ISSN 2161-3974
Full-Length Paper

Journal of eScience Librarianship 13 (1): e804
DOI: https://doi.org/10.7191/jeslib.804

Journal of eScience Librarianship
putting the pieces together: theory and practice

Open Science Recommendation Systems for Academic Libraries

Lencia Beltran, Carnegie Mellon University, Pittsburgh, PA, USA, lbeltran@andrew.cmu.edu
Chasz Griego, Carnegie Mellon University, Pittsburgh, PA, USA
Lauren Herckis, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

An interdisciplinary academic team offers a comprehensive case study describing the development of a predictive model as the cornerstone for an open science recommendation system tailored to the Carnegie Mellon University community. This initiative will empower users in choosing open science services that align with their academic requirements, introduce academics to resources they find valuable, and bridge gaps within academic library service offerings.

As an institution with a longstanding commitment to a science-informed approach and a focus on computer science, engineering, and artificial intelligence, Carnegie Mellon University has enthusiastically embraced open science practices. The Carnegie Mellon University's Libraries has been instrumental in bringing these practices into our academic landscape.

Received: October 29, 2023 **Accepted**: February 5, 2024 **Published**: March 5, 2024

Keywords: open science, artificial intelligence, AI, recommendation system, higher education, academic library services, ethical considerations

Citation: Beltran, Lencia, Chasz Griego, and Lauren Herckis. "Open Science Recommendation Systems for Academic Libraries." *Journal of eScience Librarianship* 13 (1): e804. https://doi.org/10.7191/jeslib.804.

Data Availability: Beltran, Lencia, Chasz Griego, and Lauren Herckis. 2024. "Open Science Recommendation Systems for Academic Libraries." OSF. https://doi.org/10.17605/OSF.IO/PX6HJ.

Abstract continued

The authors strive to develop a predictive model which will evolve into a recommendation system. The pursuit of this endeavor has led the authors through several ethical considerations, such as data privacy, the involvement of student contributors, and the design of a persuasive recommendation system. We are committed to exploring ethical approaches for delivering user-centered recommendations and to preserving individual autonomy.

The authors have actively engaged with diverse academic departments, students, and faculty, embarking on data exploration, and applying open science principles throughout the process. The resulting system will raise awareness of library services and deliver tailored recommendations for the adoption of proven research tools and practices.

This case study serves as an exemplar of how universities can enact open science principles and develop systems that prioritize the user's interests, navigate institutional complexities to forge interdisciplinary collaboration, and muster resources to support innovative, multi-disciplinary efforts.

Introduction

A Carnegie Mellon University team aims to build a predictive model to act as a foundation for the development of an open science recommendation system for the campus community. This model will employ user characteristics to identify services that are a good fit for users' academic needs from a universe of the Library's Open Science resources. Developing this model will help us understand which users engage with each service and identify potential users who would benefit from additional Open Science resources. The recommendation system will introduce naive users to services they are likely to find valuable and, in parallel, introduce current users to alternative capabilities that are likely to be of value. At the same time, our team will consider and speak to the ethical implications of generating a novel system of this kind in an academic setting. This system will also serve as a proof of concept for other academic library service recommendation systems.

Project Details

Carnegie Mellon University (CMU) is a private, global research university that has championed a science-informed approach for more than five decades and is consistently ranked among international leaders in computer science, engineering, and artificial intelligence. Nobel laureate, Turing award winner and father of artificial intelligence Herb Simon created interdisciplinary pathways for research and innovation that are still characteristic of Carnegie Mellon University's unique academic ecosystem today. Today, Carnegie Mellon University offers degrees with a focus on artificial intelligence at the bachelors, masters, and doctoral levels. A transdisciplinary effort focused on artificial intelligence, CMU-AI unites students, faculty, and

staff from all areas of the university to engage with complex challenges and to partner with corporations, non-profits, and research institutions around the world. In addition to extensive research and development efforts, CMU actively fosters entrepreneurship. This has resulted in a number of spin-off corporations which embed CMU-developed, AI-enabled innovations ranging from self-driving cars to smart textbooks.

This case study is authored by the project team, Lencia Beltran, Chasz Griego, and Lauren Herckis. Lencia Beltran is the Open Science Program Coordinator for Carnegie Mellon University Libraries Open Science Program. Her educational background is in Linguistics, Speech-Language Pathology, and Librarianship and Archival studies. Beltran received training in data science from Drexel University's LEADING and AI applications from the IDEA Institute on AI. Her research falls within the spheres of AI, geospatial mapping, social networks, and implications of technology on language, including diversity, identifying, and belonging in higher education and academia. For this project, Beltran supported project establishment, initiation and management, research design, documentation, and the building of institutional collaborations. Dr. Chasz Griego is a Science and Engineering Liaison Librarian and formerly an Open Science Postdoctoral Associate at the Carnegie Mellon University Libraries. His educational background is in Chemical Engineering, with a focus in computational chemistry and catalysis. His doctoral work focused on physical models coupled with machine learning to expedite catalyst screening projects. His research focuses on the influence of open science tools on reproducibility in computational research related to AI, simulations, and modeling. For this project, Dr. Griego supported research design, data curation, and technical recommendations for model development. Dr. Lauren Herckis is an anthropologist by training and has a faculty appointment in the University Library and the School of Computer Science's Human-Computer Interaction Institute. Her research explores the adoption and use of AI-augmented and collaborative learning tools, the digitalization of higher education, and the design of tools to help faculty employ effective technology-enhanced learning tools with fidelity. For this project, Dr. Herckis supported research design and data analysis, and co-developed strategies for tool deployment, data curation, service delivery, and evaluation, as well as facilitating partnerships with institutional collaborators.

Background

Carnegie Mellon has championed a science-informed approach for more than five decades and is committed to designing and facilitating transformative educational experiences, accelerating research and creative inquiry, developing innovative library infrastructure, and evolving to enable students, staff, and faculty to discover, access, and use scholarly information. Core project team members are affiliated with the University Library and have a professional interest in enhancing Library services. Carnegie Mellon has invested in an Open Science Program in recent years, and project personnel are both personal and professional champions of open science practices.

The proposal of a recommendation system derived from the idea to create a predictive model that would shed light on usage patterns of open science services. In 2021, members of the Open Science team ran an

analysis to evaluate the program's impact, using data collected over the span of two years, which included service offerings and, in most cases, high-level user information like departmental affiliation. The analysis of the preliminary data showed the effectiveness of the program in the campus community, yet it did not provide further details on why users opted for specific services and their motivations behind those decisions.

In an effort to understand our users and their motivations, we pursued the next step to build a predictive model. The predictive model will not only give us insight into these essential areas but has the potential to identify probable user groups who would benefit from open science resources. The findings from the predictive analysis can lead to discussions on where gaps in our services lie in terms of which departments are not using our services so that we can begin to develop resources to meet the needs of those departments.

The inspiration for developing a recommendation system arose as we considered strategies for simplifying the discovery process of our services to users. The recommendation system built from this predictive model will extend how the Open Science Program delivers services and how users discover these services. This recommendation system will facilitate how information is accessed/retrieved and alleviate information overload that students, as well as faculty and staff, may feel as a product of having too many options and not knowing which services will be the most helpful.

The mechanics of the recommendation system will work similarly to other well-known models, like Pinterest, Amazon, and Netflix, by providing users with a curated list of results. How the services are delivered to users is an aspect we are thinking through. As we move forward and look to other projects for guidance, we intend to keep our users at the center of each approach.

Members of the Open Science Program in the University Libraries have spearheaded the effort to implement this recommendation system. The Libraries serve the efforts of the University to continually innovate education and research by supporting the curriculum as well as faculty and student research. One area in which the Libraries are leading is in innovation around open science, a fairly new concept in the United States. Our Open Science Program has helped propel the integration of many open science elements into the education landscape of our community. This team is composed of several faculty and staff in the Libraries, and the members of this program who are actively involved in this project include Lencia Beltran, the Open Science Program Coordinator and Chasz Griego, a Science and Engineering Librarian who was formerly an Open Science Postdoctoral Associate. Along with these associates, Lauren Herckis, an anthropologist and affiliate of the Libraries, Simon Initiative, and the Human Computer Interaction Institute at CMU, has contributed in our efforts to identify collaborators and develop strategies to assess how users will engage with open science service and tool delivery in educational and research settings. We also recruited an undergraduate student, Zhijin Wu, majoring in information systems, human-computer interaction, and business administration, who is volunteering their efforts as a project manager and coordinator to gain academic and professional experience.

The Open Science Program offers many resources, which include data and research consultations and Libraries workshops in data and software Carpentries. Along with these resources, our patrons also have access to several tools and platforms that facilitate open science practices. These include KiltHub, our institutional repository, LabArchives, a platform for electronic lab notebooks, protocols.io, a repository for sharing records of research methods, and Open Science Framework, a platform for collaborative research management. With such resources and tools in place, Open Science Program personnel have initiated an ambitious plan to amplify engagement at Carnegie Mellon. In 2021, the CMU Open Science team, including past and present members, explored and gathered user data dating back to 2019 of over 900 users at CMU that counted usage and numbers of items uploaded on digital platforms as well as interactions with our other tools and services (Wang et al. 2019). This dataset has driven initial insight into our efforts to establish an open science recommendation system in the Libraries.

In the spring 2023 semester, we partnered with a faculty-led team of four Master of Statistical Practice (MSP) capstone students who agreed to undertake data exploration and develop a proof-of-concept AI-enabled predictive model that would identify likely use cases for open science tools and resources at CMU as a Master's degree capstone project. Currently, this team has described the distribution of past open science tool users among schools and departments at CMU, with faculty and Ph.D. students being the most common academic positions held by users. As these efforts continue, this team is helping us identify the features of our current dataset that will provide statistically significant predictions. Before establishing this partnership, we requested, as one of the deliverables, a written report describing the development, decision-making process, outputs, and findings so that we, including others, could reproduce their work. In keeping with our posture of openness, we also asked the team to apply open science practices such as reproducible tools and code and version control. The students have since shared their research materials, including documentation of data, code, and analysis using open platforms like Open Science Framework. In addition to the existing usage data for open science tools and services, a subset of the metrics data for students along with de-identified demographic information from the CMU Registrar, and TartanDataSource (TDS) from the University Institutional Research and Analysis Office were used in the analysis, all of which can be shared. As library and information professionals, among other titles, we understand the significance of an individual's right to privacy and, as we carry on, preserving this right will be at the forefront of our mind. As long as our research materials do not contain personally identifiable information that is unable to be anonymized, our team plans to share any code and scripts, documentation, and other information openly since one of our objectives is for this system to serve as a blueprint for other academic libraries.

In order to design service models that accommodate recommendations from our system and accurately meet the needs of the campus community, our team is also investigating perspectives from researchers and implementing open science tools into educational settings. Through 2023, the Libraries Open Science Program is conducting a needs assessment and environmental scan that includes focus group interviews of research with diverse areas of study and identifying open science services and practices among peer

institutions and other units at CMU. In summer 2023, Chasz Griego will lead an eight-week undergraduate course, hosted through the Office of the Vice Provost for Education, delivering opportunities for students to use open science tools and assess how these tools influence collaboration and reproducibility in research. This course will serve as a testing space to investigate how open sciences tools, and practices, can be implemented in practice within educational settings at CMU.

Our findings from these assessments and collaborations will aid in guiding us to develop a proof-of-concept recommendation system that will introduce naive Carnegie Mellon University students, faculty, and staff to existing open science tools and resources. Developing a predictive model and associated recommendation system is a novel approach to scaling adoption of, and engagement with, academic tools at a university like Carnegie Mellon. Eventually, we will need to instantiate a functional version of the predictive model-driven recommendation system so that it can begin effectively delivering resource recommendations to educators. The work proposed here will meet this substantial challenge and make successful implementation possible.

Ethical considerations

Already within our preliminary exploration, our team has identified ethical challenges that relate to developing both a predictive model and a recommendation system in three areas. First, many of our concerns connect to users' rights, such as using personally identifiable information and privacy. Second, this project leverages student labor in exchange for academic credit and learning opportunities. Finally, this project is designed to promote specific tools and practices through persuasive design.

Data collection for our system will include personally identifiable information that can be used to describe patron behavior with respect to Library tools and services. Generally speaking, predictive and recommendation systems take information about a user's preferences as input and predict an output of an item that is likely to meet the user's needs. As a result of the underlying nature, the collection and curation of vast amounts of personal information are inevitable for generating personalized recommendations. On the surface, these systems appear to be user-centered, because they generate curated content, but many of them are driven by business objectives and applications. Consequently, this leads to less consideration of the user and their privacy. More often than not, user data is being collected and analyzed without the consent or knowledge of the user. If users are aware data is being collected, then it is likely they do not understand its actual or intended uses. In our pursuit, we are seeking approaches that will allow us to design a recommendation system that curates open science resources, takes into account the users' rights, and carefully balances the risks of user privacy and accuracy, as well as fairness and explainability without merely shifting the responsibility to the users. For example, a solution might be to embrace a macro-ethical approach which considers ethical problems related to data, algorithms, and practices and how the problems relate, depend on, and impact each other (Milano, Taddeo, and Floridi 2020).

This project currently entails collaboration with students individually and in teams and will likely expand in the coming year to include other graduate and undergraduate students working for credit, hourly, and for free. Students regularly engage in generative activity as part of their educational experiences, and there is substantial literature about the effective design of capstone courses (Tenhunen et al. 2023) and recent literature also addresses the ownership and intellectual property associated with work that students have completed in the course of their education (Allen 2021). As this case study is being written, an undergraduate triple-major in information systems, human-computer interaction, and business administration is volunteering approximately five hours per week to serve as a project manager on this project. This student has a background in AI research and development and an interest in developing project management skills. In order to ensure that this student is gaining useful professional and/or academic experience, we worked with her to identify learning outcomes and desired skills and to agree on mutually agreeable communication and collaboration strategies. We have asked her to create project plans and visualizations, such as Gantt charts, maintain records and manage communications. In order to ensure that data and products of work are handled ethically, we used a collaboration agreement and discussed the need for explicit communication about future use of project assets. We expect that she will use visualizations and other assets as part of her portfolio. This student will gain substantial educational benefit through the hands-on learning experiences that our collaboration requires. The project will gain several durable assets which will outlast the student's collaboration on the project.

This project was integrated into the Master of Statistical Practice (MSP) graduate curriculum in the Spring of 2023. A student team working under faculty supervision undertook data exploration, developed project constraints and documentation, and built a proof of concept predictive model that met our specifications. The project team is positioned as a client, and related student efforts will be evaluated and graded as a capstone project to meet Masters degree requirements. Faculty associated with the capstone course and project will guide student work and frame the experience to best serve students' educational goals. While these efforts represent curricular and educational benefits to the students, they can also be understood as an appropriation of student labor to produce university assets. Development of a predictive model is a non-trivial task which requires substantial investment of time and resources. These resources are being extracted from students as a component of requirements for degree completion and can be understood as appropriation of student labor.

Following the pattern of other recommendation designs, our system seeks to help users discover new services and minimize the cognitive information overload that exists in academic settings. Yet we are grappling with the inherent persuasive design of recommendation models. How can we build a model that does not invite the undue or unwanted influence of library services or introduce bias but ultimately is helpful and protects the users' autonomy? Fortunately, there are a number of different approaches we can explore for building a recommendation system, yet they all generally involve constructing a user model or profile. A user profile is a set of characteristics and/or preferences for a given user and is used by the

system to make personalized recommendations. Although we do want users to receive curated services, these constructions can limit the range of options recommended to users as it places them into categories (e.g., department, academic level, etc.). As a result of this algorithmic classification, an individual's ability to make self-driven and reflected decisions on which services are extended to them is hindered, and ultimately users are nudged toward a particular outcome. Being that the area of Artificial Intelligence is still in its infancy, there is extensive research that explores the ethicality of recommendation systems and their effect on citizens as a by-product. Some research suggests strategies to assuage bias and quell autonomy issues, such as deploying a conversational recommender system that provides users with explanations for why a particular recommendation was made (Musto et al. 2019). Musto et al. (2019) found the selections were better received when explanations were provided to users.

This general description illustrates how these systems can potentially shape an individual's experience of the digital world. As echoed throughout, our intent is to aid in facilitating the discovery of these tools within our community, which may be beneficial for their academic and personal goals, rather than participate in influencing the choices or altering the perception of what services are readily available to them. At this current stage, we have more questions than answers. The ethical challenges presented here, including others we encounter as we move forward, shape how we approach conversations with individuals who have experience with designing and deploying predictive and recommendation systems.

Who is affected by this project?

Over the course of the project, we expect many individuals, services, and programs to be involved to some degree. The Libraries Dean has been the essence of support for open science practices in our community, including this project. In 2018, he endorsed the development of the Open Science and Data Collaborations Program spearheaded by three library liaisons. As the highest level of support within the Libraries, he has helped advocate for the many open science resources we offer and paved the way for our Open Science team to hold discussions and establish relationships with Deans from the Schools/Colleges. Through these many conversations, we have already seen an uptick in interest in open science practices and services from disciplines (e.g., Language Technologies Institute and Statistics) across campus. Since the Open Science Program was initiated by three library liaisons it has helped increase internal support from other library service providers (e.g., functional specialists, liaisons, and staff). The support from library service providers has been valuable for raising awareness about our services and building internal and external partnerships as each person engages with distinct departments and individuals on campus.

As implied to some degree, the Open Science team embodies a range of specialists whose work is diverse. Our team includes a staff who manages and supports the Institutional Repository, a functional specialist who provides training and support on data curation and literacy, and three library liaisons who support tools and provide training on topics related to open access, data management, scholarly communications, and more. Their involvement in our project is indirect but essential for the ongoing success of the program

and the resources we can offer our community. Our project has many moving parts, including partnering with faculty and educators to integrate our services into their curriculum, which will discursively involve the students who engage with the open science resources through their professors' class transformation.

Altogether our project can increase awareness of library services, begin to glean user motivations, and open up new approaches for gathering information on how users practice open science at an institution and measure satisfaction/success with each tool. As our project unfolds, there are potential opportunities to partner with University service areas, including the Student Academic Success Center and the Eberly Center for Teaching and Excellence. The Student Academic Success Center facilitates student learning by providing academic coaching, subject-specific tutoring, effective communication strategies, accommodations for students with disabilities, and language support for multilingual learners. The Eberly Center supports faculty, graduate students, and other educators that aim to design courses and curricula that put students at the center of the teaching process. Our effort will enhance our understanding of Carnegie Mellon Library service use and provide more effective open science support to the Carnegie Mellon community. More broadly, this project can serve as a proof of concept for other academic libraries, which we hope will build on our work.

Lessons learned and future work

While this project is still in preliminary stages and much of the work is ongoing, we have already learned several lessons to improve our approach to create a robust recommendation system while considering the ethics of data usage and implementation of AI. Many of these lessons were learned through our work with the Master of Statistical Practice (MSP) capstone team. Collaborations with Statistics faculty and students first revealed challenges for us to communicate our goals and intentions in a way that aligns with the knowledge base of these subject specialists. We chose terminology that helped translate our goals into action items that could reasonably be executed by statistic students. However, we did observe some disconnects with information exchange. For instance, the students treated the variables in our dataset as arbitrary. When analyzing trends such as academic departments that are more likely to use our institutional repository, Kilthub, the students would tend to focus on how these trends contribute to model parameters and not question the reasoning behind why a certain department would be more drawn to Kilthub. These points were usually addressed during team meetings where the faculty adviser led the efforts to ask the more subject-specific questions. Overall, the statistics students successfully applied their education to real-world problems and data, but the team encountered challenges connecting the data analysis to the context of the problems that were specific to the university libraries.

The Master of Statistical Practice (MSP) capstone team delivered preliminary results that signaled a need for a larger dataset so that the predictive model can deliver results with higher confidence. Challenges arose when considering ways to obtain expansive data that represents Library users at CMU. A major consideration is the privacy of individuals currently or previously affiliated with the university. While our aim is to develop

user profiles that describe prior or potential attraction to open science practices, we plan to only rely on user information that is publicly or internally available. However, there are challenges specific to library user data. Within the rights of the Fourth Amendment of the United States Constitution, the privacy of patrons that access information from libraries is protected. For instance, these laws protected patrons from seizures of library records from the Federal Bureau of Investigation under the USA PATRIOT act. Appropriately, the CMU libraries does not record user-specific circulation information, which teaches us that a predictive model built for library services cannot be established with such data, but other, carefully considered records that describe academic behavior and motivations.

Future work will address approaches to evaluate the effectiveness of recommendations to users. We will develop strategies to assess user responses to the recommended tools or services and how they influence research and/or educational outcomes. This will include developing metrics to measure how the recommendation system supports decision making. To evaluate the performance of decisions, we can refer to the strategies outlined by Jameson (2015) that identify choice patterns based upon attributes, consequences, experience, social conditions, policies, or trial-and-error (Jameson et al. 2015). We will survey responses from users in a variety of settings including electronic surveys, focus groups, and case studies. In case studies we will analyze changes in educational outcomes in academic courses that incorporate recommended open science tools and services.

Data Availability

Many of the materials mentioned within the case study can be found on our Open Science Framework project, Open Science Recommendation Systems for Academic Libraries (Beltran, Griego, and Herckis 2024). Please reach out to our team if you have any questions.

Acknowledgements

The research case study was developed as part of an IMLS-funded Responsible AI project, through grant number LG-252307-OLS-22.

Competing Interests

The authors declare that they have no competing interests.

References

Allen, Genevera. 2021. "Experiential Learning in Data Science: Developing an Interdisciplinary, Client-Sponsored Capstone Program." *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, March, 516–522.

Beltran, Lencia, Chasz Griego, and Lauren Herckis. 2024. "Open Science Recommendation Systems for Academic Libraries." OSF. https://doi.org/10.17605/OSF.IO/PX6HJ.

Jameson, Anthony, Martijn C. Willemsen, Alexander Felfernig, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. "Human Decision Making and Recommender Systems." In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Rokach, and Bracha Shapira, 611–648. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7637-6_18.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. "Recommender Systems and Their Ethical Challenges." *AI & SOCIETY* 35 (4): 957–967. https://doi.org/10.1007/s00146-020-00950-y.

Musto, Cataldo, Fedelucio Narducci, Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2019. "Linked Open Data-Based Explanations for Transparent Recommender Systems." *International Journal of Human-Computer Studies* 121 (January): 93–107. https://doi.org/10.1016/j.ijhcs.2018.03.003.

Tenhunen, Saara, Tomi Männistö, Matti Luukkainen, and Petri Ihantola. 2023. "A Systematic Literature Review of Capstone Courses in Software Engineering." arXiv. http://arxiv.org/abs/2301.03554.

Wang, Huajin, Melanie Gainey, Patrick Campbell, Sarah Young, and Katie Behrman. "Implementation and assessment of an end-to-end Open Science & Data Collaborations program [version 2; peer review: 2 approved]. *F1000Research* 2022, 11:501. https://doi.org/10.12688/f1000research.110355.2.