AI & Machine Learning Project FAQ

Project: Government Documents

As part of the Ontario Council of University Libraries (OCUL) AI and Machine Learning Initiative, this project builds on large-scale digitization efforts at the University of Toronto to improve access to a collection of government documents hosted on the Internet Archive.

While the collection is publicly available, metadata is often sparse or inconsistent. AI and machine learning (AIML) technologies are being applied to improve optical character recognition of scanned materials, enrich metadata, and develop new discovery tools, significantly improving the accessibility and usability of these resources for researchers and the public.

December 2025

What are these documents?

All of the documents used in this project come from the same collection: the University of Toronto government publications digitization collection hosted on the Internet Archive. This collection represents just over 50,000 documents originally housed on the fifth floor of Robarts Library. Starting in 2021, these documents were digitized, made freely available online, and then moved to long-term storage.

How are intellectual property, copyright, and privacy being considered and preserved?

The materials we are working with are documents that are either public domain or under crown copyright, and have already been made freely available online by the University of Toronto. As these documents have long been made publicly available by the government, and we do not foresee any privacy concerns. No personal, identifiable information about individuals who may use the collection or access the generated metadata is collected as part of this work.

How is the needed compute capacity for this project being procured?

The compute requirements for this project were initially being catered by OCUL's in-house AI server equipped with two Nvidia H100 GPUs. Initial performance numbers revealed that optical character recognition — or OCR — on the entire collection would take around a year to complete. To address this bottleneck, the team explored multiple strategies including the purchase of additional GPU servers. The breakthrough came through partnering with the Digital Research Alliance of Canada, which provides Canadian researchers access to national GPU supercomputing clusters. By re-engineering our pipeline using Ray, a distributed AI compute framework, we were able to leverage dozens of GPUs simultaneously across the Alliance's infrastructure. This approach reduced our processing timeline from one year to a few days. The project now operates on a hybrid model: OCUL's two H100 and two 5090 servers, each handling development, testing, and hosting the GovDocs-Meta platform large language models (LLMs), while production batch processing of the 50,000-document collection runs on Alliance infrastructure.

What has been learned so far from the project?

OCR technology is developing and improving rapidly thanks to LLM integration, and OCR tools that were best-in-class even a few months ago are being replaced by faster and more accurate tools. This has consequences for development; our workflows must be modular so that we can fairly quickly test and swap out older models for new ones over time. Our Government Documents project demonstrates how libraries can leverage collaborative infrastructure across Canada for practical applications. What traditionally required either years of processing time or prohibitive hardware purchases can now be accomplished in days using shared national resources.

AI implementation requires humans, and humans' specialized knowledge and critical thinking are essential to achieving the results we want.

What aspects of this project might be of use to OCUL member libraries?

Many libraries have digitized collections or special collections that are under- or un-described. Our testing of OCR models revealed many models that could be useful for such collections and adapted across different scenarios. Furthermore, enhancing the OCR process with AI assistance can greatly improve accuracy for digitized text materials.

Have questions or project feedback?

Reach out to Program Manager Kari D. Weaver:
kari.weaver@ocul.on.ca.

OCUL
Ontario Council of University Libraries

ocul.on.ca/aiml-program