Crowdsourced and Automated Educational Resource Metadata

Vidminas Vizgirda
University of Edinburgh
Edinburgh, United Kingdom
s1750767@ed.ac.uk

Abstract

Educational metadata can facilitate search for educational resources, however adding metadata requires time and effort from resource creators and is an often overlooked part of publishing. Previous studies have investigated automated metadata generation and crowdsourced social metadata tagging. This demo presents a prototype browser plugin featuring both automated and crowdsourced metadata to enhance educational resource search. Prototype source code is available at github.com/Vidminas/educational-search-filters.

CCS Concepts

• **Information systems** → **Web and social media search**; • **Human-centered computing** → *Computer supported cooperative work*; *Social content sharing*.

Keywords

Search, Relevance, Metadata, OER, Linked Data, Materials, Discoverability, K-12 Education, Schools

ACM Reference Format:
Vidminas Vizgirda. 2025. Crowdsourced and Automated Educational Resource Metadata. In *2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25), March 24–28, 2025, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3698204.3716472

1 Introduction

There are numerous repositories of educational resources, like MIT OpenCourseWare, OER Commons, and MERLOT. However, the top places where teachers look for resources are usually Google and YouTube [3, 6, 14], which are not specialised for educational resource search. A 2021 survey with 5442 schoolteachers in Australia found that most (~88%) participants thought they could save time if shared high-quality resources were available [8]. Teachers want "one place to find all the material, sorted by subject and standards" [2, p.9] and such places exist, so how can we explain the apparent paradox that the vast majority of teachers do not use them?

Previous studies have identified a multitude of barriers to using educational resource repositories: copyright issues, reusability concerns, difficulty of finding high-quality materials, and others [5, 13]. On the other hand, one of the key strengths of specialised educational resource repositories is that they store education domain-specific metadata, like resource subject area and intended audience level, allowing users to filter and sort results by such criteria. The main limiting factor is that tagging educational resources with metadata imposes an additional burden on the content creators, who are often already going out of their way to publish resources. Commonly, only mandatory metadata are filled in to satisfy the bare minimum requirements of resource repositories.

Automated metadata tagging could potentially alleviate content creator burden, but previous efforts to automate metadata tagging had only limited success. Recent advances in Natural Language Processing and Computer Vision show renewed potential for automatic metadata inference. Another approach could be to leverage crowd wisdom and build a Metadata Commons or Social Metadata that anyone can contribute to, shifting the burden of tagging from content authors to users.

This demonstration presents a proof-of-concept browser extension – Educational Resource Search Filters – that augments Google search results pages and result sites' interfaces with elements that enable both automated and crowdsourced educational metadata tagging.

2 Related work

The filtering and tagging functionality implemented by the Educational Resource Search Filters plugin is already available in most educational resource repositories, for example, MIT OpenCourseWare, OER Commons, MERLOT, TES Resources, and Twinkl. All these examples, however, rely on resource authors and content curators to provide all the metadata.

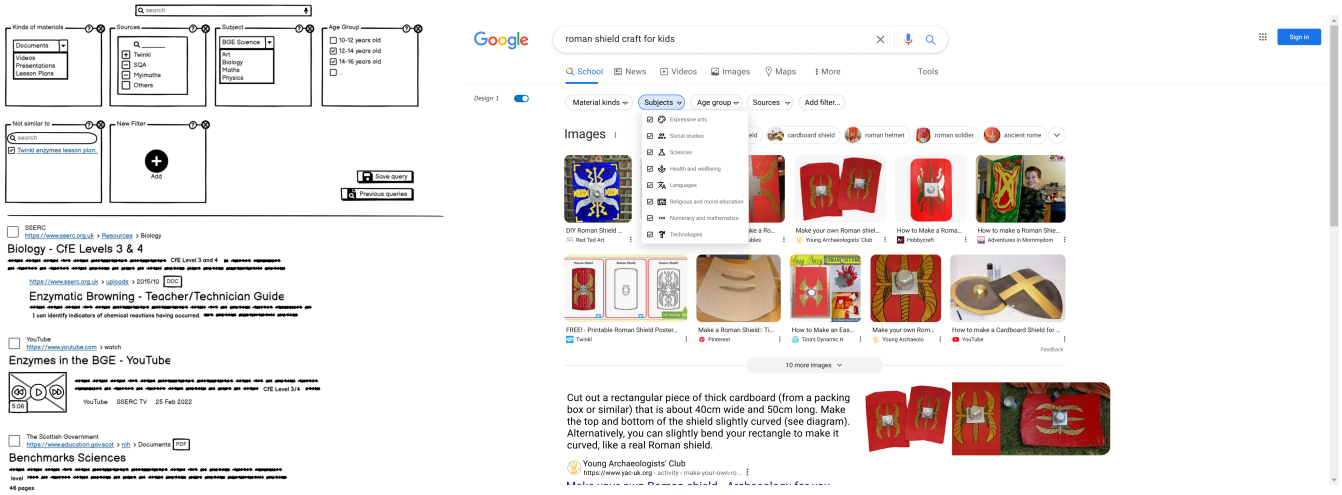
There is an abundance of previous research on social tagging [17], where users could collaborate on tagging resources and establish "folksonomies" – flexible ontologies to describe content. Most social tagging systems were very open-ended and fully manual, making them labour intensive to use. The design of this project borrows inspiration from earlier social tagging systems, however in an attempt to minimise user effort, tags are limited to a fixed vocabulary and are complemented with automatic generation.

The Learning Registry previously explored the idea of Social Metadata, with users contributing tags for others' content [1] before it was shut down due to lack of funding in 2018 [19]. Using a Learning Registry data backup, Cortinovis et al. [3] created a proof-of-concept browser extension that augments Google search results with Learning Registry metadata, showcasing the potential usefulness of domain-specific metadata in general search. Several other projects explored automatically generating educational resource metadata [7, 12, 21, 22] with varying levels of success but none achieved widespread deployment.

Hoffmann et al. [9] investigated whether an automatic information extraction system, Kylin, could be useful for filling in missing summary information in Wikipedia infoboxes. The system would generate suggestions from page content and prompt page visitors to

CHIIR '25, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1290-6/25/03
https://doi.org/10.1145/3698204.3716472



(a) Low-fidelity Balsamiq wireframe mock-up

(b) High-fidelity Figma design mock-up

Figure 1: Earlier prototype designs used to gather formative feedback

verify the accuracy of suggestions. This combined automation and crowdsourcing approach allowed not only to engage more people in editing Wikipedia articles and quality assure infobox information but also to improve Kylin's extraction performance over time. However, the authors noted that the system struggled with extracting information that was not directly mentioned in page text. This is important in educational resource metadata because concepts like curriculum subject area are often not specified but can be inferred from the context.

On a larger scale, `schema.org`, an initiative founded by Google, Microsoft, Yahoo, and Yandex, provides a common schema that website creators can use to specify information about their pages and content. Search engines can interpret this metadata to deliver more meaningful search results. For educational resources, the Dublin Core Learning Resource Metadata Initiative (LRMI) [4] is one of the most well-known metadata schemata, and is partially integrated into `schema.org` as an extension. This allows search engines that support `schema.org` to query specifically for learning resources and filter by educational criteria, for example, quizzes that match a given education standard alignment. Recalde et al. [15] created a prototype user interface to write JSON-LD linked data metadata for online educational resources based on `schema.org`. In practice, few websites (including educational resource repositories) tag their content using the LRMI metadata extension [16], which means search engines cannot distinguish most educational content from other web content. Furthermore, the subset of educational metadata supported by `schema.org` is limited compared to the full LRMI or other standards like the IEEE Learning Object Metadata [10].

Investigating the kinds of filters that teachers use in practice, Yacobson and Alexandron [20], in a study with 465 physics teachers, found that 41.5% of the participants used search filters when looking for resources in a specialised learning resource repository. The most used filters were for resource type, subject, and whether resources had been quality assured.

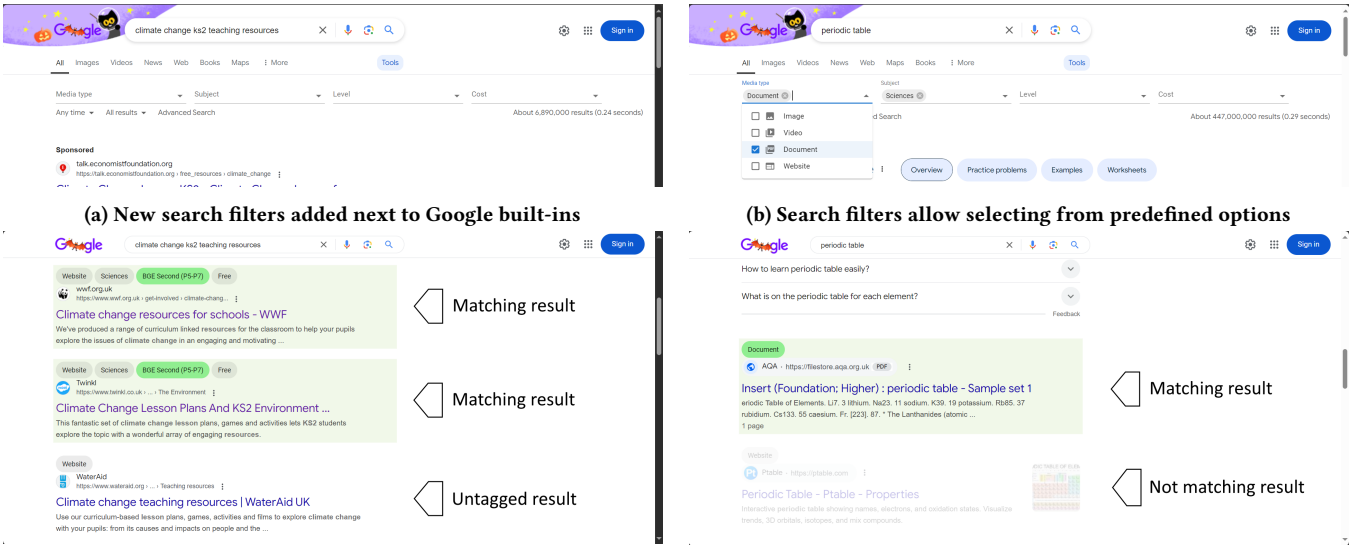
The Educational Resource Search Filters plugin is built on top of an open-source browser extension "Highlight or Hide Search Engine Results"[1]. The plugin source code is available on GitHub[2] and includes setup instructions.

3 User-Centred Design

The idea for this plugin was developed together with teachers through several user-centred design iterations. User studies were approved according to the University of Edinburgh School of Informatics Research Ethics Process, reference number `2022/42906`.

First, the idea was inspired by conversations with teachers about how they search for resources online – searching keywords to narrow down results by curriculum, by level, and by source/author were common strategies. I made low-fidelity wireframe mock-ups using Balsamiq (Figure 1a), printed them out, and gathered in-person feedback during August-October 2023 in four focus group workshops: two groups in two schools in Scotland and two groups at the Open Education Global 2023 conference. The participants included 4 teachers and 14 education experts. Participant feedback ("having lots of filters is useful but also makes this difficult to use. Need a balance") led me to refine the design, making it more streamlined with the Google user interface and more intuitive to use. In June-July 2024, I met 3 more Scottish schoolteachers over four online meetings to review a high-fidelity Figma design mock-up (Figure 1b). Teacher feedback shaped important design decisions, such as not hiding any search results and not reordering results so that wrongly tagged or untagged results would not get lost (participants were especially concerned about missing potentially useful results if they were hidden by the plugin). Finally, in October 2024, I met 2 students with experience in design and 2 teachers

[1] https://github.com/pistom/hohser
[2] https://github.com/Vidminas/educational-search-filters



(a) New search filters added next to Google built-ins

(b) Search filters allow selecting from predefined options

(c) Search results that match filters are highlighted

(d) Search results that do not match filters are semi-transparent

Figure 2: Screenshots of Google search results pages with the Educational Resource Search Filters plugin

over five meetings to refine an early version of the browser plugin. This involved redesigning plugin components and adding a top bar on every page, so that metadata tags would be easy to find (users would not need to remember to find them in a context submenu or plugin menu) but not too intrusive (like a pop-up). Participating teachers suggested several other filters that could be useful, for example, Scottish Curriculum for Excellence benchmarks and whether resources can be accessed without a login.

4 Educational Resource Search Filters

The Educational Resource Search Filters browser extension injects new user interface elements into Google search results pages and all other websites. On search results pages, it automatically opens Advanced Search options (which, by default, are hidden inside the "Tools" section). This makes built-in filters (for recency of results and verbatim or default query matching) visible. Above the built-in filters, new ones are added – the screenshot in Figure 2a shows filters for media type, subject, level, and cost. Each of the filters includes a predefined set of values to choose from, for example, media type includes options for "Image", "Video", "Document", and "Website" (as shown in Figure 2b). Above each search result, metadata tags for that result are displayed. Results that have tags that match filter selections are prominently highlighted (like the two top results in Figure 2c). Results that have relevant tags that do not match are made less prominent by applying a layer of semi-transparency (like the bottom result in Figure 2d) but not completely hidden. And results that do not have relevant tags remain neutral (like the bottom result in Figure 2c).

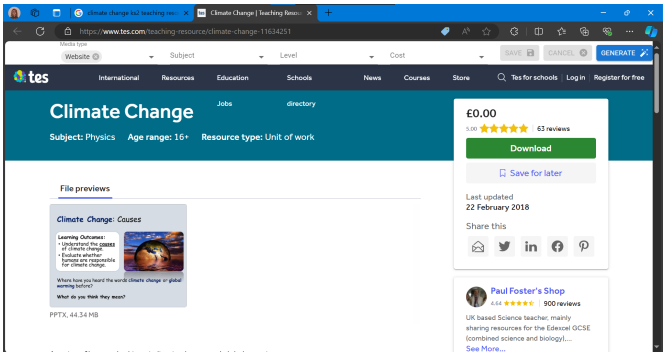
The browser extension also adds a top bar with metadata tags to any visited page that is not a search engine page. The tags mirror options available in search filters. Page media types are automatically determined from their respective `Content-Type` HTTP headers. On visiting any page, users can manually modify tags assigned to that page or generate them automatically. Automatic generation sends a request to OpenAI's `GPT-4o-mini` model API with the resource text contents and, if the resource is an HTML file and metadata are present, the page metadata, requesting it to assign the most appropriate tag for each category or "Unknown" if it cannot be determined. OpenAI API Structured Outputs mode [11] is used to ensure that only valid tag values are generated. When "Unknown" is returned, corresponding tags are left unassigned.

When assigned metadata tags are saved, they appear above the respective website if it comes up in the search results page in the future. The key idea of this plugin is that all assigned metadata would be shared between users in a Metadata Commons, so that anyone's work assigning tags would benefit everyone else.

5 Discussion and Conclusion

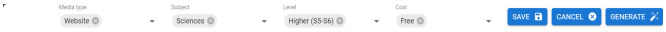
Although the current version of the browser plugin is a useful artifact and design probe for exploring whether a Metadata Commons could be possible and what it would take to make it work, it is not a finished end-user product, just an early prototype. A further user study, potentially with teachers trying simulated work tasks with the prototype, would be required to get a better understanding of how it fits with teachers' information interaction tasks.

Anecdotally, the metadata generated by `GPT-4o-mini` is fairly accurate, but most of the time the model can only determine the resource subject area, leaving level and cost up to the user to fill in. Accuracy is not a major concern because of the plugin's human-in-the-loop design, which means users would be able to catch and correct any errors. With more complex filters, automatic metadata generation may need more inputs, for example curriculum alignments might not be possible to determine from resource content alone because they frequently depend on authors' intended use for their resource.

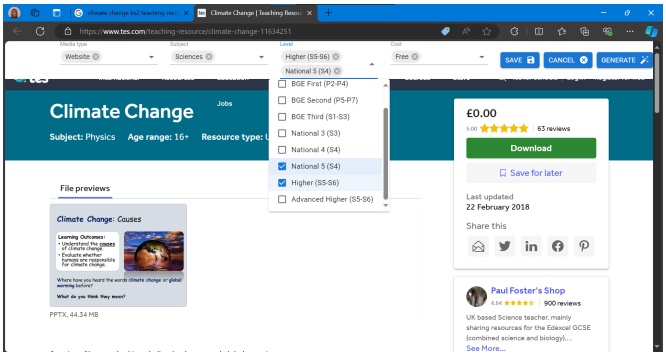


(a) Example resource page with new top bar for metadata tags

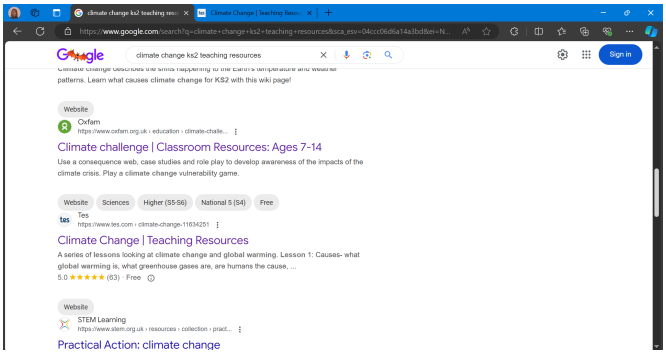
(b) Generating new tags automatically takes a few seconds



(c) Generated tags populate the top bar



(d) Metadata tags can be manually adjusted



(e) Assigned tags appear above linked search result in Google

Figure 3: Screenshots from an example result page with top bar tags

From an environmental ethics perspective, if this tool were to be scaled up to widespread usage, the energy consumption of large language model inference may become an issue. This was one of the reasons for choosing `GPT-4o-mini`, the smaller and less compute-intensive model, rather than the full-scale `GPT-4o` model, in addition to differences in API call pricing and response times.

The user studies have been carried out mostly in the Scottish education context, which influenced the choices for specific filters and possible tags. If implemented in another context, different filters may be required – for example, educational outcome alignments, subject area delineations, and level classifications would be different. In theory, the same idea could be extended to contexts outside of the education domain too, as long as there were domain-specific metadata that would be helpful for narrowing down search results.

Many open questions remain: would it be worth the effort and time for users to tag resources? Where would shared metadata be stored? Would it all be in one place (thus requiring consensus on a common metadata schema) or could there be multiple decentralised hosts for different metadata in different contexts? Would all metadata be shared between everyone or should some tags remain private? How might we prevent users from removing or changing useful tags? Who would bear the cost of running Large Language Models to generate metadata? Could it be done with local models?

In conclusion, this demo presents a novel way to augment generic search results. Further research is still required to understand the full extent of its potential impact and implications.

Acknowledgments

This work was done as part of a larger PhD project [18] at the University of Edinburgh. OpenAI API usage costs were sponsored by the University of Edinburgh. Thank you to my supervisors Dr Fiona McNeill and Prof Judy Robertson, the Global OER Graduate Network, and all the teachers, friends, and colleagues who contributed to this project through their advice or participation.

The author acknowledges the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR CHIIR 2025 was hosted. I pay my respect to their Elders past and present and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their connections to land, sea, sky, and community.

References

[1] Marie Bienkowski and James Klo. 2014. The Learning Registry: Applying Social Metadata for Learning Resource Recommendations. In *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications*, Nikos Manouselis, Hendrik Drachsler, Katrien Verbert, and Olga C. Santos (Eds.). Springer, New York, NY, 77–95. doi:10.1007/978-1-4939-0530-0_4

[2] Boston Consulting Group. 2013. *The OER Knowledge Cloud: The Open Education Resources Ecosystem An Evaluation Of The OER Movement's Current State And Its Progress Toward Mainstream Adoption.* Technical Report. William and Flora Hewlett Foundation. 1–23 pages. Retrieved 2024-11-01 from https://www.oerknowledgecloud.org/record765

[3] Renato Cortinovis, Alexander Mikroyannidis, John Domingue, Paul Mulholland, and Robert Farrow. 2019. Supporting the Discoverability of Open Educational Resources. *Education and Information Technologies* 24, 5 (Sept. 2019), 3129–3161. doi:10.1007/S10639-019-09921-3

[4] DCMI. 2020. Dublin Core Metadata Initiative: Learning Resource Metadata Initiative. Retrieved 2021-12-17 from https://www.dublincore.org/specifications/lrmi/

[5] Beatriz De Los Arcos, Robert Farrow, Rebecca Pitt, Leigh-Anne Perryman, Martin Weller, and Patrick McAndrew. 2015. OER Research Hub Data 2013-2015: Educators. http://oro.open.ac.uk/47931/

[6] Beatriz De Los Arcos, Robert Farrow, Rebecca Pitt, Martin Weller, and Patrick Mcandrew. 2016. Adapting the Curriculum: How K-12 Teachers Perceive the Role of Open Educational Resources. *Journal of Online Learning Research* 2, 1 (2016), 23–40. Retrieved 2022-02-05 from https://www.learntechlib.org/primary/p/151664/

[7] Stefan Dietze, Honq Qing Yu, Daniela Giordano, Eleni Kaldoudi, Nikolas Dovrolis, and Davide Taibi. 2012. Linked Education: Interlinking Educational Resources and the Web of Data. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12)*. Association for Computing Machinery, New York, NY, USA, 366–371. doi:10.1145/2245276.2245347

[8] Grattan Institute. 2022. Making Time for Great Teaching: How Better Government Policy Can Help. Retrieved 2022-06-01 from https://grattan.edu.au/report/making-time-for-great-teaching-how-better-government-policy-can-help/

[9] Raphael Hoffmann, Saleema Amershi, Kayur Patel, Fei Wu, James Fogarty, and Daniel S. Weld. 2009. Amplifying Community Content Creation with Mixed Initiative Information Extraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1849–1858. doi:10.1145/1518701.1518986

[10] IEEE. 2020. IEEE Standard for Learning Object Metadata. *IEEE Std 1484.12.1-2020* (Nov. 2020), 1–50. doi:10.1109/IEEESTD.2020.9262118

[11] OpenAI. [n. d.]. Structured Outputs - OpenAI API. Retrieved 2024-10-31 from https://platform.openai.com/docs/guides/structured-outputs/

[12] Saurabh Pal, Pijush Kanti Dutta Pramanik, and Prasenjit Choudhury. 2021. Enhanced Metadata Modelling and Extraction Methods to Acquire Contextual Pedagogical Information from E-Learning Contents for Personalised Learning Systems. *Multimedia Tools and Applications* 80, 16 (July 2021), 25309–25366. doi:10.1007/s11042-020-10380-z

[13] Maria Perifanou and Anastasios A. Economides. 2022. Discoverability of OER: The Case of Language OER. *Smart Innovation, Systems and Technologies* 249 (2022), 55–66. doi:10.1007/978-981-16-3930-2_5

[14] Kristen Purcell, Alan Heaps, Judy Buchanan, and Linda Friedrich. 2013. *How Teachers Are Using Technology at Home and in Their Classrooms*. Technical Report. Pew Research Center.

[15] Lorena Recalde, Rosa Navarrete, and Fernando Pogo. 2021. Making Open Educational Resources Discoverable: A JSON-LD Generator for OER Semantic Annotation. In *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, Quito, Ecuador, 182–187. doi:10.1109/ICEDEG52154.2021.9530872

[16] Ratan Sebastian and Anett Hoppe. 2024. An Updated Analysis of Learning Resource Metadata Usage on the Web. In *Linking Theory and Practice of Digital Libraries*, Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz (Eds.). Springer Nature Switzerland, Cham, 85–94. doi:10.1007/978-3-031-72440-4_8

[17] J. Trant. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10, 1 (Jan. 2009), 1–44. Retrieved 2025-02-09 from https://jodi-ojs-tdl.tdl.org/jodi/article/view/269

[18] Vidminas Vizgirda. 2024. Educational Resource Search in Scottish Schools. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Sheffield United Kingdom, 449–452. doi:10.1145/3627508.3638320

[19] John Watson. 2018. Lessons from the Rise and Fall of the Federal Learning Registry. Retrieved 2024-11-04 from https://www.digitallearningcollab.com/blog/2018/12/12/lessons-from-the-rise-and-fall-of-the-federal-learning-registry

[20] Elad Yacobson and Giora Alexandron. 2023. How Do Teachers Search for Learning Resources? A Mixed Method Field Study. In *Responsive and Sustainable Educational Futures*, Olga Viberg, Ioana Jivet, Pedro J. Muñoz-Merino, Maria Perifanou, and Tina Papathoma (Eds.). Springer Nature Switzerland, Cham, 489–503. doi:10.1007/978-3-031-42682-7_33

[21] Tolga Yilmaz, Rifat Ozcan, Ismail Sengor Altingovde, and Özgür Ulusoy. 2019. Improving Educational Web Search for Question-like Queries through Subject Classification. *Information Processing & Management* 56, 1 (Jan. 2019), 228–246. doi:10.1016/j.ipm.2018.10.013

[22] Ozgur Yilmazel, Christina M. Finneran, and Elizabeth D. Liddy. 2004. Metaextract: An NLP System to Automatically Assign Metadata. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '04)*. Association for Computing Machinery, New York, NY, USA, 241–242. doi:10.1145/996350.996405