The CDL: An Online Platform for Creating Community-based Digital Libraries

Kevin Ros
kjros2@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

ChengXiang Zhai
czhai@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

We present the Community Digital Library (CDL), a novel and extensible platform for collaborative information seeking which enables any group of users to (1) describe and save webpages relevant to their shared interests, (2) share and search the saved webpages, and (3) discover content via recommendation. The CDL is free-to-use, can be accessed online, and the source code is publicly available.

KEYWORDS

information retrieval; collaborative search; social bookmarking; recommendation

ACM Reference Format:
Kevin Ros and ChengXiang Zhai. 2023. The CDL: An Online Platform for Creating Community-based Digital Libraries. In *Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion), October 14–18, 2023, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3584931.3607495

1 INTRODUCTION

As a "universal" digital library, the Internet has vast amounts of useful information in the form of webpages, and people around the world use search engines daily to find relevant information on all kinds of topics. Despite the high utility of search engines, it is still quite difficult for a user to find relevant content. Challenges such as vocabulary gaps, a lack of domain knowledge, the fast growth and variable quality of new content (compounded by generative large language models [4]), search engine-optimized webpages, and advertisement-driven result pages can limit a user's capability of finding high-quality information and/or the best answer to their information need.

One potential way to address these challenges in certain settings is to leverage a community-based collaborative information-seeking system [5]. When a community of users who share a similar goal seek information together, they are able to directly mitigate many of the aforementioned challenges. For example, an expert user can collect and recommend information to a novice user, thus addressing the vocabulary gap and the lack of domain knowledge. Moreover, a community-based collaborative information-seeking system would enable a user to archive and easily re-find information. And with the support of adding descriptions to saved webpages, such a system would further allow a user to more easily find content in a personalized manner, thus mitigating various content quality and popularity limitations.

Ideally, we would like to enable members of any community to create an online community-based digital library. This requires a platform explicitly designed to collaboratively curate, organize, persist, and find content which helps resolve shared information needs. The platform should be easy to use and general enough to support communities in any domain (e.g., education, industry, and personal) and on any medium (e.g., PDFs, videos, and news articles), and it should be able to support general search and recommendation for efficient information discovery. However, there is currently no general infrastructure to support collaboratively creating such community digital libraries.

There are various academic projects which attempt to support collaborative information seeking [1–3, 6–8]. There are also various industry-based tools which support social navigation and collaborative information seeking, such as Pinboard[1], Hypothesis[2], Zotero[3], and Mendeley[4]. And more generally, personal note-taking Wikipedia-like services such as Obsidian[5] help support collaborative information organization. However, many of the aforementioned project and services are limited in scope (e.g., only for education or reference management), are limited in features (e.g., do not provide contextual search or general recommendation), are not collaborative, or are not available for use or open-source.

To fill in the gap, we propose the Community Digital Library (CDL). The CDL is a novel online and open-sourced social bookmarking platform designed to enable individuals of any community to collaboratively describe and save, search for, and discover online content. The target users of the CDL are any group of individuals who share a common goal and wish to collaboratively persist, recall, and discover online information relevant to the goal. As a personal information management tool, the CDL also enables users to bookmark, annotate, and organize useful webpages into multiple communities to facilitate retrieval and re-discovery. Additionally, the CDL itself is a general research infrastructure that can be conveniently used to study information-seeking behavior and user collaboration in a shared information environment. Finally, the open-source nature of the CDL promotes trust regarding data collection and allows individuals to contribute features according to their specific needs. Here, we describe the CDL's current set of

CSCW '23 Companion, October 14–18, 2023, Minneapolis, MN, USA

ACM ISBN 979-8-4007-0129-0/23/10…$15.00
https://doi.org/10.1145/3584931.3607495

[1] https://pinboard.in/
[2] https://web.hypothes.is/
[3] https://www.zotero.org/
[4] https://www.mendeley.com/
[5] https://obsidian.md/

features, various use cases, and research questions that we hope to answer in future studies. The CDL is accessible online[6] and the source code is publicly available on GitHub[7].

2 CDL DESCRIPTION

In this section, we describe the features of the CDL. The CDL consists of two main components: a website and a Chrome browser extension. The website is the initial point of access for the CDL user, and the extension can be downloaded from the Chrome Web Store.[8] The frontend is implemented with React and the backend is implemented with Python. The primary database is MongoDB, and both search and recommendation are done through OpenSearch.

2.1 Joining, Creating, and Submitting to a Community

After creating an account on the CDL website, a user has the ability to create or join a community. Once a community is created or joined, then the user can make submissions (i.e., bookmarks) to the community. A submission consists of a URL, a title, a description and/or highlighted webpage text, and a selected community. Submissions can be made via the Chrome browser extension or via the CDL website. In Figure 1, we depict the extension submission process. Here, the extension was opened on the "Memex" Wikipedia page[9] after the second paragraph on the page was highlighted. The user is now able to provide a title to the submission (e.g., summarize or describe how it is relevant to the community) and click "Submit". Clicking "Submit" will save the current URL, any highlighted text and description, and the user-provided title to the selected community (in this case, the community consists of students in a specific class, CS510 Spring 2023).

2.2 Searching over Communities

Once a submission is made to a community, any member of that community will be able to instantly search for and view the submission. Figure 2 depicts the CDL website after the user, who is a part of the CS510 Spring 2023 community, enters the query "memex". The top search result is the submission created after clicking "Submit" in Figure 1. Clicking the submission title will bring the user to the Memex Wikipedia page. Moreover, the user may provide relevance judgment feedback by clicking the "Thumbs Up" or "Thumbs Down" buttons below each search result. In the future, we plan on incorporating these judgments into the retrieval algorithms.

The user can search all submissions in their joined communities by using either the extension "Search" tab in Figure 1 or the website search bar in Figure 2. The user can also leverage hashtags (e.g., "memex #Lecture1.2") to filter submissions by a hashtag in the submission's title or description. Additionally, the user can restrict their search to specific communities by using the dropdown menu to the right of the search bar.

[6]https://textdata.org
[7]https://github.com/thecommunitydigitallibrary/cdl-platform
[8]https://chrome.google.com/webstore/detail/the-community-digital-lib/didjjbenidcdopncjajdoeniaplicdee
[9]https://en.wikipedia.org/wiki/Memex

2.3 Browsing and Interacting with Submissions

For additional submission-based actions, the user may click the three vertical dots at the top right of each search result. From there, the user has the ability to provide more fine-grained feedback (e.g., if the link is broken or the result is inappropriate). Or, the user can navigate to the submission's page, which shows the full submission text, provides the user with the ability to delete (if allowed) or share the submission to other communities, and displays the submission's statistics (i.e., clicks, views, and shares). The CDL also supports basic browsing of submissions. On the website, the user can select "Communities" from the top header, select a community, and they will be presented with a list of all submissions to that community by all members in reverse chronological order. Alternatively, the user can select "Submissions" from the top header and they will be presented with a list of all submissions that they have made across all of their joined communities, again in reverse chronological order.

2.4 Contextual-based Search, Recommendation, and Note-taking

Beyond traditional information retrieval, the CDL provides various contextual-based search and basic recommendation features. For example, when the extension is opened to the "Search Tab", the CDL provides the user with automatic search results by using the title and description meta-tags of the current webpage along with any highlighted text as a search query. Additionally, the landing page of the CDL recommends to the user either the most recent submissions made by others to their communities or submissions related to the content of their most recent submissions.

The submit functionality can be seen as a way to annotate and save webpages. Complementarity, the CDL website also supports basic markdown note-taking functionality independent of a specific webpage or community. By selecting "Notes" from the top header, the user can view, create, and save private hierarchical notes on any topic. The notes page also provides automatic search results based on the note line being editing, updating every 30 seconds by using the most recent edits made by the user as a query.

Figure 3 depicts both an example note page being edited and the respective automatic search results. The notes panel is split into three segments, from left to right: a hierarchy of note pages, editable markdown, and a page preview. Once the save button is clicked, the preview expands to replace the editable section. The automatic search results are displayed below the notes panel.

3 SMALL-SCALE EXPERIMENT AND USE CASES

During the spring of 2023, we deployed the CDL to a large computer science course on information retrieval at the University of Illinois Urbana-Champaign to facilitate collaborative learning. A community was created for the students, and the students were asked to submit, per lecture, a webpage that they found helpful for understanding the lecture's content. Moreover, they were asked to include a hashtag relating the submission to the corresponding lecture (e.g., "#L1.3", or "#L7.2") to enable efficient search of the saved content for a specific lecture by any student in the class.

The results of this deployment were positive: by the end of the semester, the students collected over 2,100 webpages, many of

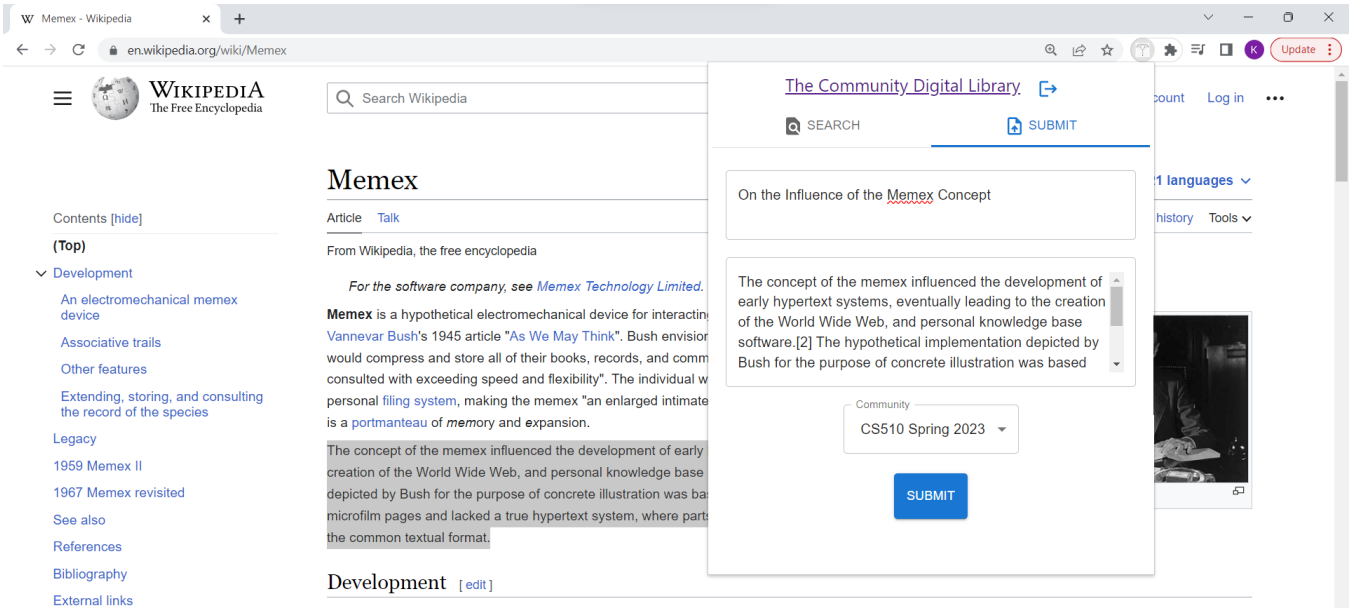

Figure 1: The CDL Chrome browser extension, opened on the "Submit" tab. When the user highlights webpage text and opens the extension, they can use the "Submit" tab to submit the webpage, title, and highlighted text to a selected community.

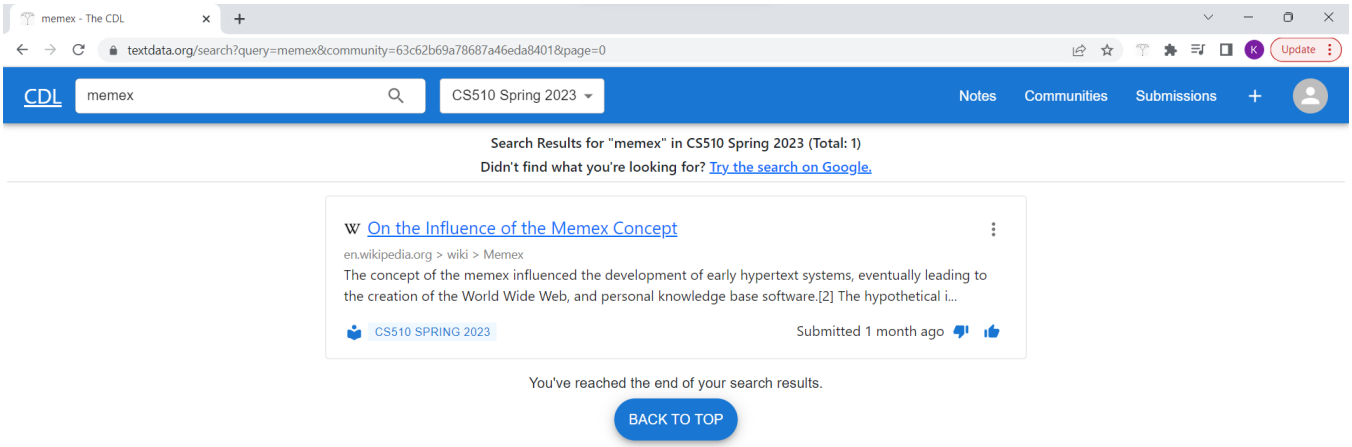

Figure 2: The CDL website, opened on the search results page. The user-entered query is "memex", and the top result is the submission created in Figure 1. The user-provided title becomes the hyperlink.

which were diverse, unique, and helpful for understanding lecture material. The saved pages were also useful for the instructor to use in the future to enrich the lectures. Moreover, multiple student groups built their course projects on top of the CDL infrastructure and the collected submissions. Project topics included integrating chat-based large language models, submission visualization, and tailoring recommendations for specific contexts. We have received IRB approval and numerous students' consent for analyzing the collected data, and we hope to study and report research results on our findings in the near future.

Beyond a classroom setting, the CDL supports many use cases. For example, the CDL can be used to support personal bookmarking and note-taking, enabling a user to collect and organize webpages in a private community. Or, the CDL can be used to support small-group collaboration in research or learning. For example, a Ph.D. student and their advisor can leverage the CDL to collect publications related to the student's thesis, and they can tag these publications accordingly (e.g., "#theoretical", "#experiment", or "#survey"). Similarly, a reading group can use the CDL to conveniently collect and annotate discussed papers. In the long run, we envision that the CDL may serve as a general framework for building and deploying next-generation office automation tools on the internet to improve productivity of many information-intensive tasks.

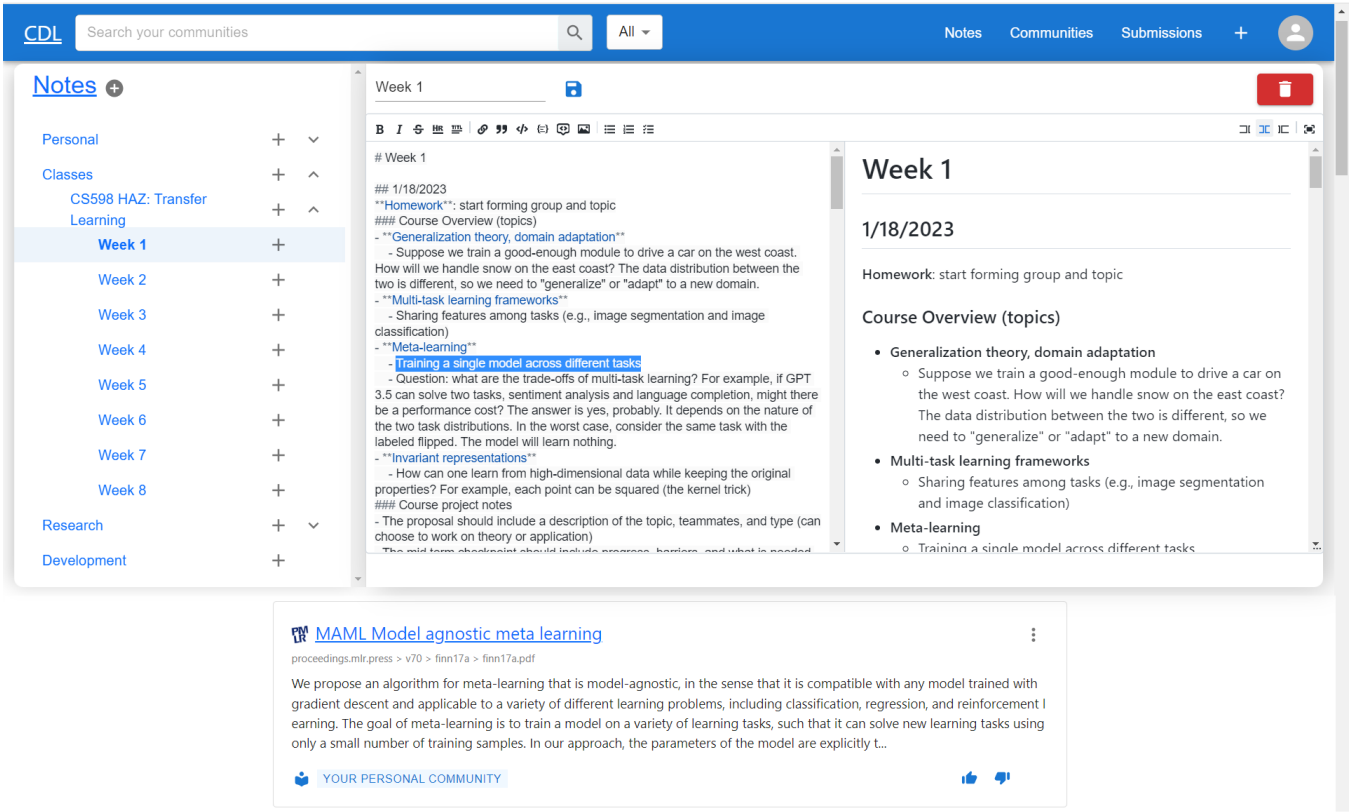

Figure 3: An example of a note page with an automatic search result. The highlighted line, "Training a single model across different tasks", is being edited, and the top automatic search result is about model-agnostic meta learning.

4 CONCLUSION AND FUTURE WORK

We proposed and developed the Community Digital Library, a novel and general platform that enables any community of users to work together to save, share, and discover online content. Due to its generality and extensibility, the CDL can potentially support numerous novel applications and opens up new opportunities for researchers to study novel paradigms of collaborative information seeking. The utility of CDL can be further increased by developing and adding novel algorithms for personalized search, contextual recommendation, and user interest prediction. The CDL can also be leveraged for studying how individuals collaboratively seek information, especially in educational settings. Finally, we hope that the open-source nature of the CDL encourages students, researchers, and developers to build on the code base and APIs, thus fostering a collaborative environment encompassing many use cases.

ACKNOWLEDGMENTS

We would like to thank all of the developers who have helped build the Community Digital Library. A complete list of developers and their respective contributions can be found on GitHub.[10] This work is supported in part by the Jump ARCHES program and the IBM-Illinois Discovery Accelerator Institute at UIUC, and by the National Science Foundation under Grant No. 1801652.

[10]https://github.com/thecommunitydigitallibrary/cdl-platform

REFERENCES

[1] Ran Cheng and Julita Vassileva. 2005. Adaptive Reward Mechanism for Sustainable Online Learning Community.. In *AIED*. 152–159.
[2] Rosta Farzan and Peter Brusilovsky. 2008. AnnotatEd: A Social Navigation and Annotation Service for Web-based Educational Resources. *New Review of Hypermedia and Multimedia* 14, 1 (2008), 3–32.
[3] Rosta Farzan and Peter Brusilovsky. 2018. Social navigation. *Social information access* (2018), 142–180.
[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9.
[5] Chirag Shah. 2014. Collaborative Information Seeking. *Journal of the Association for Information Science and Technology* 65, 2 (2014), 215–236.
[6] Diana Soltani, Matthew Mitsui, and Chirag Shah. 2019. Coagmento v3. 0: Rapid Prototyping of Web Search Experiments. In *Proceedings of the 2019 conference on human information interaction and retrieval*. 367–371.
[7] Julita Vassileva, Ran Cheng, Lingling Sun, and Weidong Han. 2004. Stimulating User Participation in a File-sharing P2P System Supporting University Classes. *P2P Journal* (2004), 14–23.
[8] Alan Wexelblat and Pattie Maes. 1999. Footprints: History-rich Tools for Information Foraging. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 270–277.