OCUL Ontario Council of University Libraries

OCUL Machine Learning/Artificial Intelligence Report and Strategy

INTERIM REPORT

March 2024

The purpose of the Interim Report from the *Ontario Council of University Libraries (OCUL) Task Force on Machine Learning/AI* is to present the current work and deliberations of the Task Force, and to encourage engagement in reviewing and critiquing the ideas and actions proposed.

Feedback on this report is welcome via email to ocul@ocul.on.ca or via an online form by March 19, 2024.

This report describes use cases for machine learning relevant to the OCUL consortium and recommends projects utilizing machine learning technologies. It also considers key contextual issues such as ethical concerns, technical capacity, available expertise, and infrastructure needs. All sections are drafts with some sections more fully developed than others.

This report uses “AI/ML” as a short form for artificial intelligence and machine learning technologies and applications. It signals that while the term AI is in common usage, it is generally machine learning techniques that are being considered. Unsurprisingly, this report has been assisted with the use of ChatGPT.

Table of Contents

1. Introduction

The goal of the *OCUL Machine Learning/Artificial Intelligence Report and Strategy* is not to provide a tutorial regarding machine learning or to review the literature of AI/ML and research libraries, but rather to:

- provide a compelling case for collaborative action,
- identify relevant opportunities, issues, and challenges,
- present and validate alternative action strategies, and
- outline a process for strategic decision-making for OCUL.

An OCUL Task Force, with the following membership, was created to undertake this work (see Appendix 1 for Terms of Reference):

- Mark Asberg, Queen's University
- Talia Chung, University of Ottawa
- Scott Gillies, Wilfrid Laurier University
- Vivian Lewis, McMaster University
- Mark Robertson, Toronto Metropolitan University
- Catherine Steeves, Western University (Chair)
- Amy Greenberg, OCUL
- Kate Davis, Scholars Portal
- Michael Ridley, University of Guelph

The Task Force recognizes that a reasonable option for OCUL is to do nothing collectively regarding machine learning while individual members chart their own directions and actions. However, the premise of the *Report and Strategy* is that machine learning will have a transformational impact on research libraries that is best understood, managed, and realized if OCUL libraries work together on a select number of core initiatives. All the proposed strategies move OCUL and member libraries from passive consumers of machine learning and machine learning-based products and services to active participants in evaluating, using, and developing applications. While transformative, machine learning is not without serious limitations and concerns; the path OCUL chooses must respect and advance the values and principles of academic librarianship and the academy.

The core objectives of this report and strategy are to:

- identify AI/ML use cases that can be implemented as operational projects for the benefit of library users, library staff or both,
- implement these projects as consortial, collaborative initiatives,
- utilize these projects and related work to build capacity in libraries regarding AI/ML awareness, knowledge, and technical skills, and
- enable transformational change to further the mission of libraries.

A final report will be presented to OCUL in May 2024.

2. Consultation and Review Process

The development of the Interim Report was guided by the mandate and deliberations of the Task Force. The process included a review of current developments and best practices, and broad consultations within OCUL and with experts from a variety of constituencies (see Appendix 2 for a list of individuals consulted as well as discussions at relevant meetings and conferences).

The many ideas and issues gathered were synthesized and prioritized into the provisional set of use cases and projects presented here. The descriptions of some of these are very preliminary, and many require further details. They are presented to indicate interest and to spark discussion. As deliberations continue, the specifics of individual projects will be refined.

An important part of the next consultation phase of this initiative is to engage with other research libraries, consortia, and related organizations.

Feedback from this Interim Report, the associated online Summit to be held March 20, 2024, and other consultations will inform the final version of the report to be presented to the OCUL Directors in May 2024.

3. Collaboration

This Interim Report outlines an OCUL perspective on the use of AI/ML in a consortial setting. While it arises primarily from OCUL conversations and deliberations, the release of this report is intended to broaden the dialogue and encourage wider collaboration. While some of the projects are specific to OCUL, others have application beyond OCUL and will be of interest to other consortia and academic libraries in Canada.

The Task Force welcomes opportunities to work with others to find common ground and develop collaborative projects.

4. Guiding Principles

The development and use of AI/ML arising from this report and strategy will be informed by a set of principles that align with the overall values of academic libraries generally and OCUL libraries specifically.

The OCUL Operating Principles indicate that OCUL projects and initiatives will a) encourage new ideas, enable collaboration, and build consensus, b) balance risk and opportunity, and c) comply with the strategic plan and OCUL's core principles. The latter highlight, advancing research, teaching, and learning; leadership in social change and inclusivity; developing and supporting robust infrastructure; and demonstrating value.

Particularly valuable among the documents outlining key principles regarding AI/ML are the UNESCO report, AI and Education: Guidance for Policy-Makers (2023), the UK Russell Group report on the use of generative AI tools in the academic setting, the National Library of the Netherlands article, AI in Libraries: Seven Principles (2020, and the Cornell University report, Generative AI in Academic Research (2023).

From these and other deliberations, the following principles are identified as important:

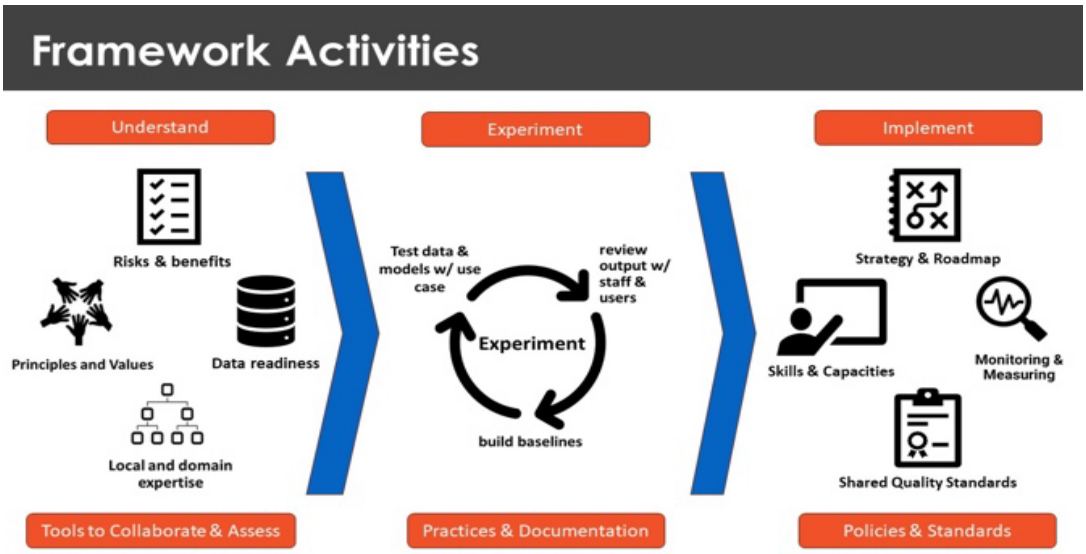
- emphasizing collaboration
- designing and implementing with humans-in-the-loop
- designing and building around diversity, equity, and inclusion
- enhancing accessibility
- advancing environmental sustainability
- ensuring transparency and explainability
- aspiring to adopt open-source tools and solutions

5. Program and Project Management

The nature and complexity of this strategy suggests the need for both program and project management. A Program Manager will oversee the overall set of projects (each guided by its own Project Manager) and coordinate among the various internal and external groups. This role requires administrative experience and could be provided via secondment from an OCUL library.

Each project will require a project team consisting of expertise in the appropriate technology, the service or resource domain, and project management expertise. These could be sourced from Scholars Portal (the digital service arm of OCUL), OCUL libraries, OCUL institutions, or beyond. Scholars Portal and OCUL libraries have a wealth of expertise in library applications and services, IT systems, and project management. However, each project will have different requirements and necessitate a management structure appropriate to those needs. The Task Force will work with the OCUL Executive and Executive Director to begin scoping and drafting requirements for these positions

The Library of Congress report, LC Labs Artificial Intelligence Planning Framework (2023) (with additional detail and tools on the LC GitHub site), provides a valuable "planning framework to support the responsible exploration and potential adoption of AI." The framework identifies three planning phases (Understand, Experiment, and Implement) with each phase evaluating the core elements of machine learning (Data, Models, and People). The LC Labs framework offers the AI/ML projects identified in this report and strategy guidance on criteria, process, evaluation, and oversight.



1An Overview of the LC Labs AI Planning Framework

6. Capacity Building

One of the observations arising from the discussions with OCUL members has been a reduction in technologically focused staff in OCUL libraries. OCUL libraries lack programmers, system designers, application developers, and other expertise necessary to explore and implement new technologies. Staff expertise in even the basic technical underpinnings of AI/ML tools and services is rare.

To a certain extent, these technical resources have moved to Scholars Portal or to central university IT. However, Scholars Portal staff (and central IT staff) are linked to specific services and resources and are generally not available for exploratory work or to assist with local library initiatives. Building technological capacity back into OCUL libraries should be a central objective arising from this Task Force. Specifically, libraries must find the means to undertake exploratory IT work that may lead to transformative changes in the way libraries operate. Such capacity provides libraries with the independence to chart their own exploration and direction, and to contribute to the academy more effectively.

The need is not just for specialized technology staff. AI literacy training for library staff is an active discussion on *AIMLComm*, the new Community of Practice (see below). From this, models and materials for staff training that can be implemented on a consortial level are expected. However, library staff in many roles need to move beyond AI literacy training to embrace increasingly levels of technical competencies and policy expertise.

Training for information professionals interested in gaining greater machine learning technical expertise (in the context of library applications) is sparse at best. There is an opportunity for OCUL to coordinate a machine learning training program that would raise the expertise in this crucial area and to develop a cohort of staff who could use their skills locally and consortially. Discussions have begun with university CIOs (through CUCCIO) to explore possible joint training programs.

6.1. Community of Practice: AIMLComm

In consultations across Canada regarding AI and machine learning in libraries, a consistent theme emerged: a desire for people to engage and learn collaboratively with their peers. In response, a new online Community of Practice (CoP) was launched in January 2024 with the objective to build capacity within Canadian research libraries to better respond to the challenges and opportunities of this technology.

The CoP, *AIMLComm*, is directed by moderators and guided by a code of conduct that models acceptable behaviour on the site. As with all communities of practice, the participants determine the nature and direction of the discussions. General topic areas include public services, technical services, and systems with participants drawn from a variety of backgrounds and interests. Additional topics will emerge as the discussions deepen and broaden.

AIMLComm uses an instance of the Mattermost software hosted and maintained by Scholars Portal. At the end of February, *AIMLComm* had over 280 members. Those interested in participating are encouraged to visit the registration page.

AIMLComm and the Virtual ML Lab (see section 8.3) are complementary initiatives. Discussions in *AIMLComm* could lead to project groups utilizing the Lab for specific applications. Similarly, project groups may use the CoP to engage an interested community.

6.2. Hackfests

On February 23, 2024, OCUL hosted a Whisper Hackfest at the Sherman Centre, McMaster University Library. Whisper is an open-source model for transcribing audio to text (see section 14.2 for this use case). A total of 35 people participated representing 13 of the 21 OCUL libraries. Staff from Scholars Portal and the OCUL office also participated. The central objective was capacity building regarding this technology and use case among a group of library staff with diverse skills, experience, expertise, and interests.

Participants worked together in various groups to implement Whisper, test against sample audio or video, advance techniques such as user-friendly processes for transcription and diarization (speaker identification), explore georeferencing output, and

video captioning. Groups also discussed policy issues, accessibility options, project management, concerns about bias and inaccuracies, and the importance of guardrails.

The Whisper Hackfest was widely viewed as successful with participants from all backgrounds and expertise finding the day valuable. OCUL was encouraged to host future hackfests on different use cases or issues. Suggestions included building large language models (LLM), testing scripting in different models, data curation, metadata creation, and virtual reference.

7. Machine Learning Expertise

Specific technical expertise in machine learning (ML) is an area where Scholars Portal and OCUL have limited experience and knowledge. Strategies to provide ML expertise are critical to the success of the projects. There are several viable alternatives to access this expertise, none of which are mutually exclusive. It is also important to note that should collaborative projects emerge involving libraries and organizations beyond OCUL this expertise could be shared and jointly funded.

Employee: hire an ML developer at Scholars Portal. This person would be the technical lead for all the selected projects. Such developers are in demand and expensive. Limited-term contracts (which may be all OCUL can commit to) will be viewed as undesirable. Not only would an ML developer on staff provide the technical foundation for the key projects, but they would also be an important resource to the training program mentioned above.

Consultant: contract with an ML developer as a consultant/advisor. This is somewhat less expensive than a full-time hire. However, there is more flexibility in terms of engagement and specific responsibilities. It is possible OCUL could cost-share an ML developer with another organization or research group.

Technical consultants/advisors are increasingly available from agencies and service providers. The challenge for this option would be to contract for the specific technical needs of the active projects and to ensure continuity and sustainability in the projects. This option should be used selectively and in limited situations.

Post-Docs: post-doctoral researchers with ML expertise could be attracted, through the LIS schools, to contribute to specific projects. This may be an area where shared funding (LIS schools and OCUL) would advance common initiatives. The policies and practices regarding post-doctoral appointments would need to be reviewed to determine the suitability of this option.

Visiting Researcher Program: the new OCUL Visiting Researcher Program could provide an opportunity to engage librarians with specific projects arising from the OCUL AI/ML strategy. This program could proactively seek librarians interested in developing AI/ML research agendas but providing research questions or issues in need of investigation.

Students: engage graduate and undergraduate students with ML expertise. Students from various programs (e.g., computer science, information science) could be involved with specific projects for a limited duration. These arrangements could be internships, practicums or short term, part-time hires. Assessing expertise, coordinating recruitment, managing work, time constraints, and arranging funding and/or course credit would be some of the complications. However, giving students real-world work experience would benefit all parties.

Global Colleagues: engage with colleagues throughout the international library sector. While ML expertise with respect to library applications is limited in OCUL, there are staff associated with projects throughout the world that could be helpful. Engaging these people as colleagues and with reciprocity would be beneficial to all.

Training Programs: arrange for in-service training programs. Regardless of the above approaches, a training program in ML for OCUL and Scholars Portal staff should be implemented to build capacity in current staff. This would be a technical approach suitable for existing systems staff or others with an interest in expanding their technical skills. Building capacity in this manner is a long-term objective that recognizes the ongoing need for expertise in this area.

8. Infrastructure

8.1. Role of Scholars Portal

Scholars Portal will provide the key infrastructure necessary to start and support work on the various AI/ML use cases. The details of this infrastructure will emerge as the use cases are further defined. However, it is important to acknowledge that new and expanded infrastructure at Scholars Portal will require new investments.

The Permafrost service provides a model for supporting OCUL AI/ML projects. Scholars Portal would centrally host the tools, services, and data. They would also provide training support. Individual projects would be managed through project teams drawn primarily from OCUL libraries.

The challenge of supporting AI/ML development and services is shared by other key infrastructure providers in Canada such as Érudit, CRKN's Canadiana, the partnership Coalition Publica, and PKP. Discussions among these providers is welcome and will contribute to a more robust environment.

8.2. Pipelines

The development and management of AI/ML pipelines is an important aspect of the strategies proposed in this report. Pipelines can be thought of as technology recipes. Pipelines define standard steps, procedures, tools, and services for the completion of specific types of AI/ML use cases (e.g., audio-to-text, document summarization, image recognition). They provide user-friendly DIY procedures that empower staff with basic training to adopt AI/ML solutions.

While the pipelines will require management by Scholars Portal, and staff in OCUL libraries may require consultations in some cases, the objective is to build staff capacity and independence. While not all use cases can be implemented as pipelines, it is expected that parts of the typical AI/ML development process (e.g., data acquisition and preparation, accessing necessary technology, retrieving and verifying output) can be supported with pipeline components.

8.3. Virtual Machine Learning Lab

The Machine Learning Lab is a proposal to centrally, but virtually, provide access to tools, resources, and expertise necessary for OCUL members to explore and/or implement specific ML solutions. The Lab would license and make available access to approved ML and related software, provide storage via the Ontario Library Research Cloud (OLRC), and offer training and consultation.

Only modest compute options are expected to be available through Scholars Portal infrastructure. It is anticipated that large-scale LLM or model training will require access to external compute facilities such as Digital Research Alliance of Canada (the Alliance), OpenAI, Amazon's AWS, or Microsoft's Azure. The advantages and costs of compute through these services are project dependent and have yet to be estimated (see section below on Compute Requirements).

Several library-based ML labs offer different models. Development and experimentation is focus of many labs: LC Labs (Library of Congress), KBLab (National Library of Sweden), AI-Lab (National Library of Norway), Library Innovation Lab (Harvard) and Library AI Trials (Stanford Libraries). Labs in universities typically emphasize collaboration with faculty or graduate student research: TMU Collaboratory (Toronto Metropolitan University Library) and Library AI Lab (University of Rhode Island).

8.4. Compute Requirements and Sources

The cloud service from the Alliance is an “infrastructure as a service” offering. Virtual machines are created allowing the installation of the necessary software, defining storage options, and accessing required GPU compute. Storage could be provided through the Alliance or potentially using OCUL’s OLRC. Access to GPU compute is available as needed (up to a finite threshold). It is anticipated in the pilot stages that this compute limit will not problematic. This option assumes OCUL projects have the necessary ML technical expertise. The Alliance services and resources are freely available to Canadian researchers.

A second option is to engage with a full-service commercial provider, such as AWS or Azure, and utilize their standard processes (including access to relevant software and compute requirements). This approach would allow OCUL to draw on the expertise available through the support services from these companies. However, the cost models for these services are complex since each component and process is costed separately. AWS and Azure offer some “free tier” services that could be a cost-effective option for small-scale pilots.

9. Training Data and Open Access

Data are at the core of all AI/ML applications. The use of various data sources as training data, most notably for LLM, has raised numerous issues. This is a very fluid area with some arguing that existing intellectual property frameworks (such as copyright, IP, patents) are not adequate to address the concerns and opportunities generated by AI/ML. However, the Canadian Association of Research Libraries and others (e.g., Michael Geist, University of Ottawa and Carys Craig, Osgoode Hall Law School) argue that fair dealing for research purposes applies with respect to text and data mining.

During the early phases of its AI/ML projects, OCUL will focus on using open access resources and public metadata for training data. This would include open access resources in Scholars Portal Books and Journals, Institutional Repositories, government documents, and collections in archives and special collections. A motivating factor is to provide enhanced access to these important resources.

10. Licensing

AI/ML is being incorporated widely into many information resource products, some existing and others new. Examples include scite.ai (smart citations and literature search), Scopus AI (scholarly resource database), Transkribus (manuscript transcription), and systematic reviewing tools such as DistillerSR, which already incorporates AI/ML capability, and Covidence, which is likely do so shortly. Adobe has just added AI features to Adobe Acrobat.

In these and other cases, there is no compelling argument to create a similar tool or resource. Rather, the decision is whether to seek a consortial license. It is not known at this point if there is sufficient interest in any of these by OCUL libraries and if a consortial license is available (both conditions for OCUL licensing or acquisition). It is recommended that such decisions follow the existing process enabled by the OCUL Information Resources Committee (OCUL-IR). This would facilitate the process of collecting interest, establishing evaluation criteria, determining benefits and acquisition details, and moving forward with negotiations should the situation warrant it. Given that these tools differ from resources typically reviewed by OCUL-IR, it is suggested that those seeking consortial licenses work with OCUL-IR to help define the need and criteria. Many of these tools or services may benefit from short-term, pre-negotiation trials combined with basic training to assist OCUL libraries in assessing interest.

An interesting experiment would be to consider licensing a generative AI tool (e.g., ChatGPT or a similar tool) giving “pro” access to all students in the OCUL community (~400K students). A cost-benefit analysis of such an option is not considered as part of this report. However, as a thought experiment it would be instructive to see how transformational (or detrimental) this would be.

Some of the use cases identified below can also be accomplished through purchased or licensed products or services. There are several questions to be considered. Why should OCUL pursue some of the in-house use cases if a commercial product is available? Why build when we could buy or license? While some of the use cases involve commercial components (e.g., access to compute and specialized models for large scale training), many incorporate open-source and involve local development or integration. This approach is consistent with a more hands-on attitude towards AI/ML development that builds capacity among OCUL staff, controls and manages the OCUL technology stack, and exercises a preference for open-source solutions.

11. Accessibility

Furthering accessibility is a significant use case for AI/ML. Responding to AODA standards is an objective, and part of the regulatory requirements, of all OCUL libraries. However, many collections, especially those in archives and special collections areas, are difficult to make accessible. These limitations have precluded some collections being made available online and in formats that lend themselves to broader use.

Various AI/ML techniques can enhance accessibility. For example, changes in format (audio to text; video to text), plain language (summarization), and multilingual capabilities (presentation in language of choice). A particular benefit is applying AI/ML to data remediation so that documents are compatible with assistive technology such as screen readers. OCUL's Accessible Content ePortal (ACE) underscores the consortial commitment to accessibility. The use of AI/ML in specific areas can enhance this service.

12. Agency, Environment, and Explainability

While the Guiding Principles identify foundational values supporting the AI/ML strategy and implementation, there are other concerns that deserve attention and consideration. Three issues are highlighted here: human agency, environmental impact, and explainability.

A "humans-in-the-loop" philosophy ensures that the technology is always guided and informed by human involvement and control. Human agency is important at the design stage and at the point of use. Designers must create AI/ML tools and services that allow for human direction and intervention.

Training LLM, which is a component of some of the use cases discussed below, has a notable environmental impact. While mitigation efforts are ongoing (e.g., smaller models, green data centres, more efficient compute), the environmental impact must be acknowledged. Responding to this challenge will require balancing objectives and supporting efforts to address these concerns.

Accountability in the provision of AI/ML tools and services requires both transparency in how they were created (e.g., training data, model characteristics) and explainability in the results provided (e.g., the ability for users to question and contest results). Transparency contributes to user trust and explainability supports system veracity.

13. Discrimination, Bias, and Veracity

The widely identified limitations and “hallucinations” of generative AI highlight the critical issues of discrimination, bias, and veracity in AI/ML technologies. Working with AI/ML is a crucial opportunity for libraries to recommit to advancing Indigenization, equity, diversity, inclusion, anti-racism, and accessibility. Providing tools and services that undermine trust will have serious implications.

Increasingly there are techniques for data acquisition, model selection, fine-tuning, and guardrails that mitigate these concerns. However, none are perfect. These issues will continue to occupy our concern and attention. The use of warnings and disclaimers are important signals to our users, but they hardly resolve the issues.

What is most important is that research libraries engage with the development and use of AI/ML to advocate for change and to adopt effective strategies. One method to do this is the idea of “critical making.” This approach to technology development recommends not simply building and deploying solutions but doing so with a constant critical lens that identifies the social implications of the solution and seeks to implement mitigations.

14. Use Cases and Projects

The use cases described below are selective and illustrative. Inclusion here does not mean that OCUL will pursue any of, or only, these use cases or projects. Responses from the review of this Interim Report and future consultations could identify additional relevant and compelling use cases and projects that will be incorporated into the full report. Each section defines the value of the use case and outlines projects to implement them. Some of these sections will be more detailed than others as further investigation is still required.

It is reasonable to think of various use cases arising from generative AI tools and services (e.g., ChatGPT, Gemini, Perplexity). As standalone systems these tools have the capability to do many things relevant to library work or user benefits. However, they are an example of a general purpose technology (ironically known as a GPT). They can be used for a wide variety of functions and purposes.

Libraries can use generative AI for format translation, error checking, metadata creation, translation, literature reviews, and so many other tasks. However, while beneficial, these examples use AI/ML as individual productivity tools; very helpful and encouraged but not consortial applications worthy of collaborative solutions. Through AI literacy initiatives, library staff need to understand how using these widely accessible generative AI tools can positively influence their work.

General purpose productivity tools utilizing AI/ML, such as Microsoft 365 CoPilot, are also excluded from consideration even though they have beneficial uses. The focus is on use cases that are specific to the library and utilized for consortial or collective actions.

14.1. AI Literacy and Instruction

The need for AI literacy and instruction for students, faculty, and library staff was widely identified during the early consultation phase. Emphasis was placed on working with library staff (viewed as a prerequisite to engage staff in AI issues and services) but the need for programs to support students and faculty was also required.

The link to institutional teaching and learning strategies is important as is connecting this to the larger issues of information literacy, trust, misinformation, and attribution. The work of the recently created subcommittee on AI in teaching and learning by the CARL Advancing Teaching & Learning Committee will be important to this.

AIMLComm, the newly launched Community of Practice regarding AI/ML in academic libraries, has a channel focused on AI literacy and instruction. It is still early but discussion has begun, and participants have been sharing their local initiatives. It is anticipated that this channel will enable collaborative work on programs and materials allowing libraries to share and work together rather. Engagement in this area has already included people from outside OCUL. This is an excellent indication that working at a national scale in this area is possible. Furthering this collaboration will likely require leadership (and motivation) from committed individuals. Acknowledging this work as a formal project would further this.

While there is discussion related to the Virtual Reference use case focusing on systems with a pedagogical perspective (i.e., not just question answering but responding with learning strategies to equip the inquirer with new or enhanced skills or understanding), there is also interest in creating an AI literacy tool delivered as an LLM. The LLM would act as a teacher or tutor with respect to AI literacy (see AI Tutor Pro and Contact North's Digital Skills Chatbot for an example). Such a tool could respond to students, faculty, and library staff.

14.2. Audio-to-Text Transcription

Many OCUL libraries have spoken audio files (e.g., oral histories, meetings, podcasts, speeches, broadcasts) that have not been transcribed. Using AI/ML transcription models, these audio files can become more accessible. The resulting transcriptions are indexed to the audio for easy referencing and a corpus of related audio files could be trained for enhanced discovery. An important outcome is an alternative format that furthers AODA objectives. While transcribed audio still requires verification, one

promising option is to use corrected transcripts to train (fine-tune) the model to identify and correct errors.

While there are many audio-to-text transcription tools or services available, the opportunity to explore this use case early in the Task Force investigations resulted in a pilot initiative. An open-source tool with widespread use and success (Whisper; created by OpenAI and extended by many others) was the focus of a hackfest discussed earlier.

A key outcome of the hackfest was the viability of Whisper for a broad variety of audio corpora transcriptions and video captioning. In addition, participants developed a user-friendly pipeline for library staff to easily request and receive audio transcriptions through centrally managed resources at Scholars Portal. Extensions to this pipeline are expected to leverage compute resources at the Alliance for projects that require GPU capabilities. It was also clear, for example, that this pipeline could be used by faculty and graduate students to obtain transcriptions of research interviews.

14.3. Metadata Creation

Using generative AI for metadata creation has received considerable attention. The limitations of metadata accompanying the purchase of large e-resource collections suggests an application for AI/ML to generate MARC or JSON records for inclusion in Omni or other integrated library systems.

Glenn Greenly at Capilano University has created *CataloguerGPT* using the OpenAI API for GPT4 to generate catalogue records from a variety of input (eBook contents, images of book pages, or eBooks themselves) and create MARC records as output. Knowledge of cataloging standards, LCSH, LC and Dewey, and the MARC format is inherent in the LLM's training. The program supplements this with information, examples, and prompts to guide and instruct the LLM output. A less advanced version is available publicly. This and other initiatives offer promising, consortial-scale opportunities.

A different metadata application is provided by a service prototyped for the institutional repository (IR) at the Los Alamos Research Library. The service semi-automates the creation of subject metadata for submission ingest that could be used for the emergent National Institutional Repository Service. While the author of an IR submission is typically encouraged to add subject metadata (usually keywords), the LANL Research Library pipeline provides a means to augment that metadata using preferred descriptors and with the participation of the author.

From the text of the submission, GPT is used to generate possible subject terms from one or more thesauri (e.g., LCSH, specialized vocabularies). These terms are emailed to the submitting author. They are asked to identify relevant terms by checking off items

on the list and returning the email. The email is processed resulting in the selected terms being added to the submission record.

14.4. Data Curation

Curation of a research dataset is the process of reviewing an entire dataset (files, code, documentation, and metadata) to assess whether it meets the FAIR Principles (Findable, Accessible, Interoperable, and Reusable) and making recommendations on how the dataset could be altered or updated to achieve FAIRness. In Ontario (and Canada), not all institutions offer a dataset curation service due to lack of resources and expertise.

AI/ML offers substantial opportunities in data curation by automating standardized tasks, ensuring accuracy and completeness in dataset review, handling technical file assessments, and streamlining communication and documentation processes. These advancements can significantly enhance the efficiency and effectiveness of data curation practices. The following provide specific examples:

a) Review datasets for accuracy and completeness: AI/ML can assess the consistency between files in the dataset and their documentation, recommend appropriate metadata schemas and verify the completion of metadata fields, and verify and recommend links to related publications, datasets, and resources.

b) Review technical aspects of dataset files: AI/ML can be instrumental in ensuring all files in the dataset open correctly, converting proprietary file types to non-proprietary formats (if possible), checking for appropriate file structure and naming conventions and running included code to confirm it delivers expected results.

c) Facilitate communication with dataset owners and support personnel: AI/ML can streamline communication by summarizing feedback and recommendations, prioritizing them from most critical to least, sending review results to researchers for action or approval, and documenting the entire curation process within the dataset for transparency and provenance.

14.5. Virtual Reference

The concept of AI-based Virtual Reference (VR) has been around since the 1960s, but recent advancements in generative AI have renewed interest in its practical application. "Ask a Librarian" is a successful VR service, utilized by a majority of OCUL institutions, offering services in English and French for 67 hours a week and handling approximately 27,000 requests annually. The service uses the LibraryH3lp platform.

The service is more than just a question-and-answer system; it involves around 350 operators who incorporate information literacy instruction into their responses. This highlights that any AI/ML implementation should include not only a human-in-the-loop approach but also integrate a pedagogical strategy.

The potential of AI/ML in this setting is significant. While AI/ML can handle straightforward "ready reference" inquiries, there is also the potential to manage more complex dialogues. Inspirations from existing AI educational tools like the Contact North AI Tutor and the Khan Academy suggest the feasibility of a tutor model in AI/ML for VR services.

When considering the implementation of AI in VR, there are three strategic options: building a new service, buying an existing product, or augmenting current services. Currently, there is no ready-made ML VR product available, and building a new service from scratch is a complex and challenging endeavor. The augmentation approach, which involves enhancing existing services like "Ask a Librarian" with AI/ML capabilities, offers several advantages. These include collaboration with experienced partners, phased and modular development, and effective management of financial and human resources. Importantly, this approach views AI/ML as a complement to library staff, not a replacement, recognizing the value of human expertise in the service.

The proposed strategy for implementing AI/ML in VR services is divided into four stages:

- Stage 1: Analyze "Ask a Librarian" logs to understand current requests and expertise.
- Stage 2: Engage development partners, combining LibraryH3lp's library expertise with companies experienced with chat technology.
- Stage 3: Train a local LLM using "Ask a Librarian" materials (e.g., logs, training materials, best practices documentation), initially as a tool for operators or after-hours service.
- Stage 4: If successful, release the VR model publicly with a pedagogical model and human-in-the-loop design, operating 24/7.

This use case and project outlines a structured approach to integrating AI/ML into the "Ask a Librarian" service. It emphasizes the importance of augmentation, partnership, and maintaining a balance between technological innovation and the invaluable role of human expertise in library services.

14.6. ACE Accessible Tagging and Book Summaries

The ACE book collection from Scholars Portal makes alternative formats available for users at universities and colleges in Ontario with print disabilities. Data remediation is an important issue for ACE content. Accessible tagging of ACE books would ensure these resources are compliant with assistive technologies such as screen readers. Tools to remediate existing and future ACE books requires further investigation.

For users of this collection, it is often difficult and time consuming to determine the nature of the content and if it is relevant to their needs. Alternative format book summaries would be helpful; these could be done at the book and/or chapter levels. A variety of summarization models are available using extractive and abstractive summarization techniques. Extractive summaries are created from a selection of sentences in the document while abstractive summaries, typical of generative AI approaches, are interpreted from the text provided. Accessible summaries could be generated from the scanned books, added to the ACE book metadata, and made available as a filtering tool for users.

14.7. Canadian Census Data

Canadian Census Data Discovery Partnership is an inventory of all known census data including census data files (all media and formats) and data tables, maps, and figures, embedded in scanned census publications. The scanned publications contain data tables, maps, and figures. The use of pre-trained and locally trained LLM, could generate usable data files (e.g., CSV, XML) and extract metadata for discovery and reuse.

14.8. Canadian Historical Maps

The GeoPortal at Scholars Portal has a collection of ~15,000 digitized and georeferenced Canadian maps from the 1:50,000 series. AI/ML tools, a trained LLM and related tools for example, could be used to identify similar feature patterns across large maps or image corpora. This can lead to extraction of historical topographic features (forest coverage, cheese factories, road networks, railways, landscape features, mining and industrial development, land use, urban development) as well as for the generation of keywords and other metadata. A model for such a project is Mäyrä, J., et al. (2023) Utilizing historical maps in identification of long-term land use and land cover changes. *Ambio* **52**, 1777–1792. doi.org/10.1007/s13280-023-01838-z.

14.9. Canadian & Ontario Government Documents

The following collections at Scholars Portal and the Internet Archive contain ~325,000 Canadian and Ontario government documents in PDF format:

Canadian Public Documents Collection 60,000 (Federal)

Statistics Canada 110,000 (Federal)

National Round Table on the Environment and the Economy 245 (Federal)

University of Toronto Government Digitization Collection 75,000 (Federal & Ontario)

Canadian Think Tanks Collection 15,000 (National; not necessarily Federal)

Ontario Government Documents Collection 46,000 (Ontario)

Internet Archives Government Documents Collection PDFs (Ontario)

These historic documents, some as early as the 1840s, have relevance to many research fields. Unfortunately, searchable metadata is often either unavailable or uneven in extent and quality. Providing enhanced metadata and new discovery tools would substantially increase accessibility and useability of these resources.

Preliminary experiments with ChatGPT have successfully extracted useable metadata and document summaries from the scanned images. Even complex data presentations (e.g., a table printed sideways on the page) yielded a worthwhile summary. With the GPT API and other tools it will be possible to: 1) create a pipeline to query the corpus sequentially and generate metadata and document summaries, and 2) use the JSON output to upload and modify existing metadata. A locally trained LLM, on either the core metadata data plus summaries or the full text, would enable natural language queries and semantic search capabilities.

14.10. Canadian Scholarly Publications: Books and Journals

A large corpus of Canadian scholarly publications is contained in Scholars Portal Books and Journals. Examples include the All Canadian University Presses (ACUP) Book Collection and the journals associated with the Canadian Association of Learned Journals. Training on these resources as a distinct corpus would enhance discovery through natural language queries, semantic searching, and summarization.

For example, the ACUP collection consists over 13,000 PDFs (some EPUBs) from 16 of the leading university presses from across the country. Both the subject matter and metadata quality vary greatly between publishers. This collection is currently distributed by Degruyter who have made the content fully downloadable since taking it over in 2023. The presses in the collection currently include Alberta, Athabasca, Calgary, ISER, Laurier, Laval, Manitoba, Montreal, MQUP, Ottawa, PIMS, Quebec, Regina, Toronto, and UBC.

This initiative would be undertaken with the participation of Canadian scholarly resource publishers to align the objectives with those of these providers. Discussions with these groups will begin shortly.

14.11. Scholaris: National Institutional Repository Service

Institutional repositories are important scholarly resources that contribute to open scholarship. The recently announced Scholaris, the National Institutional Repository Service, promises to significantly expand access to these resources and to further encourage open access. Applying AI/ML techniques to this corpus would enable metadata creation, natural language processing queries, semantic search, and document summarization. Discussions regarding AI/ML pilot projects with Scholaris are at the very early stages.

14.12. Scholars Portal Books and Journals

Applying AI/ML tools to the large Scholars Portal Books and Journals corpora by training on the full text or on the metadata would provide new ways to explore this important scholarly content. However, in the immediate term, OCUL will focus on open access resources (e.g., government documents, institutional repositories) and Canadian scholarship (e.g., books and journals from Canadian publishers) to enhance access to these collections and to build expertise in using AI/ML techniques on large corpora. With experience, a next step might be to train on the metadata from collections and enable natural language queries and semantic searching.

15. Recommendations: Actions and Timelines

[details to follow in the final report – May 2024]

16. Funding

The OCUL AI/ML strategy's funding is diversified across several channels:

Startup Funding: The OCUL New Initiatives Fund has allocated $440,000 over two years for start-up costs:

- $220,000 in May 2024
- $220,000 in May 2025

These funds are designated for project coordination ($100,000 annually) and analytical/technical support ($120,000 annually).

External Funding Sources: Opportunities exist to obtain funding for specific AI/ML initiatives, particularly through SSHRC's Insight Development Grants and their Partnership Engage Grants. Another possible funding source, for AI literacy and instruction projects, is eCampusOntario. The specifics of these funding opportunities warrant further exploration.

In-Kind Support (OCUL Libraries): In-kind contributions from OCUL libraries form a crucial part of the funding strategy. This includes staff allocation for project roles, providing both technical and domain expertise.

Collaborative Partnerships: Engaging with partners outside OCUL, especially for specific project use cases, is planned. These partnerships might involve shared costs and in-kind contributions, with potential partners including commercial groups as well as research libraries and related organizations. Formal co-development agreements will be essential.

Sustainability Focus: Developing a sustainable funding approach is critical. After the initial phase, the financial responsibility is likely to shift to the libraries. Strategies might include securing new institutional funding and reallocating existing library budgets.

Strategic Communications: To ensure the successful adoption and sustained use of AI/ML in OCUL libraries, it is essential to engage with various stakeholders within OCUL institutions (such as senior administration, faculty, students, and library staff). Making a compelling case for the adoption of these technologies and demonstrating their benefits to the local institution, the Ontario university system, and potentially the global academic community is crucial.

APPENDICES

Appendix 1: OCUL Task Force on Machine Learning/AI Terms of Reference

Background

Machine learning (ML) is a transformative technology that has been widely adopted by various organizations. Applications in research libraries have been identified but adoption has been limited. While time, funding, and access to tools are common barriers to adoption, the lack of expertise or experience is often the first barrier identified. To increase understanding and to help identify specific paths forward, OCUL Directors had several conversations amongst themselves, with library experts and with CARL and CRKN throughout 2023. There is consensus that the time is right to move ahead with a collaborative project for Canadian research libraries, and that OCUL can take a lead role in defining this project. An application to the New Initiatives Fund has been submitted to secure funding but the shape of the project remains undefined due to lack of clarity around specific options, issues, and use cases.

Purpose

This Task Force will take the lead in identifying and organizing immediate next steps to provide an informed perspective on ML to allow OCUL to make strategic choices on utilizing and piloting aspects of this technology. Suggestions for immediate next steps as surfaced in prior conversations include an Interim Report, a summit, and a Final Report and Strategy, with each informing the other.

The objectives of the interim report and summit would include:

- Reviewing the landscape of ML applications in academic libraries
- Identifying relevant legal, ethical, technical, expertise, capacity, infrastructure and investment issues that impact adoption and use of ML
- Describing potential ML use cases for academic libraries that provide services, address current problems, or present new opportunities
- Proposing multiple options for pathways forward

The Task Force could extend its mandate, if charged by OCUL Directors, to include scoping and oversight of the ML project itself (likely in partnership with stakeholders outside of OCUL).

Responsibilities

The Task Force will:

- Continue the work and conversations to date by developing and implementing an interim report, summit, and final strategy including outcomes and timelines (see Deliverables for more details)
- Identify partners and stakeholders within and external to OCUL who should contribute to each of these steps
- Identify a project manager and funding as needed

Membership

Due to the time-sensitive nature of the work, membership should be of a manageable size with no more than 5 Directors plus representation from the OCUL office and Scholars Portal team. At least one Director should be on the Executive. The project manager is also a member.

Appointment Process

Self-nomination

Term

Approximately 6 months (until spring 2024 Directors Meeting)

Chair

The group will appoint a Chair.

Meetings

Remote or hybrid using OCUL Zoom. SPOTDocs and listservs organized by the OCUL office are available, but the group may prefer to explore other options for ease of communication and document sharing.

Reporting

Regular reports to Executive with a final report at the spring Directors Meeting.

Resources and Budget

It is likely that this work will need to be coordinated by a dedicated project manager. This position could be a secondment from a member library or an external hire. For the latter, possibilities within the OCUL budget need to be explored (bridging NIF funds, surplus, etc.)

Deliverables and Timeline

- November 2023 to January 2024: Task Force deliberations
- Summit planning and interim report drafting: Nov-Dec. 2023
- Summit agenda and speakers finalized: January (topics would be informed by interim report)
- February 2024: Virtual Summit
- March 2024: Task Force reflection and deliberations
- April 2024: Final Report and Strategy preparation
- May 2024: Task Force Report and Strategy, including recommendations and ML project direction, distribution to OCUL Directors; presentation at OCUL meeting.

Approved by: OCUL Executive Committee

Approved on: September 18, 2023, Revised by the Task Force Nov. 6, 2023

Appendix 2: People Consulted/Conference and Meeting Discussions

As part of the investigations to date, the following people gave generously of their time and provided important insights that have shaped this report. The Task Force thanks them for their contributions.

Clare Appavoo, Executive Director, CRKN

Jasmine Bouchard, ADM, User Experience and Engagement Sector, LAC

Mish Boutet, Digital Literacy Librarian, University of Ottawa Library

Jacquelyn Burkell, Associate VP Research, Western University

Alicia Cappello, Engineering & Science Librarian, Queen's University Library

Mark Daley, VP Artificial Intelligence, Western University

John Fink, McMaster University Library

Michael Flierl, Student Learning Librarian, Ohio State University Libraries

Christa Foley, Assistant Director, Information Resources and Collections, OCUL

Valerie Gibbons, Chair and member of the VR Coordinators Meeting, Ask a Librarian

Gayleen Gray, CTO, McMaster University

Majela Guzmán, Research Librarian Social Sciences, University of Ottawa Library

Susan Haigh, Executive Director, CARL

Rui Han, OTA Technologies

Jeremy Heil, Digital & Private Records Archivist, Queen's University Library

Elizabeth Kalbfleisch, Project Officer, CARL

Bart Kawula, Web & Discovery Services Librarian, Scholars Portal

David Kemper, McMaster University Library

Amber Leahey, Data & GIS Librarian, Scholars Portal

Karen Linauskas, Director General, Access and Services, LAC

Amber Lannon, University Librarian, Carleton University

Beth LaPensee, Senior Product Manager, ITHAKA

Judith Logan, Head, Research & Education Robarts, University of Toronto Libraries

James MacGregor, Manager, Research Infrastructure and Development, CRKN

Steve Marks, Digital Preservation Librarian, University of Toronto Libraries

Kevin Matsui, Managing Director, CARE-AI, University of Guelph

Giovanna Mingarelli, CEO, MC3

Guinsly Mondésir, Coordinator, Virtual Reference Services, Scholars Portal

Javad Mostafa, Dean, iSchool, University of Toronto

Craig Olsvik, Senior Manager, Licensing & Member Services, CRKN

Sabina Pagotto, Assessment & Member Engagement, Scholars Portal

Danica Pawlick Potts, Indigenous RDM Project, Western University

Jordan Pedersen, Research and Scholarship Librarian, University of Guelph

Jennifer Peters, Manager of Library Literacy, Instruction, and Outreach, Seneca Polytechnic

Brendan Quinn, Senior Developer, Northwestern University Libraries

Art Rhyno, Systems & Liaison Librarian, University of Windsor Library

Tim Ribaric, Brock University Library

Bryan Ryder, Senior Machine Learning Engineer, ITHAKA

Curtis Sassur, Head of Archival and Special Collections, University of Guelph Library

Joel Serino, Head of Technology, MC3

Carol Shepstone, Research and Impact Assessment, Toronto Metropolitan Library

Harpinder Singh, Senior Systems Administrator, Scholars Portal

Adam Smith, Rise Up Strategies, Ottawa

Emily Sommers, Digital Archivist, University of Toronto Libraries

Jean Tak, OTA Technologies

Amaz Taufique, Manager, Enterprise Infrastructure and Staff Technology, University of Toronto Libraries

André Vellino, Director, School of Information Studies, University of Ottawa

Fangmin Wang, Former Collaboratory Director, Toronto Metropolitan University Library

Shellen Wang, OTA Technologies

Jacqueline Whyte Appleby, Interim Associate Director, Scholars Portal

Conference and Meeting Discussions

The following conferences and meetings provided an opportunity to engage participants in discussions relevant to this investigation.

Access (Halifax) October 23-25, 2023.

Fantastic Futures; AI for Libraries, Archives, and Museums (Vancouver) November 15-17, 2023.

OLA Copyright Symposium (online) December 7, 2023.

Ontario Library Association Super Conference (Toronto) January 26, 2024.

Whisper Hackfest (McMaster University) February 24, 2024.