CLARA Search Engine: Linking Licensed Educational Resources

Manoé Kieffer[1,2], Hugo Chabane[1,3], Matthéo Lécrivain[1,3], and Patricia Serrano-Alvarado[1,2,4]

[1] Nantes University, LS2N, CNRS, UMR6004, 44000 Nantes, France
[2] {name.lastname@}univ-nantes.fr
[3] {name.lastname@}etu.univ-nantes.fr
[4] ORCiD: 0000-0003-2367-5361 *

Abstract. A basic keyword search is the most frequent technique used by teachers looking for reusable Educational Resources (ER). Despite the abundance of ERs on the Web, many remain undiscovered because they are not well connected. In addition, usage rights issues arising from incompatible licences of ERs are a barrier for teachers. In this demonstration, we introduce the CLARA search engine, a web application based on a knowledge graph designed to help teachers in discovering relevant and license-compatible ERs. Based on a set of subjects given by a teacher, the CLARA search engine provides a ranked set of ERs that can be bookmarked to be reused in a new course. The licenses of bookmarked ERs are organised within a compatibility graph, which suggests the licenses that could protect the new course based on the compatibility of the licenses of the bookmarked ERs.

1 Introduction and motivation

In the CLARA project[5], we aim to help teachers create content reusing relevant Educational Resources (ER) without having to delve into licensing aspects. Despite the abundance of valuable resources available on the Web, ranging from slides, videos, figures, text, and code, many remain undiscovered because they are not well connected [1]. In addition, there are usage rights problems since it is not legal to combine resources in a teacher's course if their licenses are not compatible with the course's license. These legal issues represent a barrier for the teacher himself and the institution that is hosting his course. Ideally, analysing available resources to match a course plan and the license verification should not be time-consuming.

There are platforms to help learners construct personalized learning paths[6], to allow teachers to share their ERs[7], or even platforms to help teachers cre-

[5] https://project.inria.fr/clara/

[6] https://labs.tib.eu/edoer/

[7] https://www.merlot.org/

ate new ERs[8][9]. However, a solution to help teachers find relevant and license-compatible ERs is missing. CC Search Portal[10], finds images licensed under a Creative Commons license. Similarly, the Google search engine can filter images by access rights according to licenses (Creative Commons licenses and other commercial licenses). However, if a user wants to reuse multiple images protected by different licenses, these search engines do not suggest the license that might protect the image mashup. Moreover, there are tens of open and free licenses[11] and they are not considered in these solutions.

An effective way to enhance the discoverability and reusability of ERs is by using Linked Data principles. Thus, we have developed a search engine based on a knowledge graph (KG) to help teachers find ERs. Its functionalities are the following: (1) It returns an ordered set of the most relevant ERs to a subject search. (2) Each ER indicates its accessible URL, authors, license, year of publication and most relevant subjects. (3) The search can be filtered by language, format, and license. (4) A visual graph with the resulting ERs linked with their most relevant subjects is provided. (5) ERs that are interesting for a new course can be bookmarked and the compatibility graph of their licenses is provided [4]. The goal is to highlight licenses capable of protecting new content based on the licenses of the bookmarked ERs. If a suitable license cannot be found, our tool provides explanations regarding incompatible licenses. This strategic feature gives teachers the means to design new resources that integrate ERs without the hassle of incompatible licensing issues.

In this paper, we describe the design of the CLARA search engine.

2 CLARA search engine

ERs can be described with several properties such as title, authors, language, license, etc., as well as the subjects they cover. While ERs may cover multiple subjects, not all subjects are equally significant to the resource. Some subjects serve as primary focal points, while others are briefly mentioned. Consequently, it is essential to assess the relevance of each subject and weigh their relationship with each ER accordingly. Statement-level reification[12], allows us to annotate with scores the relation between ERs and the subjects they treat. As an example, the following triple (in RDF-star[13]) states that an ER focuses, with a score of 0.4, on a subject : `« :ER1 dct:subject :Query_Language » uno:pageRank "0.4"`.

---

[8] `https://www.oercommons.org`

[9] See this metasearch engine of OERs `https://oer.deepwebaccess.com/`

[10] `https://search.creativecommons.org/`

[11] `https://en.wikipedia.org/wiki/Free_license/`

[12] Reification allows to write RDF statements about RDF statements. We focus on statement-level reification which allows to write RDF statements about a single RDF statement.

[13] `https://w3c.github.io/rdf-star/cg-spec/2021-12-17.html`

Various reification models exist, each with distinct syntax and performance implications for storage and query processing. In [2], we compare four statement-based reification models (including standard reification, singleton properties, named graphs, and RDF-star) on four triplestores (Virtuoso, Jena, Oxigraph, and GraphDB) to determine the most pertinent choice for our use case. The four versions of the analyzed knowledge graphs (KG) are available in [3]. Our experiments indicate that both, standard reification and named graphs, when used with a Virtuoso triplestore, exhibit the best performance with our KG. While RDF-star presents an elegant and concise model for statement-based annotations, the efficiency of its implementations should be improved if *quoted triples* are included in RDF 1.2.

Currently, the CLARA KG links 45,000 ERs (licenced under twelve different open licences) with 135,000 subjects, collected in the European project X5GON[14]. Statement-level reification is used to annotate the `dct:subject` relation with a PageRank score, ranging from 0 to 1, obtained through a wikification process[15]. On average, each ER is linked to 184 subjects (with a median of 171 subjects per ER). In contrast, subjects are more sparsely distributed, with an average of 61 ERs linked to each subject (with a median of 2 ERs per subject). Therefore, linking ERs through their relevant subjects enables the discovery of relevant and related ERs that can be reused in a course given a list of subjects.

Normalization of PageRank scores. The PageRank score of a subject is local to an ER, as it depends on the number of subjects associated with this ER. The sum of the PageRank scores of all subjects linked to an ER is 1. Consequently, the more subjects linked to an ER, the lower their individual PageRank scores. This local nature of the metric complicates the comparison of resources by subject. For example, an ER highly relevant to a particular subject but linked to multiple other subjects may have a lower PageRank for that subject compared to a less relevant ER linked to fewer subjects. This issue makes necessary a ranking function that uses normalized PageRank scores to account for a proportional distribution of these values by subject.

The normalized PageRank score of a subject linked to an ER uses the magnitude of the vector which is composed of all the PageRank scores of the ER. We define the magnitude of the vector of an ER as:

Definition 1. $magnitude_{ER_i} = \sqrt{PageRank_{s1}^2 + PageRank_{s2}^2 + ... + PageRank_{sn}^2}$

where s_i are the subjects linked to the ER. This magnitude allows the definition of normalized vectors as:

Definition 2. $normVector_{ER_i} = \left[\frac{PageRank_{s_1}}{magnitude_{ER_i}}, \frac{PageRank_{s_2}}{magnitude_{ER_i}}, \ldots, \frac{PageRank_{s_n}}{magnitude_{ER_i}}\right].$

The normalized vectors are then used to compare ERs by subject in a manner that is not influenced by the length of the vectors, as the magnitude of a normalized vector is always 1.

[14] `https://www.x5gon.org`

[15] Wikifier tool `https://wikifier.org`

Ranking function. From a set of subjects provided by the user, the ranking function creates a hypothetical ER (a goal) which serves as the ideal response to the query. Currently, this hypothetical ER does not distinguish the relevance of subjects. In other words, all subjects are considered equally relevant, with identical PageRank scores, and their sum equals 1. The search engine then ranks the existing ERs based on their distance (cosine similarity) from this hypothetical goal. The cosine similarity function is used to calculate the pairwise similarity between the ER representing the goal and all other ERs. Cosine similarity is defined as the dot product of the vectors divided by the product of their magnitudes. Since the magnitude of normalized vectors is always 1, the similarity is simplified to just the dot product of the vectors.

The web application. The web application of the CLARA search engine is implemented with the framework Vue.js[16] in TypeScript.

- The backend calculates the similarity measures using Elasticsearch[17] to index the subjects, the related ERs and the corresponding normalized PageRank scores. The produced ranking is used to request the ERs to our Virtuoso endpoint with SPARQL queries. Results contain several properties of each ER as their title, authors, license, language, abstract, etc. The backend is accessible from a public API and can be used independently.
- The frontend displays results in pages of twenty ERs each. Additionally, a graph tab provides a visualization containing the same twenty ERs along with their three best-linked subjects. To ensure continuity of teacher activities, bookmarked ERs are saved between sessions using the browser's local storage, facilitating continued selection of ERs over time.

3 Demonstration

Figure 1 shows a screenshot of the CLARA search engine. In this example, four ERs are bookmarked to be reused in a new course. On the right, the compatibility graph of licenses shows in green the license that is compliant to the licenses of the bookmarked resources. If all the licenses are red, a problem of license compatibility exists. During the demonstration, attendes will be able to use the CLARA search engine.

All our resources are publicly available under open licenses.

- The CLARA search engine: `https://clara.univ-nantes.fr/`.
- The associated API: `https://clara.univ-nantes.fr/api-docs`.
- The source code: `https://gitlab.univ-nantes.fr/clara/CLARA-Backend` and `https://gitlab.univ-nantes.fr/clara/CLARA-front`
- A SPARQL endpoint to access the knowledge graph (in standard reification): `https://clara.univ-nantes.fr/sparql`.

[16] `https://vuejs.org`

[17] `https://github.com/elastic/elasticsearch`

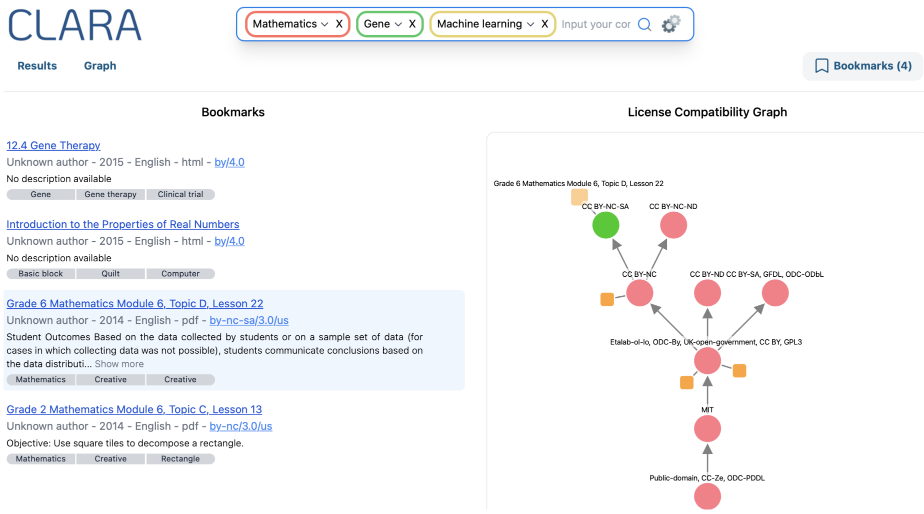

Fig. 1. Screenshot of the CLARA search engine. It shows the compatibility graph of licences of some bookmarked ERs. The license in green can protect a new course mixing-up the bookmarked ERs.

- The CLARA knowledge graph is available at `https://doi.org/10.5281/zenodo.8403142`
- A video of the CLARA search engine: `https://youtu.be/2MEZd5Wr-IE`.

Acknowledgment. This work has received a French government support granted to the Labex Cominlabs excellence laboratory and managed by the French National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01. The authors acknowledge, Matthieu Perrin for his help in defining the ranking function, as well as Ginwa Fakih, and Master's students in Computer Science at Nantes University for their involvement in this project.

References

1. Cortinovis, R., Mikroyannidis, A., Domingue, J., Mulholland, P., Farrow, R.: Supporting the Discoverability of Open Educational Resources. Education and Information Technologies **24**, 3129–3161 (2019)
2. Kieffer, M., Fakih, G., Serrano-Alvarado, P.: Evaluating Reification with Multi-valued Properties in a Knowledge Graph of Licensed Educational Resources. In: 19th International Conference on Semantic Systems (SEMANTiCS). Leipzig, Germany (Sep 2023). https://doi.org/10.3233/SSW230008
3. Kieffer, M., Fakih, G., Serrano Alvarado, P.: CLARA Knowledge Graph of Licensed Educational Resources (Oct 2023). https://doi.org/10.5281/zenodo.8403142
4. Moreau, B., Serrano-Alvarado, P., Perrin, M., Emmanuel, D.: Modelling the Compatibility of Licenses. In: 16th Extended Semantic Web Conference (ESWC). Springer, Portoroz, Slovenia (Jun 2019). https://doi.org/10.1007/978-3-030-21348-0_17, `https://hal.archives-ouvertes.fr/hal-02069076`