

# What can I do? Academics perceptions on issues concerning licences

Hamdy Michael Ayas<sup>1</sup>, Linda Erlenhov<sup>1</sup>, and Hannes Thörn<sup>2</sup>

<sup>1</sup>Chalmers University of Technology and the University of Gothenburg

<sup>2</sup>Chalmers University of Technology

## ABSTRACT

In this study we aim to understand what people in academia (academics, researchers and students) find challenging regarding the topic of licences. Specifically, we collect discussions from Q&A fora and manually analyse them to unravel what our stakeholders feel compelled to ask about and discuss online. We collect 407 questions and after filtering we analyze 212 of them using thematic analysis techniques. We identify 8 clusters of challenges, that we further determine based on 6 types of questions they can be concerned with (i.e., What, How, Why, Should, Can, and Is there questions). We find that most questions ask what license to pick for different scenarios and the tools available to help with this selection do not take in to consideration that there might already be licensed material in use. Our findings can be used as input to educational material regarding licences or future "licence helper tools" development, in order to cover what people in academia typically find challenging and ask about in expert Q&A fora.

This work is licensed under a Creative Commons "Attribution 4.0 International" license.



## INTRODUCTION

Open Science is gaining popularity among scientific disciplines, with different schools of thought being evolved on systematically maintaining openness in scientific knowledge (Fecher and Friesike, 2014). The basic premise of openness in science is that research should be done transparently with openness and availability in the developed scientific knowledge, scientific infrastructure, scientific dialogue, and engagement of society with science (UNESCO, 2021). To achieve this openness there are policies that regulate the distribution of scientific content (Burgelman et al., 2019). Such policies are applied via the scientists that publish their research artifacts in open access — scientists that come across several challenges on how to share the knowledge they develop (Olejniczak and Wilson, 2020).

With the growth of internet, knowledge sharing takes many forms and thus, academics, students and researchers are required to learn these new paradigms. Democratization of knowledge brings a lot of content closer to us, but it is not clear where to find information on how to use the content and specifically: what usage is allowed by the authors of the content. For example, open-source software exists in repositories that are publicly available and can be re-used in several ways. However, there are limitations imposed on the resource from the creators and maintainers. These limitations are expressed typically in the form of licences. However, these licences are not always easily understandable.

In this project, we investigate how licences are perceived by academics, students, and researchers in the university realm. To get the perceptions of those stakeholders, we will analyse discussions they post and talk about in online forums that are specific for academics. Our analysis results to a comprehensive taxonomy of challenges that the stakeholders have on the topic of licences. The main research question (RQ) we address is:

***RQ: What are the challenges that researchers face regarding licences?***

To answer this RQ, we query the database of StackExchange Academia (Q&A forum where academics discuss issues publicly) for questions regarding licensing. We filter the resulted raw data to get to a dataset

that we can analyse (using inclusion and exclusion criteria). Then we qualitatively analyse the results from that query (using thematic analysis). This results in a taxonomy of identified challenges. From the analyzed questions, we identify a taxonomy of 8 clusters of challenges. Each of these challenges are further deconstructed into 6 types of questions that researchers have. Specifically, we identify *What, How, Why, Should, Can, and Is there* types of questions per each challenge cluster. We then investigate one of the most common questions in most of these categories, *Which* license that fit in different scenarios and find that the current tools available to assist our stakeholders is not very helpful when it comes to combining different material from different sources - which is a common scenario for the targeted group.

## METHODOLOGY

To address our questions we are applying empirical methods with a qualitative analysis. We query the database of StackExchange Academia (Q&A forum where academics discuss issues publicly) for questions regarding licensing. We then filter the resulted raw data to get a dataset that we can analyse (using inclusion and exclusion criteria). For the analysis we are using card sorting. The approach described by Zimmerman (Menzies et al., 2016) to derive themes from text through card sorting presents a method commonly used for creating mental models and establishing taxonomies from data. This technique aims to extract a higher level of abstraction and identify prevalent themes within textual content. The fundamental principle involves transcribing text onto index cards and subsequently organizing these cards into groups that represent distinct themes or patterns.

Utilizing card sorting offers various advantages over alternative methods, such as annotating text within software like Excel. One significant advantage is the flexibility it provides; physical cards allow for effortless splitting or merging of groups, facilitating dynamic adjustments during the sorting process. Additionally, the tangible nature of the cards enables easy access to review and read the content within each group, enhancing comprehension and analysis.

### Data Gathering

Firstly, we identified the expert discussion exchange site that was within scope of our study. StackExchange maintains more than of 175 Q&A communities<sup>1</sup>. We limited our data collection to the StackExchange Academia community<sup>2</sup>. Next, we gathered data from StackExchange using SQL queries in the StackExchange Data Explorer (SEDE). The queries were designed to gathering questions that initiate discussions on the topic of licences. Therefore, the keywords "*licence*" (Brittish English) and "*license*" (American English) were searched for in the title or body of posts. Below we present the verbatim American English query.

The initial search resulted to 407 question posts.

```
select Id, AcceptedAnswerId, CreationDate, Score,
       ViewCount, Title, Body, Tags
from Posts
where (
    (Title like concat('%', 'license', '%') or
     Body like concat('%', 'license', '%')) and
     ParentId is null
)
```

### Data Processing

The first and third author distributed the questions equally between them and coded them by looking at the title only, for inclusion or exclusion for analysis. The questions had to be about licenses and academia related. This also meant that questions that did not contain "a question" (where the field was blank for some reason) was excluded automatically. The second author then randomly picked 15 questions of each of the other authors questions, to see how the agreement overlapped. We found one disagreement for each combination of authors within those 30 posts, the disagreement was on the same topic and we chose to include them. Based on the discussion that followed the second author methodically examined the previously coded "not include" to see if any of those should also be included in alignment with the

<sup>1</sup><https://meta.stackexchange.com/questions/289233/how-many-sites-are-there-on-stack-exchange>

<sup>2</sup><https://academia.stackexchange.com/>



miro

**Figure 1.** Legend of initial themes used in the card sorting exercise

newly discovered topic, and found one. As a second inclusion criterion, we only use questions that have a positive total voting score. By this we include questions which are somewhat validated by the community, as the aggregate score is at least not negative or neutral. This resulting data set contains 212 questions.

### Data Analysis

The qualitative data analysis adhered to Braun and Clarke's thematic analysis framework, focusing on transcripts of semi-structured interviews categorized by color themes based on distinct types of questions: "What/Which," "How," "Why," "Should," "Can I," and "Is there." The legend for the Miro board can be found in Figure 1

Initially, immersion into the transcripts facilitated a comprehensive understanding of the data without formal coding, allowing the researcher to note initial impressions and key segments. Subsequently, coding of relevant segments according to the various question types commenced, employing an open coding approach to enable the emergence of themes without predetermined categories.

As similar codes were clustered we created themes that aligned with the different question types. These themes encapsulated diverse facets such as preferences and choices; methods or strategies; motivations and justifications; normative considerations; feasibility and opportunities; and queries about existence or availability.

Following theme identification, reviews were undertaken to ensure the coherence and relevance of the themes, verifying their alignment with the original data segments. Clear definitions were then formulated for each theme, and descriptive labels were assigned to encapsulate the core content related to specific question types.

### Data Management Plan

This Data Management Plan outlines the approach for storing openly accessible and reproducible data obtained from Stack Exchange, a network of Q&A websites covering various topics. The data collected from Stack Exchange will be used for research purposes related to licenses.

### Data Source and Licensing

The data utilized in this project are mined from Stack Exchange, which operates under the Creative Commons Attribution-ShareAlike (CC BY-SA) license. The specific licensing terms for each question and answer revision are as follows:

- Content contributed before 2011-04-08 (UTC) is licensed under CC BY-SA 2.5.
- Content contributed from 2011-04-08 up to but not including 2018-05-02 (UTC) is licensed under CC BY-SA 3.0.
- Content contributed on or after 2018-05-02 (UTC) is licensed under CC BY-SA 4.0.

The source of the licences is: CC BY-SA 2.5 Deed — Attribution-ShareAlike 2.5 Generic — Creative Commons Public Network Terms of Service - Stack Overflow

### **Data Collection and Storage**

The data is collected using the Stack Exchange Data Explorer, as mentioned in the Data Gathering section, to extract questions, answers, and user contributions from Stack Exchange sites.

The collected data are stored in a structured format (i.e., CSV) to maintain the integrity of the information. The data is stored in an openly accessible repository that adheres to the principles of open science and ensures reproducibility. We use Zenodo to host the dataset and intermediary analysis results. The dataset is made publicly available and accessible to anyone interested in the research subject. (Michael Ayas et al., 2023). No restrictions are imposed on accessing the data.

A metadata file describing the dataset structure, variables, and any transformations applied are included. Additionally, detailed documentation on data collection methods and processing steps are provided to facilitate reproducibility.

The dataset are released under the same CC BY-SA license as the original Stack Exchange content. Proper attribution and adherence to the ShareAlike conditions are required for any derivative works or publications using this dataset.

Users accessing the dataset are encouraged to cite the original Stack Exchange content and provide appropriate attribution in their work. Clear instructions on citing the dataset are provided in the repository.

### **Data Preservation and Long-Term Availability**

The repository hosting the dataset ensure long-term preservation and availability of the data, maintaining its accessibility and usability for future research. Contact details of the project owner or principal investigator are provided in the repository for inquiries or collaboration opportunities related to the dataset.

### **Ethical Considerations**

Personal identifiers and sensitive information of Stack Exchange users are anonymized or removed to ensure privacy and compliance with ethical standards. In addition, the data are used solely for research purposes, adhering to ethical guidelines and legal regulations, and will not be utilized for any commercial or unlawful activities.

## **RESULTS**

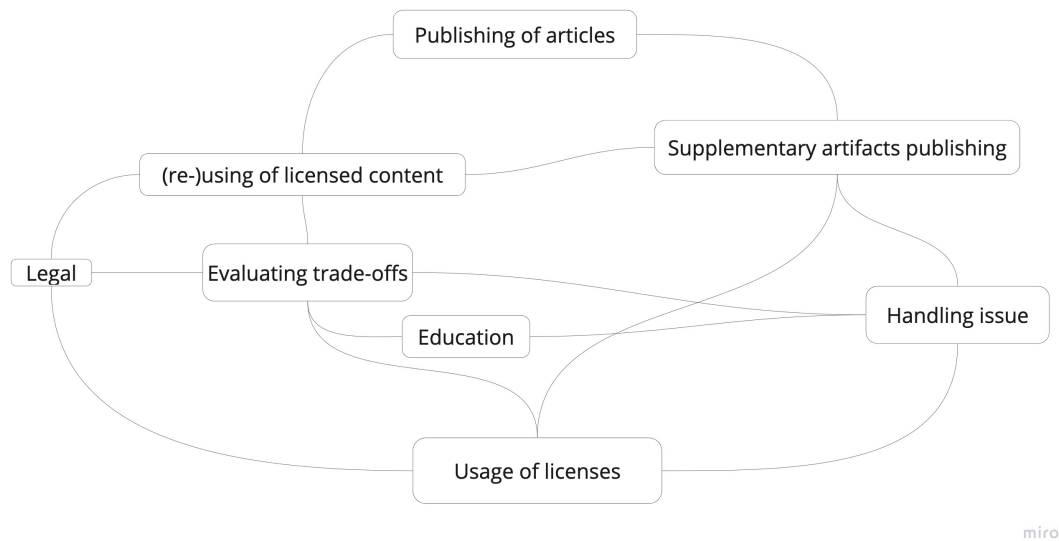
After analyzing the data, we identify the following themes of areas where researchers have challenges, based on their discussions on the Q&A sites. The mindmap of those themes is presented in Figure 2

**Publishing articles:** This theme concerns licenses when publishing papers. For example, questions about how to publish pre-prints of papers and how to handle cases of different versions of publications.

In the exploration of licensing complexities within academic publishing, the investigation centered on various queries related to disseminating papers through diverse channels. Predominantly, these inquiries revolved around pre-print publication and managing different versions of research papers. The analysis categorized these questions based on distinct question types prevalent in this domain.

The most commonly encountered question type, framed as *What/which*, primarily sought clarification on specific choices related to licensing and publication avenues. For instance, researchers sought guidance on platforms permitting pre-print publication or suitable licensing options for varying versions of their papers; or selecting appropriate licenses when submitting papers to repositories such as arXiv. An example question was, "Which license should be chosen in arXiv for a paper to be published in IEEE TPDS?" This inquiry aimed to align the chosen license with both arXiv's policies and the publication requirements of IEEE Transactions on Parallel and Distributed Systems (TPDS). Another example question posed was, "What license to use while putting papers on the Arxiv?" This query aimed to understand the permissible licensing options available for papers uploaded to arXiv, aligning with platform guidelines while ensuring proper dissemination permissions.

The subsequent frequently encountered question type was framed as *Can I*. These inquiries focused on the permissibly or feasibility of specific actions within the publishing domain. For example, researchers inquired about protecting their questionnaire's unauthorized use through publication under a specific license or submitting a paper to a journal already present on arXiv under a public domain license.



**Figure 2.** Mindmap of identified Themes

Another prevalent question type was formulated as *Is there/it*. These questions sought the existence or availability of specific publishing opportunities or guidelines. For instance, researchers queried whether open-source licenses enforced citations or if there were attribution obligations for arXiv pre-prints under the standard license.

Furthermore, researchers sought procedural guidance through questions framed as *How*. These inquiries had the main objective of navigating different versions of a publication or understanding the future-proof nature of arXiv licenses.

Questions categorized as *Why* were less frequent and less prevalent. These aimed to understand the rationale behind certain publishing norms or requirements, such as dollar sign listed near copyright information or the terms of services in platforms like academia.edu.

**Supplementary artifacts publishing:** This theme is derived from questions about publishing supplementary artifacts of publications such as dataset, code and others.

The most frequently encountered question type pertained to inquiries framed as *What/which*. These queries sought clarification or comparison between various aspects related to supplementary artifact publication. For instance, questions emerged regarding the distinctions between an "electronic preprint server" and a "subject repository" or the considerations involved in making externally conceived code for an experimental paradigm publicly available.

Following this, the second most prevalent question type observed was framed as *Is there/it*. Although less frequent than the *What/which* inquiries, these questions sought information about the existence or availability of specific resources or platforms related to publishing supplementary artifacts. For instance, queries arose regarding the availability of internet-based Git-like repositories for paper collaboration or free web hosting services tailored for academics.

The subsequent frequently encountered question type was categorized as *Can I*. These inquiries focused on permission-seeking or feasibility regarding specific actions related to publishing supplementary artifacts. For example, individuals sought advice on publishing source code with potential copyright issues or changing access settings for archives like Zenodo from closed to open.

Additionally, the fourth most common question type, equally frequent as the *Can I* inquiries, was framed as *How*. These questions sought guidance on procedural aspects or methodologies related to the sharing or citation of supplementary artifacts. For instance, inquiries focused on sharing computer code effectively or ensuring proper citation practices for software.

The least common question type encountered among these inquiries was framed as *Should I*. These

queries aimed at seeking recommendations or best practices when deciding on aspects related to publishing supplementary artifacts, such as selecting a Creative Commons license for data publication.

**Usage of licences:** This theme encapsulates questions regarding how to use licences in a correct way. Questions here focus on understanding and learning about different licences or terms and conditions. For example, when are certain clauses triggered, or what type of license to use.

The most popular type of questions regarding the usage of licences is *Is it* or *Does it* questions. Specifically, such questions ask for the existence of licences that can be used, or if licences are triggered in specific contexts. For example, *Is there a CC licence for academic library?*. In addition, such questions ask whether it is possible to use tools or artifacts (e.g., OpenStreetMaps) in research publications. A related question type is the *Can I* queries. Such queries usually refer to investigating potential ways of licensing content.

Furthermore, some queries are *What* questions. Specifically, these are typically asking for clarifications on what exactly a specific licence is, or which licence to use for a specific context.

Finally, there are few questions that are asking *Why*, *How*, and *Should I*. *Why* questions ask for the reasoning on how the licensing system works, whereas *Should I* questions ask for what licensing information should be used. *How* questions ask details on the ways that licence information should be written when used, or when citing licensed content.

**Re-using content:** This theme focus on the re-use of material, such as images; texts; figures; and data. How and when can copyrighted and published materials be used and which licenses permits non-commercial use and otherwise how is it possible to obtain and prove permission to use.

The most recurrently encountered question type centered on inquiries framed as *How*. These inquiries primarily sought insights into two areas i) copyright laws concerning specific scenarios - for instance, questions were posed about the intricacies of copyright for screen captures or guidelines on citing code sourced from MATLAB File Exchange. ii) How to credit the used material - specific questions on where to put the credit and how to phrase it.

Following this, the second most prevalent question type was categorized as *Can I*. These inquiries predominantly focused on seeking permission or determining the permissibility of using copyrighted materials. For example, individuals sought clarification on reusing figures they created and retained the copyright for after publication or inquiring about the permissibility of using figures from their articles on personal websites, especially when published open access while retaining copyright. Additionally there were questions about material where how copyright is applied to the content is not clearly understood. For example, *Can you use quotes or poems in published papers?*.

The third most common was the *What/which* question, for example. These inquiries sought specific details about licensing concerns and permissions when utilizing content from diverse sources. For instance, individuals sought information on licensing concerns for using content from platforms like Stack Exchange in academic papers or inquiring about permissible images from arXiv papers for inclusion in educational materials.

The least common question type encountered among these inquiries was framed as *Should I*. These queries aimed at seeking recommendations or ethical guidance regarding acknowledging licensed code in the context of a thesis or academic work.

**Evaluating trade-offs:** This theme is regarding researchers' concerns on evaluating the pros and cons of different licences and copyright aspects, as well as concerns for balancing economic and privacy concerns with licensing. Queries about evaluating trade-offs are predominantly *What* questions. Specifically, researchers ask about the costs of having open educational resources, or about the advantages and disadvantages of in following reproducible research practices.

**Education:** This theme aggregates questions about teaching the values and benefits of licenses to students, as well as questions about licensing education material. These questions were segmented into two primary categories based on their prevalent question structures.

The most recurrently encountered question type within this educational context featured inquiries framed as *Can I*. These inquiries primarily sought permission or evaluated the feasibility of specific actions within the realm of licensing education. For instance, individuals sought clarification on whether they could utilize GitHub code in their thesis, albeit without a specified license in the repository, or if they could incorporate their designed schematics and layouts from Altium CircuitMaker into their thesis.

Additionally, queries arose regarding the permissibility of purchasing licenses for e-books and lending them to students.

The second most prevalent question type observed in this domain was categorized as *Should I*. These inquiries aimed at soliciting recommendations, best practices, or authoritative guidance regarding the approach to teaching licensing concepts or using licensed educational materials. For instance, individuals sought advice on the instructor's permissible choice of license when having students scribe lectures or the attributive practices concerning teaching material under the GPL (General Public License).

In summary, the analysis of inquiries related to licensing education in academic settings identified two predominant question types. *Can I* inquiries predominantly focused on seeking permission or evaluating feasibility, while *Should I* inquiries revolved around seeking recommendations or best practices in teaching licensing concepts or utilizing licensed educational materials.

**Handling issues:** This theme aggregates several issues that individuals of the academic community had regarding licenses and asked for public wisdom to help resolve them in the best way possible. The queries in this challenge are predominantly about *How*, *Should I*, and *What* questions, even though there are few *Can I* and *Is there* questions.

Most *How* questions are about unclarities in handling specific cases of problematic issues. Specifically, it seems that researchers pose open-ended questions, asking the community on the best way to handle illicit behaviors, disagreements between collaborators on licensing, and dividing economic benefits of licensed patent.

Similarly to *How* questions, *What* questions are open-ended and ask the community ideas on what course of action to take for a specific case. For example, *I chose wrong license during submission on arXiv, what shall I do?* is a query showcasing this.

On the contrary, in *Should I* queries, researchers ask the community in a more close-ended form of questions, if they should perform a specific action. For example, *Should I report my colleagues action, or should i give all intellectual property?*

Finally, researchers ask *Is there* and *Can I* queries to typically clarify a specific issue they face. For example, *Is it reasonable to get the university's legal team to advise on an open source licensing issue that affects software developed for research?*

**Legal & Ethical concerns:** This theme collects concerns that researchers communicate in the community regarding legal or ethical aspects of their research. This challenge is mostly contained by *Is there* and *How* questions, even though there are few *What*, *Can I*, *Why* and *Should* questions.

Most *Is there* queries are about exploring potential legal issues for specific cases such as for releasing crawled data or using pirated software in research. *How* queries are concerned with how to use licences in order to be covered legally, or how to use content without infringing any legal concerns. *Should I* queries are targeting advice regarding course of action.

The *What* and *Can I* questions are open questions on what licences exist for covering any legal concerns, or what can be legally done with research outcomes that are licensed. For example, if publication images of a thesis can be used in academic fairs. Finally, *Why* questions seek explanations on the reasoning regarding specific clauses forbidding the distribution of research artifacts.

## DISCUSSION

From the analysis we see that various challenges associated with licensing, several prominent themes and prevalent question types exist across the academic domain. By clustering similar inquiries we highlight researchers' diverse concerns and inquiries within each domain.

In the realm of publishing articles, the predominant question type, *What/which*, encapsulated queries seeking clarifications on specific choices regarding licensing and publication avenues. Researchers aimed to comprehend suitable platforms allowing pre-print publication and appropriate licensing options for various versions of their papers. The subsequent question types, *Can I* and *Is there/it*, emphasized permissions, feasibility, and the existence of specific opportunities or guidelines. Additionally, *How* and *Why* questions delved into procedural guidance and the rationale behind certain publishing norms, albeit less frequently encountered.

Similarly, in supplementary artifact publishing, *What/which* inquiries predominated, seeking clarifications or comparisons between aspects of artifact publication. Followed by *Is there/it*, *Can I*, and *How* questions, which delved into the existence, permissions, feasibility, and procedural aspects of publishing

**Table 1.** Selection of supporting tools for license selection

| Curated by and Link to site           | Provides guidance on..                                             | Open source      |
|---------------------------------------|--------------------------------------------------------------------|------------------|
| GitHub <sup>3</sup>                   | ..choosing a license for your OSS project                          | Yes <sup>4</sup> |
| Creative Commons <sup>5</sup>         | ..choosing appropriate CC license                                  | unknown          |
| UFAL, Charles University <sup>6</sup> | ..choosing licence for both code and data                          | Yes <sup>7</sup> |
| European Commission <sup>8</sup>      | ..finding and comparing licence for code, data and other artefacts | unknown          |
| European Commission <sup>9</sup>      | ...compatibility of licenses                                       | unknown          |

supplementary artifacts. The infrequent *Should I* queries sought recommendations or best practices in this domain.

Concerning the usage of licenses, prevalent *Is there/it* questions focused on the existence and applicability of licenses in specific contexts. *Can I* queries probed into potential ways of licensing and using content, while *What/which* questions sought clarifications on specific license types or contexts. Meanwhile, *Why*, *How* and *Should I* questions, although less frequent, sought explanations, procedural details, and recommendations regarding license usage.

In re-using licensed content, *How* inquiries dominated, aiming to comprehend copyright intricacies, followed closely by *Can I* queries seeking permissions and *What/which* questions seeking specific details about licensing concerns. Infrequent *Should I* queries sought ethical guidance on acknowledging licensed content.

Education-related inquiries predominantly comprised *Can I* and *Should I* question types, focusing on permissions and best practices in teaching licensing concepts and using licensed educational materials.

In handling issues and evaluating trade-offs, *How* and *What/which* questions were prominent, seeking procedural guidance, best practices, or open-ended advice for resolving licensing issues or weighing advantages and disadvantages.

Lastly, in legal and ethical concerns, prevalent *Is there/it* and *How* questions aimed to explore legal issues, understand licensing for legal coverage, or seek clarification on permissible actions. Other question types such as *What/which*, *Can I*, *Should I*, and *Why* were less frequent, addressing various aspects of legality, permissible actions, recommendations, and reasoning behind licensing clauses.

Across the range of identified challenges, the most prevalent question type was found to be the *What/which* category, consistently accounting for a majority of inquiries. This category primarily encompassed queries seeking specific choices, comparisons, or clarifications within various aspects of licensing. The *How* question type emerged as the second most common inquiry, predominantly seeking procedural insights and a comprehensive understanding of licensing processes or methodologies in different scenarios. Moreover, the *Is there/it* question type ranked third in frequency, often seeking information about the presence, availability, or existence of specific resources, platforms, or guidelines related to licensing and its associated matters.

Keeping this in mind, we investigated what support that exist outside of the foras. There are several tools available on the internet that can assist when choosing a license. We have examined five helper tools, which you can find in Table 1. Four of these tools assume that you are starting with no licensed content involved. However, in most cases that were described by our stakeholders the problem was that several licenses were usually involved: either the material to be licensed contained materials with different licenses, or the final product would be licensed under some predefined license and in that case could other things with other licenses be included in this product. The one tool we found that could help with multiple licenses was the *Joinup Licensing Assistant*<sup>9</sup> curated and maintained by the European Commission. However upon further investigation we found that the output from the tool was still not

<sup>3</sup><https://choosealicense.com/>

<sup>4</sup><https://github.com/github/choosealicense.com>

<sup>5</sup><https://chooser-beta.creativecommons.org/>

<sup>6</sup><https://ufal.github.io/public-license-selector/>

<sup>7</sup><https://github.com/ufal/public-license-selector>

<sup>8</sup><https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-find-and-compare-software-licenses>

<sup>9</sup><https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-compatibility-checker>

easily interpreted unless you are already familiar with licenses.

### **Threats to validity**

Some threats are inherent to the methodology chosen in this study, that readers should take into account.

#### ***Internal Validity***

A threat exists regarding the information that questions posted in the Q&A forum contain. Specifically, there is the threat that questions do not contain generally relevant concerns of researchers and academics. To mitigate this threat we included in our analysis only questions with positive total voting score, ensuring that at least one more peer engineer found the question relevant. In addition, our analysis step of merging concerns under more general themes helps in identifying the more general challenge rather than individual instances of existent issues. In addition, a limitation of this study is that we cannot claim that we have the exhaustive list of challenges that exist regarding licences. Therefore, future work can expand the data source of identifying challenges and evolve the taxonomy we have started in this study, with more challenges.

#### ***External Validity***

Another threat to validity is that we cannot claim representativeness of the study demographics for the entire research and academic community. Specifically, our study is limited to a part of the community that has an online presence in Q&A fora. Therefore, we call for future research to triangulate the results of this study with different data sources, increasing the representativeness of the results. To address the threat, we kept the search term fairly broad and general, including any question that exists regarding licences, and not filtering based on a specific topic of licences (e.g., licences in software, specific licences or academic material).

## **CONCLUSION**

In this project we have investigated what academics, researchers and students' perceptions of issues regarding licenses in their line of work are. We have found that licenses affect many aspects of their tasks, whether it comes to education, conducting research or publishing. The most common questions are related to which licence to use in different cases, and even though many of these cases can be assisted with the various "choose your license"-tools available on the internet we see a gap in their function. Four out of five of the found tools assume you are starting from scratch with your content, while the reality is that in most cases, at least for our stakeholders, there are several objects and artifacts under different licenses involved in the final product and for this the tools help little. The one tool that helped with compatibility still did not provide clarity to the specific cases.

By bringing light to these recurring concerns such as permissions, feasibility, specific guidelines, and procedural guidance encapsulated in these questions, institutions and policymakers can tailor their resources and guidelines. They can curate educational materials that precisely address these prevalent queries, offering clearer guidance and tools directly aligned with the needs of the academic community. This adaptive approach promises more effective support structures that are better equipped to navigate the intricate terrain of licensing complexities within academic settings, ensuring a more seamless and informed engagement with licensing practices and regulations.

Furthermore, even though we have no evidence to support the claim, we do also suspect that these problems exist outside of our scope of people in academia. Therefore we believe that everyone would benefit if a future "choose your license"-tool adopted a workflow where you could "start at both ends". Either by i) adding current licenses on materials used in the product to get suggestions for what license to choose in the end depending on additional preferences or ii) if the end license is already predetermined (by e.g. the publisher) indicate what type of licensed content that cannot be included, and give warnings about clashes in licenses of the source material.

## **REFERENCES**

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., and Schoupe, M. (2019). Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in Big Data*, 2.

- Fecher, B. and Friesike, S. (2014). *Open Science: One Term, Five Schools of Thought*, pages 17–47. Springer International Publishing, Cham.
- Menzies, T., Williams, L., and Zimmermann, T. (2016). *Perspectives on Data Science for Software Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition.
- Michael Ayas, H., Erlenhov, L., and Thörn, H. (2023). Replication package to "What can I do? Academics perceptions on issues concerning licences". <https://doi.org/10.5281/zenodo.10418747>.
- Olejniczak, A. J. and Wilson, M. J. (2020). Who's writing open access (OA) articles? Characteristics of OA authors at Ph.D.-granting institutions in the United States. *Quantitative Science Studies*, 1(4):1429–1450.
- UNESCO (2021). Unesco recommendation on open science. <https://doi.org/10.54677/MNMH8546>.