology which changes as a result of increasing number of published regulations. Its biggest advantages are unambiguous and consistent processing of document content, compliance of user queries with document descriptors, and a network of semantic links to find content-related documents.

IV. CADIAL SEARCH ENGINE - SEARCH POSSIBILITIES

In the collection of legal regulations, in terms of content all regulations are processed with the support of the software system for automatic indexing by Eurovoc descriptors (AIDE). The project “AIDE - automatic indexing using Eurovoc descriptors” was launched in 2004 and was taken as the basis for further work in the creation of the CADIAL search engine (Computer Aided Document Indexing for Accessing Legislation). A two-year project funded by the Flemish Government and the Government of the Republic of Croatia was launched in cooperation with the research team of the Katholieke Universiteit Leuven Interdisciplinaire Centrum voor Rechten en Informatica (ICRI), experts from the University of Zagreb and the European Commission's Joint Research Centre. The aim of this project was to provide the infrastructure for permanent and easy availability of Croatian legal regulations to promote the use of official documents [7]. CADIAL has been publicly available since the end of 2008 and provides users with permanent and easy access to all national legislation. It is based on a high degree of data integration as well as improved technical capabilities.

The CADIAL search engine [8] allows users to search the full text of a document, to search by document title and by using Eurovoc descriptors. It is preferable to search for documents by using all three options at the same time, since in that case the search includes all implicit content that encompass detailed content processing (Figure 1).
In addition to the text of the regulation, the search result provides a complete overview of a legal regulation thanks to the numerous data used to describe it through the document processing.

Search results are ranked according to relevance based on a combination of several factors. Each search result offered to the user contains the title of the document, the reference or source in which the document is contained, the relevant text clip, the descriptors used with a link to Eurovoc, the notes on entry into force and the links to related documents.

It is important to note that by using the software for morphological normalization of Croatian words, CADIAL has solved the problem of locating a word in all its forms. Namely, the reduction of different forms of words (gender, number, case...) to a unique form facilitates the search, especially if we consider the morphological complexity of Croatian language in relation to, for example, the English language.

Selecting one of the offered results opens the full text of the regulation and numerous relevant data and links related to the selected regulation. This means that regulations can be viewed directly, and that navigation flows from text to text without the need for a bibliographic description, while the context of the regulation document is given with other data and links in the header.

All other data and links are a result of the formal and content processing of documents from the collection and relate to the type (legal/bylaw) and the status of regulations (valid/not in force). They also show the descriptors assigned to the document with links to Eurovoc, as well as the classification according to the field of competence. In addition, each document is accompanied by specific notes, which are extremely important in the legal interpretation of regulations. Links to related documents (basic or influential regulations, amendments, consolidated texts and corrections, and by-laws) and links to texts of the European Union legislation, which Croatian regulations are harmonized with, make an integral part of the search results [9]. Data of the regulations and links to the original document can be found below (Figure 2).

V. OBJECTIVES AND DESCRIPTION OF THE RESEARCH

One of the main objectives of the research\(^1\) was to identify and analyse the role of key words in searching for a specific search engine. At a time when most search engines support full-text search and when the user expects the required information in the first few search results, the focus of interest has been on examining the use of key words from a specialized thesaurus such as the Eurovoc Thesaurus.

The aim was to analyse the users - who they are, what they search for and how they search. When the idea for this type of research was created, we were working on the development of a search engine (primarily the development of an automatic indexing programme), but without a real insight into the information needs of users. Predefined search options were not questioned, although there was doubt about the actual use of searching by using Eurovoc key words.

A. Analysis of user queries through case studies

To analyse the habits of CADIAL search engine users, and to get a better insight into the existing search capabilities, as well as the role of the Eurovoc Thesaurus, user records in the form of logs collected from CADIAL servers were analysed. Logging began on June 1, 2017 and continued until August 31, 2017. The requested records were delivered in .txt file and raw format, unstructured and non-readable for analysis (with over two million queries). With the help of colleagues from the Faculty of Electrical Engineering and Computing (Text Analysis and Knowledge Engineering Lab), the data was structured and compressed into a .CSV file, which contained 16,804 user queries. The number of queries in the final file is much lower than the results originally obtained, because of unrelated requests that were recorded with an error.

The goal was to analyse unique sessions - all actions that a user took after typing a search term into a search engine. It was desirable to see information about the

\(^1\) The research was conducted as part of the author's doctoral dissertation, entitled “Model of a system for making legal regulations interconnected, accessible and preserved in a long-term”.

---

1. MIPRO 2021/ICTLAW

1691
search option selected by the user and information on whether the user selected the regulation after the search.

B. Survey on user satisfaction with the CADIAL search engine

Although the search engine has been publicly available since the end of 2008, and as a unique search engine provides the user easy and free access to the complex structure of national legislation, no user satisfaction survey has been conducted in the last 10 years. The search engine did not change significantly during this period, and since there was no feedback from users, there was only a presumption how easy the search engine is to use, whether the instructions on search methods help, and whether users use all the advanced search options offered by the search engine.

The aim of the survey was to examine the users' satisfaction with the search engine and most importantly, to learn whether the user found the regulation he was looking for. Since all specific user queries were collected from the logs in the same period, the survey sought to obtain final confirmation about the simplicity of search, possible suggestions, and comments in cases where the user rated the search engine as complicated to use, and the user's intention to reuse the search engine in the future.

The survey was conducted from July 3 to September 5, 2017. The survey was placed on a search engine page on demand, as a pop-up survey, for every third user on the webpage - so it opened for 33% of users. The answers of the completed surveys were delivered in a .CSV file from which the individual answers were analysed for the purpose of the research.

The survey consisted of only five questions and three minutes was enough time to complete it. In addition to answering the questions, the IP addresses of the users were also collected, as well as the exact time of completing the survey. This was necessary to be able to combine these two data sets with identical data obtained from search engine logs.

In the above-mentioned period, 875 user responses to the survey were received. However, after eliminating all spam responses that had no answers other than comments, 862 unique polls remained. Compared to the total number of users who searched for legal regulations in the same period (24,497 users who started at least one session during the specified date range) and given that the survey was conducted by every third user, we can conclude that 10.5% of users responded to the survey.

VI. RESEARCH RESULTS

Out of a total of 16,804 collected queries, 534 were not suitable for analysis (the query did not correspond to the selected regulation); therefore 16,270 user queries were included into further analysis. When looking into the search options used by the users, it is noticeable that most of them type the search term into the search box and do not change the predefined search options - by title and text. Search by Eurovoc (key words) must be selected, as well as search of valid regulations. Predefined text search options were used in 96.63% of queries and title search in 99.60% of queries.

It is important to note that only 1,508 queries out of the above-mentioned 16,270 (9.27%) had the Eurovoc search option included (key words). I believe that the reason for such a low number of queries with the selected option is partly because the search engine interface is predefined with the already included search options by text and by title.

The analysis has shown that only half of the users choose to use advanced search options to limit their results to, for example, the type of regulation, area of activity or EU chapter.

The next step in the analysis was to select several case studies to better identify individual query types. Thus, 16,270 user queries were included in the case study examination. Instead of using a computer software, each query was manually rated and categorized as a specific case study. By reviewing all collected records, five case studies were established:

- #1 – the user has entered the correct title of the regulation
- #2 – the user entered the correct title of the regulation and the number of the Official Gazette
- #3 – the user has entered part of the title of the regulation
- #4 – the user has entered the terms for which he expects the required regulation
- #5 – the user entered the direct identifier - the number of the Official Gazette.

The analysis of all collected queries demonstrated that the highest number of user queries in the case study was #1, which means that most users knew the exact title of the document they were searching for (Figure 3). The user entered the correct title of the required regulation in 9,440 queries, i.e., over 58% of all queries set as the entry of the correct title of the regulation. Queries were considered correct regardless of the cases used, while a query written in capital letters or in quotation marks was also considered correct.

![Figure 3. Display of the number of queries in all case studies](image-url)
A very low percentage of queries (only 11.78%) were related to semantic search. Case study #4 includes queries that are enclosed in quotation marks as a search term (e.g. the user enters the phrase subordinate creditors for which the required regulation is expected – “Ordinance on the capital adequacy of a pension insurance company”) and a semantic search in which the user enters terms for which he expects a relevant document (e.g. the user enters the terms precious metal for the required regulation “Ordinance on technical conditions, which, with regard to construction, must be met by articles made of precious metals”).

Further analysis considered only semantic queries (1,917 of the total number of queries). After eliminating the same or repeated queries, a total of 743 different queries were selected and then their match with the offered search options was analysed (of which the search options by text and title are pre-defined in the search engine while Eurovoc search should be enabled by the user).

Each user query was examined in the Eurovoc Thesaurus search engine in the form of singular nominatives since Eurovoc mainly contains singular terms. Thus, the query boats were found in Eurovoc, but in the form of the word boat which was evaluated as a matching word.

The total match of the search terms with the key words from Eurovoc is 218 out of 743 user queries. This means that only 29.34% of all semantic queries matched the Eurovoc descriptors.

Each user query was entered into the CADIAL search engine and then terms were searched through the text. 665 queries were found in the total number of the mentioned case study (89.50%).

Each user query was also searched by selecting on the option “Search in the title”. The match with the requested query was evaluated positively if over 50% of the words in the query corresponded to the words from the title of the text. Also, the singular and the plural in the required terms were neglected. The analysis showed that 56.12% of user queries matched the words in the document title.

It was interesting to observe which regulations users searched for the most and which queries they entered. The analysis showed that the user will search according to the words from the title of the regulation in case of a short title, e.g., “Building Act”. For regulations that have longer titles, they will search more often for parts of the words in the title or will enter terms which they expect to find in the text of the regulations. For example, the “Ordinance on fire resistance and other requirements that buildings must meet in case of fire” will more often search with queries such as: evacuation provisions, fire resistance, evacuation routes, parapet, and similar.

The user satisfaction survey with the CADIAL search engine showed interesting results. The first, and the most important question in the survey was the following: “Did you find the regulation you were looking for?”. As many as 782 out of 862 users answered “Yes”. Thus, the regulation was found by 90.72% of users who participated in the survey.

In the question that followed, the ease of use of the search engine was examined. The question was: “Please rate whether you find the search engine easy to use” and two answers were offered: “It’s simple to use”, and “It’s not simple to use”. A total of 789 out of 862 users rated the search engine as simple, i.e., 91.53% of users. Users who rated the search engine as “not simple to use” were offered a text box in which they could enter their comment. A total of 52 user comments were entered, some of which rated the search engine as simple but had a comment or suggestion. Analysis of all user comments showed that users have remarks and comments to improve search capabilities (they wish for consolidated texts or chronological search), filter results (suggestion to include period of the regulation), the ranking of the required regulation and the display options (the most common complaint is that the search engine is not adjusted for browsing on mobile devices). User comments (Table 1) can serve as a good indicator of the user experience and point out to possible improvements to make the user experience better.

The third question was related to the use of key word search functionality from Eurovoc, and the user could choose one of the three answers provided in the survey. The question was: “Did you use the option to search/filter results using key words from the Eurovoc Thesaurus in your search?”. Answers were offered: “Yes, key word search was helpful in searching/filtering results”, “No, key word search did not result in the required regulation” and “I did not use the Eurovoc Thesaurus when searching”. The highest number of users (443) did not use the possibility to search by using Eurovoc, which is 51.39% of all user responses. A total of 361 users (41.89%) answered that the key word search was helpful, and 57 users (6.62%) answered that the key word search did not result in obtaining the required regulation.

<table>
<thead>
<tr>
<th>Jeste li pretraživale/ali je trudili se pretražiti izraz ili frazic?</th>
<th>Jednostavnost korisničkog tržišta</th>
<th>Ako je vaš odgovor “Nije jednostavan” među opono svižnih reči</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA Jednostavan je</td>
<td>Prijelaz je između mogućnosti postavljanja pretraživih teksta posredovanja</td>
<td></td>
</tr>
<tr>
<td>DA Nije jednostavan</td>
<td>Jekoz.dozdaš dobli kad bi se pogledalo pripazovano</td>
<td></td>
</tr>
<tr>
<td>DA Nije jednostavan</td>
<td>Tržišča je treba jednostavnije dobi za lepšek se izgoobja v samostveno zmanjujanje diplomatiki. Na drugem delu je dnevno več posnova, ki bi bilo to odliko do se nepripravi verjame na posodobljene tečajne se vnapreku uvede in iziskovano zakonje.</td>
<td></td>
</tr>
<tr>
<td>NE Nije jednostavan</td>
<td>ne dre odnasovanih poljov</td>
<td></td>
</tr>
<tr>
<td>DA Jednostavan je</td>
<td>Jekoz.dozdaš dobli od poljov posodobljenih sklepa.</td>
<td></td>
</tr>
<tr>
<td>NE Nije jednostavan</td>
<td>od dobli vsteje poljovski sklepa, ki.</td>
<td></td>
</tr>
<tr>
<td>DA Jednostavan je</td>
<td>Zeleni oznake poljov za tržišča CADIAL kompilirati kvalitetne oznake, ki bi bile odvisne od kontrole v ose.</td>
<td></td>
</tr>
<tr>
<td>NE Nije jednostavan</td>
<td>Teka štanda poljovskih oznak, tako pa je poljovskih oznak.</td>
<td></td>
</tr>
<tr>
<td>DA Nije jednostavan</td>
<td>Priljubljena uporba tehnične znamke. Za znam.</td>
<td></td>
</tr>
<tr>
<td>DA Nije jednostavan</td>
<td>Hle iznura do oznaka izdelka/kibernetc iz delovnih zakonoch. Odzivalo se na sej o pripazovanjih, ki so posledica. Oblačna je oznaka da se prilagodi znamke izdelka.</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. User Comments on Simplicity to Use the Search Engine
The following question was about the user's workplace. The aim was to determine the environment in which users use the search engine. The answers included: civil service, public service, private sector or other. Although the assumption was that most of the users come from the state or public service since they use legal regulations daily, the survey showed that the private sector leads in the use of the search engine - a total of 53% of all search engine users.

The last question included the feedback of users who had previously used the search engine. It is an important indicator of user experience because it directly asks the users if they will use the search engine again - under the assumption of a positive answer if they have used the search engine successfully and if they have found the regulation they were searching for. The results show that as many as 94.77% of users plan to use the search engine again (817 survey users).

VII. CONCLUSION

The Eurovoc Thesaurus is undeniably a useful tool in searching for documentation, but it is questionable whether users of various information needs understand it and use it in practice. This was tested by analysing user queries. The analysis was extensive and included 16,270 queries which helped determine the user's behaviour through unique sessions - including all actions that the user took after typing the search term into the search engine. Selected search options were analysed, and each query entered the search engine was evaluated individually through different case studies, as well as the evaluation of search success that was defined as the selection of a document after entering a search term.

Unfortunately, only 9,27% of users use the Eurovoc search option, while even significantly fewer, only 0,28% users, decide to limit their search to the valid regulations in the search engine. I assume that the results would be different if all search options were immediately visible (both basic and advanced), and if the search included more "overview menus" - e.g., search by body or author, field of activity, consolidated texts, etc. The analysis which focused on semantic queries only, examined their correspondence with all basic search options. The compatibility of such user queries was highest with words in the text and lowest with Eurovoc descriptors. This was expected given the low usage of Eurovoc.

The CADIAL user satisfaction survey showed excellent results, although they may not be completely in line with the results shown by their actual search queries (it should be noted that the queries were analysed in the same period as the search engine survey). More than 90% of users answered that they found the required regulation, and as many as 94.77% of users plan to use the search engine again.

The results of the research show that searching for legislation, which is very specific, it is necessary to adapt it to the user. Along with better presentation of the regulations themselves, it is necessary to improve the existing search engine which will be able to respond to the various information needs of end users. This can be achieved through further use of artificial intelligence in morphological word processing and ranking results, so that the term entered in the search engine automatically recognizes the user's information need and retrieves the required document. Thus, the CADIAL search engine would fulfil its goal as it would enable the average user who is not versed in legal terminology, to access the requested document efficiently, quickly and without any cost.

Since 2018, SDURDD in cooperation with the Faculty of Electrical Engineering and Computing (Text Analysis and Knowledge Engineering Lab) have been working together on improving the semantic search engine as a recognizable Croatian product through an EU funded project “Establishment of an integral system for the management of the official documentation of the Republic of Croatia”. Results of the project are expected during the next two years.

REFERENCES

[8] CADIAL search engine https://sredisnjikatalogrh.gov/hr/