CEUR-WS.org/Vol-3697/paper4.pdf

Extracting licence information from web resources with a Large Language Model*

Enrico Daga[1], Jason Carvalho[1] and Alba Morales Tirado[1]

[1]The Open University, Milton Keynes (United Kingdom)

Abstract

Data catalogues play an increasing role in supporting information sharing and reuse on the Web. However, evaluating the reusability of Web resources requires an understanding of the related licence and terms of use. Recent methods for licence representation and reasoning allows to explore Web resources according to their permissions, obligations, and duties. Therefore, licence annotations should be linked to those representations in order to support users in filtering and exploring datasets according to their licencing requirements. However, populating data catalogues with licence information is a tedious and error-prone task. In this paper, we explore the suitability of a Large Language Model (LLM) to support the automatic extraction, annotation, and linking of licence information from reference Web pages of data catalogue items. The approach is evaluated for its capacity to automatically find relevant pages from within a main web page, extract data about copyright and licencing, and link licence descriptions to a knowledge graph of licences expressed in RDF/ODRL. We apply our method to extend the coverage of licence annotations of a data catalogue in the music domain.

Keywords

Licence Extraction, Knowledge Graphs, Open Digital Rights Language, Large Language Models,

1. Introduction

Data catalogues play an increasing role in supporting information sharing and reuse on the Web in many domains. However, evaluating the reusability of Web resources requires an understanding of the licence and terms of use associated with those resources. Recent methods for licence representation and reasoning allow to explore Web resources according to their permissions, obligations, and duties [1, 2]. However, licence annotations should be linked to those representations in order to support users in filtering and exploring datasets according to their licensing requirements [3, 4].

Our starting point is a registry of resources relevant to music research: the musoW catalogue [5]. MusoW is a knowledge graph and Web registry of datasets and projects annotated with crowd-sourced metadata. We analysed the coverage of licence annotations from querying the musoW SPARQL endpoint. Figure 1 shows how most resources do not have a specific licence associated with them (almost ~70% of the registry). The reasons may vary: 1. the resource does not have a specific licence; 2. the information being not available at the time the metadata

Workshop

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

enrico.daga@open.ac.uk (E. Daga); jason.carvalho@open.ac.uk (J. Carvalho); alba.morales-tirado@open.ac.uk (A. M. Tirado)

0000-0002-3184-5407 (E. Daga); 0009-0003-3458-9845 (J. Carvalho); 0000-0001-6984-5122 (A. M. Tirado)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

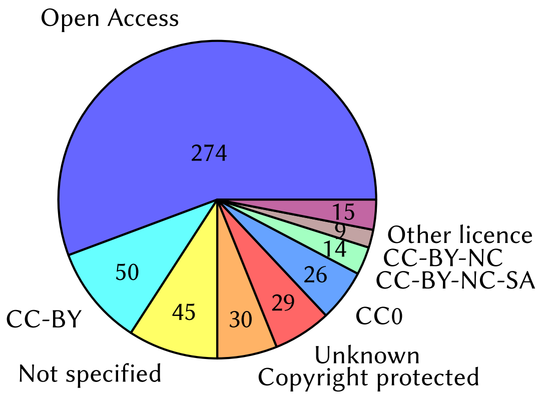

Figure 1: Summary of licence annotations in the musoW dataset. Most items are labelled as generically open access. In other cases, licences are unknown or copyright is protected with no licence.

was curated (but it may be available today); 3. the curator overlooked the information, maybe because it was hidden in secondary web pages.

The lack of sufficient licensing and terms of use information for published web resources is a well-known problem whose impact on the broader landscape of content reuse on the web cannot be underestimated. In the case of musoW catalogue, we are confident that most of the annotations are actually correct (or they were correct at the time of their retrieval). However, supporting curators in collecting such information without having to browse each one of the websites catalogued manually would certainly contribute to improving the quality and coverage of the musoW catalogue.

Large Language Models (LLM) such as OpenAI's ChatGPT, Meta's LLAMA and Google's Bard, have emerged recently providing impressive abilities in language generation, opening new opportunities for interacting with textual content, for example, for detecting and extracting structured information [6]. In this paper, we apply a Large Language Model (LLM) for extracting licence information from web resources to improve the coverage of licence metadata in Web registries, considering the case of the musoW registry. Specifically, we pose the following questions:

RQ1 Can copyright and licence information be derived automatically from web pages?

RQ2 How can copyright and licence information be derived automatically from web pages using Large Language Models (LLM)?

RQ3 How accurate would an LLM detect the copyright and licence information (in other words, is it worth pursuing this line of enquiry)?

RQ4 How much can we complete a curated catalogue of licence metadata with an automatic method based on LLMs?

The rest of the paper is structured as follows. The next Section is dedicated to related work. Next, we illustrate our methodology to apply a Large Language Model to extract and link licence information from Web resources (Section 3). We report on extensive experiments in Section 4, before discussing our results (Section 5) and concluding the work in Section 6.

2. Related Work

Licences and terms of use have been a recurrent topic of interest in Web research. Initiatives include The Creative Commons Rights Expression Language[1] proposed by the Open Data Institute, and the Open Digital Rights Language (ODRL) a W3C specification to support the definition, exchange and validation of policies [7]. Online repositories are developed to publish licences expressed in RDF, including the RDFLicense Dataset[3] [8], and DALICC [2], which we use in our work. A formal representation of licences can be of use to support the users in deciding what possible constraints they want to guarantee concerning the use of their data [9, 4]. Datasets should include information on licences facilitating researchers' decisions to reuse such resources [10]. Computational legal policies allows to reason on the applicability of terms to data derived from licenced resources [11]. The collection and curation of licence metadata is clearly a necessary step for enabling such applications. Applying natural language processing techniques, like the ones proposed in [12], can facilitate the process of data acquisition. Recently, there has been increasing work on applying Large Language Models (LLM) to aid the extraction of structured information from textual content (e.g. [13, 14, 15]). Differently from fine-tuning (e.g. in RAG [16]), in-context learning allows for tailoring the response flexibly without significant computational resources [17]. Emergent research is exploring complex tasks such as interpret the content of web pages and navigate[4] [18]. Attention has been given on evaluating the suitability of LLMs in many end-user tasks as well as raising concerns on their limitations, for example, in generating plausible but wrong information (hallucination) and propagating societal biases derived from the text they have been trained from [19]. Knowledge graphs play a key role in bridging the gap between language models and structured data models [20], including attempts to mitigate known issues in content generated by LLMs such as hallucinations and biases [21]. Similarly, LLMs are at the centre of current effort in aiding knowledge graph population in various domains [22]. In our work, we use LLMs with in-context learning to identify copyright and licence metadata from web resources, and develop a knowledge extraction pipeline that generates links between two knowledge graphs: one of catalogue items and the other of licences represented computationally.

3. Methodology

We tackle the problem by designing a methodology that engages with an LLM by asking to perform language understanding tasks. To avoid relying on the LLM embedded knowledge (which is known to be incomplete and often leads to unreliable information due to hallucinations), we design specific prompts (in-context learning) that make use of its language processing/predictive abilities but constrain them only to content that we provide. The method is structured as follows:

Data preparation We start from a list of resources published on the Web for which we want to know the associated licence. The assumption is that for each resource there is a web page which includes such information in one of the linked pages.

[1] Creative Commons rights language, the *ODRS* vocabulary[2]https://creativecommons.org/ns

[3] RDFLicense Dataset, https://rdflicense.linkeddata.es/

[4] See also tools such as vimGPT https://github.com/ishan0102/vimGPT and the BrowserPilot extension of ChatGPT https://community.openai.com/t/browserpilot-a-plugin-for-enhanced-chatgpt-interactions/297653.

Task 1 :: identify Starting from the main web page of the catalogued resource, we design a prompt for an LLM asking to find no more than three links that may include copyright, privacy, or licencing information. The resource's home page is downloaded, and all HTML tags except anchor tags (links) are removed. This step is necessary for reducing the content size, making it less expensive to be analysed by the LLM. We ask the LLM to find such information in the content provided. The expected output is a list of links potentially including copyright and licence information.
Task 2 :: extract We design a prompt for an LLM asking to derive copyright, licence, and terms of use information from a piece of textual content. For each one of the resources and links collected, we download the HTML page and remove all tags. We then send the content to the LLM, which is asked to return a structured data object with three main fields: copyright, licence, and terms of use.
Task 3 :: link In this step, we focus on linking licencing information to a catalogue of well-known licences. We designed a prompt for an LLM asking to identify a licence from a piece of text, selecting it from a list provided.
Evaluation We evaluate each one of the previous steps under a number of dimensions, including 1. the ability of the LLM to provide an answer syntactically correct (following the requested specification); 2. the ability of the LLM to make an answer semantically correct (a meaningful answer); Each task involving the LLM included a prompt engineering design phase which was essentially exploratory, starting from a prompt-as-hypothesis and resulting in a final prompt, after a short time of incremental trials with the LLM UI dashboard. In what follows, we apply the methodology to the collection of resources published in the musoW catalogue that do not have a specified licence and describe our approach in detail.

4. Experiments

In this section, we report on the experiments conducted by applying the methodology outlined so far to the musoW catalogue of resources that do not have an explicit licence in the metadata. The experiments were executed using OpenAI ChatGPT API with model `gpt-3.5-turbo-16k`. The experiments are reproducible with the source code provided in this GitHub project: https://github.com/polifonia-project/musow-licences-experiments-llm.

4.1. Data preparation

We use two main resources: 1. the musoW catalogue of musical resources on the Web [5] 2. the DALICC catalogue of licences in RDF/ODRL [2]. We start by downloading the content from the musoW SPARQL endpoint[5], specifically the resource identifier and name, the main home page of the resource, some categorical data and the licence metadata[6]. Next, we obtain the list of DALICC licences and generate a file summarising the licence description, legal text URL, and code used as a local name to identify the Linked Data entity[7].

[5] musoW endpoint: https://projects.dharc.unibo.it/musow/sparql

[6] The data file can be inspected at https://github.com/polifonia-project/musow-licences-experiments-llm/blob/main/Query-16.csv.

[7] The YAML file can be found in the experiments project folder on GitHub: https://github.com/polifonia-project/musow-licences-experiments-llm/blob/main/licences.yaml

Table 1
Example of results from Task 1. Relevance is established as follows: 1 - definitely irrelevant; 2 - probably irrelevant; 3 - cannot decide; 4 - probably relevant; 5 - definitely relevant.

web page	links found	rel.
http://www.transforming-musicology.org/	https://transforming-musicology.org/ https://www.themercurialmagpie.com/ https://plowns.com/	1
http://www.bruckner-online.at/	http://www.bruckner-online.at/?page_id=604	2
http://www.tudorpartbooks.ac.uk	http://www.ncl.ac.uk http://www.tudorpartbooks.ac.uk/newsevents/ http://www.tudorpartbooks.ac.uk/outputs/	3
https://www.metal-archives.com/	https://www.metal-archives.com/content/help	4
http://www.musicatradicional.eu	http://www.musicatradicional.eu/contact	4
http://www.sjsu.edu/beethoven/	http://www.sjsu.edu/privacy/index.html http://www.sjsu.edu/titleix/ http://www.sjsu.edu/siteindex/	5

4.2. Task 1: finding links in web pages

The first task aims to automatically retrieve links pointing to web pages potentially including information about copyright, licence, and terms and conditions.

Prompt engineering. We start with the following prompt as an initial hypothesis:

```
SYSTEM: You are an expert in licencing and terms and conditions of resources on the Web
    .
USER: Find the link to the pages describing licences , privacy policies , or terms of use
      of the content in the following HTML source code. Please respond in a JSON format
     . HTML code: {{HTMLCODE}}
```

We perform tests with sample web pages from the musoW catalogue and change the prompt to include more details regarding the expected format and strengthen the reference to HTML knowledge. The resulting prompt is the following:

```
SYSTEM: You are expert in licencing and terms and conditions of resources on the Web.
    You also know how to find information on a web page by reading its HTML content.
USER: Find the link to the pages describing licences , privacy policies , or terms of use
      of the content in the following HTML source code. Please respond ONLY with a JSON
      format with a list of maximum 3 links , resolved according to this address: {url}
    HTML code: {html}
```

We iterate over the list of resources without explicit licence information (or marked with any of the categories that do not refer to a specific licence, as discussed in Section 1). The answers are saved locally and collected into a table that we later analyse to evaluate the performance of the LLM under the two dimensions mentioned in our methodology, which we specify as follows: Q1 Are there any links returned? (Yes/No) Q2 Is the returned well-formed JSON? (Yes/No) Q3* Are any of those links relevant? We evaluate the answer on a Likert scale, from definitely not (1) to surely yes (5). While the first two questions can be answered automatically, we rely on manual supervision to answer the third one (we indicate this with the asterisk). It needs to be duly noted that we did not manually check each one of the web pages but only observed the returned links and assessed whether any of them may potentially provide useful information. A sample of the results of this task can be seen in Table 1.

4.3. Task 2: extract copyright, licence, and terms of use

The output of the previous step is a set of links for each one of the resources derived from the content of the home web page. The second task aims to extract information from each one of those web pages. We used all links returned, independently from our manual relevance assessment (270 resources and 648 links in total).

For this task, we want the information to be structured under three dimensions: copyright statement – who owns the intellectual property of the resource; licence – what is the licence associated with it (if any); and terms of use – to include any other information regarding the use of the resource.

Prompt engineering. We start with the following prompt as an initial hypothesis:

```
SYSTEM: You are expert in licencing and terms and conditions of resources on the Web.
    You also know how to find information on a web page by reading its HTML content.
USER: Please list the licences and copyright owners named in the following HTML code.
    Format the answer in JSON with two fields, ’copyright’ and ’licences’. {{HTMLCODEE
    }}
```

We perform tests with a sample of content from the web pages of the previous step and refine the prompt until we obtain sufficiently consistent results. The resulting prompt is the following:

```
SYSTEM: You are expert in licencing and terms and conditions of resources on the Web.
    You also know how to find information on a web page by reading its HTML content
    and express it in JSON format.
USER: Please list the licences, copyright owners, and terms and conditions mentioned in
     the following text. Respond only with a JSON object with 3 fields, ’copyright’, ’
    licences’, and ’terms and conditions’. The text is: {text}
```

We save the responses locally and gather them in a tabular format. Subsequently, we analyze this data to assess the effectiveness of the LLM based on syntactic and semantic accuracy, as stated in our methodology. These dimensions are elaborated as follows: Q4 Is the text returned well-formed JSON? Q5, Q8 Did the LLM find any copyright information? Q6, Q9 Did the LLM find any licence information? Q7, Q10 Did the LLM find any terms and condition information? We pose the last three questions above two times, the first considering each of the links (web pages) and associated requests to the LLM and the second aggregating all responses related to each resource and quantifying whether any provided links was useful to gather the information. While the above questions can be answered automatically, we add a qualitative, human-based assessment of the quality of the results, answering the following additional questions on a restricted sample of 100 items: Q11* Is the copyright information correct? Q12* Is the licence information correct?

Tables 2 and 3 show example annotations for questions Q11 and Q12 respectively. Crucially, we observed that for all 100 evaluated responses to Q12, the LLM never returned a wrong answer, while having some variability in the form (for example, in some cases it did not find a licence but it still returned some content). We leave the assessment of the information related to the terms and conditions to future work.

4.4. Task 3: link licence descriptions to the licences database

The expected output of the previous step is a structured JSON object with three fields: copyright, licence, and terms of use. In this task, we focus on the content returned for the field ’licence’

Table 2
Example annotations for evaluating Q11: 0 - No copyright; 1 - Somehow correct; 2 - Surely correct.

Web page	Copyright	Ann.
https://github.com/midi-ld/documentation/issues	2023 GitHub, Inc.	0
http://popmusic.mtsu.edu/ManuscriptMusic/guidelines.aspx	Copyright © 2023 – All Rights Reserved	1
https://www.uni-regensburg.de/impressum/index.html	Universität Regensburg	1
https://www.youtube.com/t/privacy	Google LLC	2
https://aaamc.indiana.edu/whats-going-on/aaamc-speaks/index.php	The Trustees of Indiana University	2

Table 3
Example annotations for evaluating Q12. The **F?** column evaluates whether the LLM found a licence in the content: 0 - Not found; 1 - Found but incorrect; 2 - Found and correct. The **Inc?** column evaluates if the response was not correct.

Web page	Licence	F?	Inc?
http://popmusic.mtsu.edu/ManuscriptMusic/guidelines.aspx	[]	0	0
https://www.youtube.com/t/terms	worldwide, non-exclusive, royalty-free, transferable, sublicensable licence to use that Content	0	0
https://github.com/midi-ld/	['MIT']	2	0
http://cantus.uwaterloo.ca	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License	2	0
http://pemdatabase.eu/	['Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)']	2	0
https://libraries.mit.edu/permissions	['CC BY-NC']	2	0

and aim to automatically link such licence descriptions with the equivalent authoritative entry derived from the Dalicc catalogue of licences expressed in RDF/ODRL. The initial prompt hypothesis is the following:

```
SYSTEM: You are expert in licencing and terms and conditions of resources on the Web.
    You also know how to find information on a web page by reading its HTML content.
    You are also proficient in reading YAML files.
USER: Given the following list of licences, can you tell me to which licence the
    following description refers to {LICENCEEXPR} {YAML}
```

We refined the prompt by testing a sample of licence descriptions identified in the previous step to improve the results. Specifically, we moved the list of licences to the SYSTEM input and asked to return 'NONE' if the description would not refer to any specific licence in the list.

```
SYSTEM: You are expert in licencing and terms and conditions of resources on the Web
    and know the following list of licences:{listOfLicences}
USER: Can you tell me to which licences the following licence description refers to?
    The description is  {description} -- Please respond by only reporting the selected
     licences from the list or 'NONE' if none is found.
```

We manually evaluate each one of the responses and annotate them as follows: (-1) The licence described is in the list, but the LLM didn't find it (or it hallucinated in some way); (0) The licence described is not in the list and the LLM correctly did not find it; (1) The licence description found is correct and in the list but the LLM did not link it properly (for example, it did not respond with the licence code); (2) The licence was found in the list and linked properly (the correct licence code was returned). In this assessment, we ignore the licence version and accept to link, for example, a CC licence Version 1 to the equivalent Version 4 in the DALICC

Table 4
Examples from the results of Task 3: -1 - The licence is in the list, the LLM didn’t find it; 0 - The licence is not in the list and the LLM did not find it; 1 - The licence found is correct but the LLM did not link it properly; 2 - The licence was found and linked.

’licence’ description from Task 2	answer of the LLM to Task 3	ann.
[”[{’licence’: ’Library of Congress’}, {’licence’: ’ReadSpeaker’}, {’licence’: ’Facebook’}, {’licence’: ’Twitter’}, {’licence’: ’Bachtrack’}, {’licence’: ’WETA Classical Conversations’}, {’licence’: ’National Public Radio’}]”]	The licences corresponding to the descriptions are:- Library of Congress: [CC0 1.0 Universal]- ReadSpeaker: [CC-BY-ND_v4]- Facebook: [CC-BY_v4]- Twitter: [CC-BY_v4]- Bachtrack: [CC-BY_NC-ND_v4]- WETA Classical Conversations: [CC-BY_NC-ND_v4]- National Public Radio: [CC-BY_v4]	-1
[”{’ELVIS Database Source’: ’https://github.com/ELVIS-Project/elvis-database/releases’}”]	NONE	0
[”[’MIT’]”]	The description ’MIT’ corresponds to the [ExpatLicense] MIT License (MIT) https://opensource.org/licenses/MIT	1
[”[{’title’: ’CC BY-NC 4.0’, ’website’: ’https://creativecommons.org/licenses/by-nc/4.0/’}]”]	[CC-BY-NC_v4] Creative Commons Attribution-NonCommercial 4.0 International (Creative Commons) https://creativecommons.org/licenses/by-nc/4.0/legalcode	2
[”[{’name’: ’Creative Commons Attribution License’, ’url’: ’http://creativecommons.org/licenses/by/1.0/’}]”]	CC-BY_v4	2

catalogue[8]. Table 4 shows a sample of annotated responses, limited to the content of Task 2 returning the field ’licence’.

We quantitatively evaluate this task as follows: Q13* How many correct decisions are made? (all except -1) Q14* How many licences are correctly not found? (0) Q15* How many licences are correctly found? (1 and 2) Q16* How many licences are linked to the list? (2) Q13 includes all answers, positive and negative, while Q14 summarise the licences that were missing from the sources and therefore not linked. Q15 counts the licences that where found and correct, even if the linking task didn’t work syntactically, while Q16 only measures the licences that were correct and properly linked to the list provided.

5. Results and discussion

In this section we present the results of our experiments and discuss them in the light of the initial assumptions and hypotheses. A summary of the results is reported in Table 5. The results of each one of the steps, the related manual annotations, and the computed statistics can be reviewed at this address: https://docs.google.com/spreadsheets/d/1wl-5YKcLVY9wDwSauPWz9NlLyeI7Ga1Da5WJXtOOp18/edit?usp=sharing.

We can first look at the results of each one of the tasks, in order to gather evidence that would allow us to answer the main questions.

Task 1 The first task is related to finding links in web pages that may include copyright or licence information. The task was executed 313 times, one for each resource home page. The vast majority of results were provided with a correct JSON syntax (this includes responses with no links). The LLM was capable of finding links in 86% of the cases, and most of the link sets are

[8] The registry typically includes only the most recent version of a licence. However, we leave the assessment of licence versions to future work.

Table 5
Summary of the results. For Task 2 some questions are repeated, the first considering each one of the links (web pages) from the previous step and the second aggregating all responses related to each resource (marked with R).

Qn	Task	Question	True	Maybe	False	Total	% True
1	T1	Are there any links returned?	270	0	43	313	86%
2	T1	Is the returned text well-formed JSON?	293	0	20	313	94%
3	T1	Are any of those links relevant? (*)	160	67	86	313	51%
4	T2	Is the returned text well-formed JSON?	485	0	163	648	75%
5	T2	Did the LLM find any copyright information?	428	0	220	648	66%
6	T2	Did the LLM find any licence information?	169	0	479	648	26%
7	T2	Did the LLM find any terms information?	235	0	413	648	36%
8	T2	Did the LLM find any copyright information? (R)	221	0	49	270	82%
9	T2	Did the LLM find any licence information? (R)	115	0	155	270	43%
10	T2	Did the LLM find any terms information? (R)	133	0	137	270	49%
11	T2	Is the copyright information correct? (*)	47	0	16	63	75%
12	T2	Is the licence information correct? (*)	26		0	26	1.00
13	T3	How many correct decisions are made? (all except -1)	104	0	11	115	90%
14	T3	How many licences are correctly not found? (0)	66	0	49	115	57%
15	T3	How many licences are correctly found? (1 and 2)	38	0	77	115	33%
16	T3	How many licences are linked to the list? (2)	29	0	86	115	25%
17	-	How many licences are automatically linked (of the ones correctly found)?	29	0	9	38	76%
18	-	How many licences are linked (of the total resources)?	38			313	12%

deemed to be potentially relevant (51% were surely relevant and 21% were deemed potentially relevant by our manual assessment).

Task 2 The second task aims at extracting textual content from the web pages mentioning copyright, licence, or terms of use information. The task was executed 648 times, one for each web page collected in the previous step. Those links covered 86% of the collection (270). A good amount of results were provided with a correct JSON syntax – 75% (this includes responses with no information). Copyright information was found in 66% of the cases (82% of the resources, 221/270), while licence information had a much lower result: 26%, corresponding to less than half of the resources for which at least one web page was returned in the previous step (43%). Terms of use are also found with a similar success rate, however, we don't delve into those now and leave an assessment of the quality of this additional information to future work. At the end of this second step, out of 313 initial resources, we obtain copyright information for 221 of them and licence information for 115 of them, approximately 70% and 36% respectively. The reasons vary from errors propagated from the previous step to the information not existing at all on the web pages. Crucially, we validate the quality of the results with a manual supervision of a sample of 100 resources, for which we find that 65% include correct copyright information and 100% include correct licence information (or did not find any when none was there). This information was checked by manually opening each one of the web pages and verifying its content. Crucially, the LLM did not hallucinate when requested to derive licence information from a web page, therefore, the returned content, when valid, is also true.

Task 3 The last task is devoted to automatically linking the licence information to the list of licences in the Dalicc catalogue. The results of this operation were performed on the 115 resources that included any form of licence information (including cases where such information was empty, missing, or non-referring to a specific licence). We evaluate the entire result set manually according to a Likert scale of 5, reflected in questions 13-16 (see Table 5). The prompt to the LLM was designed to identify licences from the list provided, starting from a text that supposedly mentions any of them. We can observe how the system made a correct decision (whether there was a licence from the list or not) in 90% of the cases. However, in more than half of the cases, there was no licence information – 57%. However, the system managed to correctly identify a licence from the Dalicc catalogue for 38 resources (33% of the cases) and in 25% of the cases it was able to report the correct licence code from the list (76% of the ones correctly found). With this approach, we managed to retrieve and link 38 licence information in an automatic (or semi-automatic) way, covering 12% of the resources which originally did not have a licence specified. We conclude this section by discussing the original research questions.

[RQ1] Can copyright and licence information be derived automatically from web pages? We can conclude that it is possible to derive such information from web pages, and automatic methods involving LLM can help in processing large amounts of web pages and gathering relevant information with little human supervision. Crucially, we gathered evidence that there is little risk of generating plausible but wrong information in the case of licencing, thus making us confident that it is possible to apply LLM for extracting licencing information from the content of web resources (see Table 3). This is not true for copyright, as shown by our evaluation of Q11 (reported in Table 2).

[RQ2] How can copyright and licence information be derived automatically from web pages using Large Language Models (LLM)? Our methodology, which was validated by our experiments, is an initial answer to this question. However, we performed our experiments with one specific LLM (ChatGPT) and we acknowledge the fact that a larger study would be needed in order to establish what kind of prompts would be most successful generally, for example considering portability across different LLMs, in achieving this task. However, our experiments are promising and open directions about how to improve the overall workflow both in terms of accuracy and coverage.

[RQ3] How accurate would an LLM detect the copyright and licence information (in other words, is it worth pursuing this line of enquiry)? By looking into the results, we can observe how most of the decrease of coverage during the pipeline was due either to difficulties in producing machine-readable content or in actually recognise that the information is not there (for example, this can be seen by comparing the results of Q9 with Q14). Increasing correct responses in the case of true negatives seems to be a challenge (sometimes the LLM returns some content that does not include relevant information in a task but then this becomes ineffective in further tasks, for example when the LLM returns a piece of text that does not describe a licence in Task 2 and the same text is correctly not linked to any licence in Task 3). Instead, we can observe how the LLM was particularly accurate in deciding, for example, whether a certain piece of text included a licence from a given list (Q13). These results are particularly encouraging and we can definitely see this as a promising research direction.

[RQ4] How much can we complete a curated catalogue of licence metadata with an automatic method based on LLMs? This final answer pertains to our case studies. We managed to find new licence information for 38 resources (12% of the set of resources without licence annotations). We cannot confidently state that those are all the existing missing ones but from the analysis of the results of intermediate steps in our pipelines, we are confident that most of the web pages scrutinised did not include licence information (see results about Q12 and Q13). This is also coherent with the original statistics in musoW, where most of the resources did not present licence information. However, our method allowed us to get more of them, inspiring us to consider opportunities for adopting LLM as an aid for curating digital libraries' metadata.

6. Conclusions

In this paper, we focused on the problem of helping data curators of Web registries to collect and link licence information. To the best of our knowledge, this is the first work focusing on extracting licence information from web resources with LLMs. The risk of LLM hallucinating is not fully dissipated in our results. In the future, we want to improve the quality of the recommendations by refining the prompts and analyse error propagation, as well as extending the evaluation to copyright and terms of use. Furthermore, we plan a larger evaluation comparing different LLMs and models and covering the whole musoW dataset. Finally, we will possibly integrate the method in the data acquisition workflow [23].

Acknowledgements

This work was supported by the EU's Horizon Europe research and innovation programme within the Polifonia project (grant agreement N. 101004746).

References

[1] R. Iannella, S. Guth, D. Pähler, A. Kasten, ODRL: Open Digital Rights Language 2.1, Technical Report, W3C, 2015. URL: https://www.w3.org/ns/odrl/2/ODRL21.

[2] T. Pellegrini, G. Havur, S. Steyskal, O. Panasiuk, A. Fensel, V. Mireles, T. Thurner, A. Polleres, S. Kirrane, A. Schönhofer, Dalicc: a license management framework for digital assets, Proceedings of the Internationales Rechtsinformatik Symposion (IRIS) 10 (2019).

[3] S. Villata, F. Gandon, Licenses compatibility and composition in the web of data, in: Proceedings of the Third International Conference on Consuming Linked Data-Volume 905, CEUR-WS. org, 2012, pp. 124–135.

[4] E. Daga, M. d'Aquin, E. Motta, A. Gangemi, A Bottom-Up Approach for Licences Classification and Selection, in: Proceedings of the International Workshop on Legal Domain And Semantic Web Applications (LeDA-SWAn), co-located with ESWC 2015., CEUR WS, 2015.

[5] M. Daquino, D. Enrico, D. Mathieu, A. Gangemi, H. Simon, L. Robin, A. Merono Penuela, M. Paul, Characterizing the landscape of musical data on the web: State of the art and challenges, in: Workshop on Humanities in the Semantic Web, co-located with ISWC., 2017.

[6] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, A. Jain, Structured information extraction from complex scientific text with fine-tuned large language models, arXiv preprint arXiv:2212.05238 (2022).

[7] M. Steidl, R. Iannella, V. Rodríguez-Doncel, S. Myles, ODRL Vocabulary & Expression 2.2, W3C Recommendation, W3C, 2018. Https://www.w3.org/TR/2018/REC-odrl-vocab-20180215/.

[8] V. Rodríguez-Doncel, S. Villata, A. Gómez-Pérez, A dataset of RDF licenses, in: R. Hoekstra (Ed.), Legal Knowledge and Information Systems. JURIX 2014: The Twenty-Seventh Annual Conference., IOS Press, 2014. doi:10.3233/978-1-61499-468-8-187.

[9] C. Cardellino, S. Villata, F. Gandon, G. Governatori, B. Lam, A. Rotolo, Licentia: a Tool for Supporting Users in Data Licensing on the Web of Data, in: ISWC 2014 Posters & Demo Track. 13th International Semantic Web Conference (ISWC), Riva del Garda, Italy, 2014.

[10] A. Wilke, A. Bannoura, A.-C. N. Ngomo, Relicensing combined datasets, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), IEEE, 2021, pp. 241–247.

[11] E. Daga, M. d'Aquin, A. Gangemi, E. Motta, Propagation of Policies in Rich Data Flows, in: Proceedings of the 8th International Conference on Knowledge Capture, ACM, 2015, p. 5.

[12] E. Cabrio, A. Palmero Aprosio, S. Villata, These Are Your Rights, in: The Semantic Web: Trends and Challenges, volume 8465 of *LNCS*, Springer International Publishing, 2014, pp. 255–269. doi:10.1007/978-3-319-07443-6_18.

[13] D. N. Ribeiro, K. Forbus, Combining analogy with language models for knowledge extraction, in: 3rd Conference on Automated Knowledge Base Construction, 2021.

[14] Z. Yao, Y. Cao, Z. Yang, V. Deshpande, H. Yu, Extracting biomedical factual knowledge using pretrained language model and electronic health record context, in: AMIA Annual Symposium Proceedings, volume 2022, American Medical Informatics Association, 2022.

[15] Y. Gu, S. Zhang, N. Usuyama, Y. Woldesenbet, C. Wong, P. Sanapathi, M. Wei, N. Valluri, E. Strandberg, T. Naumann, et al., Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events, arXiv:2307.06439 (2023).

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[17] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint arXiv:2308.07107 (2023).

[18] Q. Chen, D. Pitawela, C. Zhao, G. Zhou, H.-T. Chen, Q. Wu, Webvln: Vision-and-language navigation on websites, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 1165–1173.

[19] A. Tamkin, M. Brundage, J. Clark, D. Ganguli, Understanding the capabilities, limitations, and societal impact of large language models, arXiv preprint arXiv:2102.02503 (2021).

[20] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, arXiv preprint arXiv:2306.08302 (2023).

[21] L. Yang, H. Chen, Z. Li, X. Ding, X. Wu, Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling, arXiv preprint arXiv:2306.11489 (2023).

[22] J. Frey, L.-P. Meyer, N. Arndt, F. Brei, K. Bulert, Benchmarking the abilities of large language models for rdf knowledge graph creation and comprehension: How well do llms speak turtle?, arXiv preprint arXiv:2309.17122 (2023).

[23] M. Daquino, M. Wigham, E. Daga, L. Giagnolini, F. Tomasi, CLEF. A Linked Open Data Native System for Crowdsourcing, ACM JOCCH 16 (2023).