Open source as a gateway to artificial intelligence

Edmund Balnaves[1]
ebalnaves@prosentient.com.au ORCID: 0000-0002-0441-8151

Received: 27 April 2024
Accepted: 30 August 2024

Open source has been widely adopted across the entire solution set in the library world, including software for traditional library services, digital libraries, discovery systems, identity management and persistent identifier solutions. Open source is also foundational in the development of artificial intelligence (AI) solutions. Open source AI toolkits have driven the adoption of AI in machine learning, image recognition, video analysis and many other areas relevant to library systems. The communities supporting open source library software are among the largest and most dynamic in the library world. The Koha library management system is supported by a community of developers, librarians, service providers and participating libraries around the world, and is probably the most widely adopted library management system in the world. This article explores the evolution of open source in libraries and the ways in which the open source solutions can facilitate the introduction of AI in libraries in a way that gives libraries agency on the AI model development and implementation. Examples are provided on the use of AI toolkits in the library context.

Keywords: Artificial intelligence, open source, library management systems, digital library systems, discovery, machine learning, large language models

1 Introduction

The concept "open source" does not have a universal definition but is generally shaped around the original GNU General Public License (GPL) released in 1989 by the Free Software Foundation. It is a licensing approach to software distribution that endeavours in one way or another to provide a means of systematically distributing software in full source code in a manner that ensures its ongoing availability in that source to developers and implementers of solutions using that source code. It is no surprise that its success hinges on the ease of collaborative programming in an internet-enabled environment. Its sustainability hinges on its ability to provide service-based and reputation-based business models that give value to ongoing life of the software. Open source has been foundational in the rapid evolution of the systems used by libraries. It has also been foundational in the extraordinary rise of artificial intelligence (AI) systems in the last decade, including the recent emergence of large language models such as ChatGPT. This article explores the synergies between the two open source contexts and the ways in which open source is significant in the evolving synergies between library systems and AI systems.

Library institutions have been active participants in open source development of software and have encouraged the development of standards and the implementation of data interchange systems that support open source. The MARC (Machine-Readable Cataloguing) standard was an important early standard for open interchange of bibliographic data, and the early release of MARC modules written in Perl provided a foundation for subsequent experiments on open source library management systems. Similarly, the Z39.50 networked search standard enabled open inter-networking of library catalogues and was accompanied by open source modules supporting the standard. These came together in early 2000 with the release of the Koha Library Management system (Balnaves 2008; Balnaves 2011).

The release by the Library of Congress of utilities for conversion, validation and interchange of MARC data had a profound influence on open source adoption. This has blossomed into a rich set of tools supporting this primal data interchange standard. Since then, many other projects in library open source have emerged in the library management space, including Greenstone and Evergreen (Breeding 2008).

The open source philosophy is different from “source code provided”. For instance, IBM was obliged to release its operating system source code in the 1960s as a result of the US anti-trust actions of the time – an event unlikely to affect the current social media titans. The collegial environment of software innovation that permeated the universities of the 1960s saw the emergence of the components that emerged as the Unix operating system, an environment that has largely disappeared with the drive for outsourcing in university institutions. Powerful and cheap personal computers (PCs) sustained a new generation of computer users engaged in collaborative “social networking” and software development. The

[1] Edmund Balnaves is CEO and Founder of Prosentient Systems at Sydney Australia

forking and branching that have emerged as the evolutionary model of open source have encouraged ongoing innovation rather than speciation of open source solutions.

The larger library institutions have been active in the open source space for two decades. Acceptance has progressively moved from cautious (Dorman 2004) to enthusiastic (Chalon, Alexandre-Joaquim, Naget & Becquart 2005). The systems have gained in both functionality and reach across different library types. Library 2.0 energised the library interest in new approaches to a functionally weary library catalogue, and propelled innovation and adoption of open source to capitalise on the interest in this area.

One of the enduring synergies for libraries is the relationship between open source systems and open access (OA). Adoption of open source systems and OA in libraries has been incremental but profound over the last 20 years, with similar timeframes of uptake and community acceptance. The availability of open source solutions has enabled the adoption of OA digital libraries (Balnaves 2012). Similarly, a successful OA undertaking can be the justification for contribution to, and enhancement of, open source projects. OA and open science are seen as crucial to sustainability goals (IFLA 2023). Open source software (OSS) can provide an institution with a level of certainty in their operational costs, which is important in the context of sustainable goals. OSS in libraries has become more robust as more users adopt it and contribute to it. It offers a degree of security because there is no dependency on a single vendor, and the code is transparent (and therefore can be fixed). This security feature is enhanced by the efforts of those who use the open source model.

While it may be lower in cost, no information technology system operation is free. The ongoing nurturing of a system, software upgrades over time, support for customisations and enhancements, server administration and network costs are just a few of the baseline elements of managing an information system. Nevertheless, the amortisation of the software support across a wide installed base makes for an effective cost model for smaller institutions. Open source systems can provide a cost-viable model for the implementation of OA in smaller institutions. A common confusion is that open source means “free”; however, systems thinking provides a framework for understanding its context as service in the wider organisational and societal context (Balnaves 2005), and in this context, open source must be understood as one in a variety of the ecosystem of software available to libraries whose operational considerations overlap with commercial systems.

The library community has been effective in defining and sponsoring strong standards for data interchange and metadata sharing. In 1999, the Virginia Library Association held a workshop called “A Perl toolkit for libraries” that became the cornerstone of perl4lib (perl4lib, perl4lib 1999). This workshop formed the foundation for a set of open source tools drawing on strong library standards that formed a starting point for subsequent open source solutions for libraries.

AI is the use of computer systems to achieve tasks that would normally have required human interpretive intervention. This includes:

- interpreting images for place, facial or object recognition
- interpreting audio for language recognition and translation
- analysing data in depth and breadth to discern patterns
- controlling and managing movement of robotics in a real (physical) environment
- conversational dialogue management that recognises and interprets and response appropriately

The concept of AI dates back to the early stages of computers. The philosophical examination of the implications of emerging computer technology began in particular with a workshop held in 1956 at Dartmouth College that included early greats John McCarthy and Marvin Minsky. Expert systems gave way to deep learning models such as the ImageNet Large Scale Visual Recognition entry called AlexNet (Krizhevsky, Sutskever & Hinton 2012). Leveraging on very large collections of data to build models that can give credible predictive insight into the development of large-scale multi-processor architectures has been central to advances in AI over the last decade. As with library systems, open source has been a key element in the rapid evolution of AI.

The following sections explore the emergence of full open source applications in the context of key library software categories (the library management system and the digital library system) The wide adoption of open source among libraries presents an opportunity for the convergence of open source AI systems and open source library systems.

2 Library management systems and Koha

There are many open source library management systems (LMS) that are available adopted now; however, the Koha LMS is unique, at least in the library field, due to the diversity of its community and the functional breadth of the application. Koha represents a preeminent example of a broadly based open source community. The peripatetic nature of Koha evolution in an open source context is demonstrated by the absence of a roadmap. Rather than a long-term plan for architectural development, Koha is released on a bi-annual basis and incorporates the changes that are funded by different contributors around the world. The code base has 244 registered contributors and more than 70 have more than 500 commits during the lifetime of the git repository (16 years).

Koha was created in 1999 by Katipo Communications for the Horowhenua Library Trust in New Zealand and focused on providing a public library LMS solution (Breeding 2014). Its first live operation was in January 2000. In this way, it is the first and still the most widely adopted library systems millennial child. There are now more than 50 support vendors around the world, and Koha is implement both as a self-managed and a vendor-managed solution (Koha Community 2023). The software base remains actively developed, with agile, if somewhat random, adoption of new functions and design. The system has been moving slowly towards an application programming interface (API) platform and a more service-oriented architecture. Its base, written as a traditional Perl-based LAMP solution, is no longer “mainstream”, but its functional breadth sustains its popularity among libraries.

The widespread adoption of Koha among mid-sized libraries has seen the emergence of at-scale solutions at enterprise level, including the Evergreen system (The Evergreen Project 2023) and more recently the EBSCO-sponsored FOLIO system (Open Library Foundation 2023).

3 Digital library systems – Dspace and EPrints

The transition from physical to digital electronic resources in libraries has been in progress for the last 30 years (Butler 1999). The first digital library systems emerged about the same time as Koha, at the turn of the millennium. They were inspired by the early evolution metadata sharing standards and tools, in particular, the Open Archives Initiative Protocol for Metadata Harvesting (OAI/PMH).

The EPrints system started development in 1999 and had its first (v.1) release in January 2000. It is profoundly influenced by the principle of data sharing (EPrints Services 2023) and its ongoing development is supported by *EPrints Services,* a not-for-profit foundation based in the University of Southampton.

The DSpace digital library system emerged in 2002 from a university base (Massachusetts Institute of Technology (MIT)) as a collaboration between MIT and HP Labs and centred on resource delivery in both an OA context and as a tool for educational resource delivery. As with Koha, it proved to be popular, mainly because of its “out of the box” model. Unlike Koha, the DSpace development community was underpinned by a formal foundation with a governing board – the DSpace Foundation. This community merged with the Fedora Commons community to form DuraSpace in 2009. While the intent was to create a common model between the two systems, this merger left DSpace in the wilderness for several years between original DSpace version 1 and a regained momentum with DSpace version 3 after abandoning the efforts to merge the systems. The strategic plan of 2015 and the profound contribution of 4Science (the re-written API) and Atmire (the new Angular user interface) lead to a significant reworking of the system.

In contrast to Koha, DSpace and EPrints represent a more focused developer community with a not-for-profit governing agency. The feature depth of these systems is more narrowly cast than Koha, servicing a specific role in the digital library space. EPrints is represented by a small but diverse and dedicated team – with eight active contributors over the last five years.

The digital libraries open source scene has flourished, and there are now many alternatives to choose from. There is also a growing overlap between the digital functionality of the LMS and the e-resource management of the digital library system. Open source projects encourage the confluence of functionality across these systems.

4 Other open source in the library ecosystem

The emergence of the online realm has seen the decline of the LMS as the centrepiece of the library operation. Instead, it now represents is a complex ecosystem of digital resource management, physical asset management and learning systems. Overarching these is the discovery system as a central search resource. The glue that unifies these systems is identity management to bind the complex rules of digital rights access rules, e-resource access and physical access to resources curated by the library. Each of these areas has witnessed the emergence of strong open source solution. Keycloak, for instance, has established itself as an effective solution for identity management. Keycloak is an open source project sponsored by Red Hat and serves as their own platform for identity management (Had 2023).

5 The open source benefit

Balnaves (2008) looks at the multi-dimensional opportunities for scrutiny when adopting open source, contrasted with proprietary systems that generally give minimal visibility to:

“a) their internal development capacity;
b) their detailed roadmap for development and strategic intentions;

c) the source code for their systems;
d) the internal development dialogue; and,
e) the detailed database schema design.” (Balnaves 2008)

These dimensions include the functional design of the system, the architectural characteristics of the software, the community dimension supporting the open source system, the code design characteristics, the database design and schema elements and the security dimension of the code. It is a false argument that the obscurity of commercial software gives a higher level of safety. Code hackers have been creative in building intelligent penetration of commercial systems through software engineering and other techniques. Open source can be quickly remediated outside the long release cycle of closed source systems. Use of outdated and vulnerable commercial and opensource systems are also co-equal phenomena.

The interaction between strong library standards and the strength of evolution of open source in libraries is evident in the data interchange standards supporting traditional and digital libraries. The library has to deal with many challenges in managing digital resources, such as online collections, news feeds and digitised materials. To make it easier for the library clients to access and use these resources, the library may need to have a single portal that can search across different sources, a system that can handle the workflows of creating and maintaining electronic collections, and a way to simplify the login process for various databases. The fast pace of technology also poses problems for keeping the digital records and creations safe and usable over time.

Metadata is the information that describes the objects in the digital library, such as their title, author, size, and format. Metadata has three functions in the digital library. Firstly, descriptive metadata standards give a common platform for describing physical and digital objects, including standards for identifiers such as the International Standard Book Number (ISBN) and International Standard Serial Number (ISSN). There are standards for describing digital objects that are equivalent to the standards for traditional catalogues, such as AACR2 and MARC. Some examples of descriptive metadata standards for digital libraries are Dublin Core Metadata Initiative (DCMI), Metadata Object Description Schema (MODS) and Metadata Encoding & Transmission Standard (METS). DCMI is more common among digital libraries, but MODS and METS offer more comprehensive descriptions as successors to MARC. DSpace and Greenstone use DCMI as their descriptive metadata framework. Semantic metadata is the information that gives the subject and relationship of the objects in the library. This can be based on simple name/value pairs or sentence, like semantic statements (such as the Resource Description Framework (RDF)). Open source solutions are the glue that encourages data sharing between systems, leveraging on strong library standards. The most widely used harvesting system is Open Archives Initiative Protocol for Metadata Harvesting (OAI/PMH), which has many open source solutions.

6 The open source risks in libraries and AI

Open source, in its nature, depends on a community of volunteers. The commitment of this community may derive from the support of a major vendor, from the diversity and size of the community or from the reputation benefit of membership in the community. Rhandhawa (2013) discusses six open source full function LMSs, two of which (ABCD and NewGenLib) have been inactive for some years. Similarly, PHPMyLibrary is no longer active. Organisations supporting open source are vulnerable to collapse. FossHost served as a not-for-profit hosting agency for many library services and notable projects such as GNOME, Armbian, Debian and Free Software Foundation Europe (FSFE). The sudden demise of FossHost in December 2022 (Sharma 2022) illustrates some of the risks around open source and commercial service providers are not immune to the depredations of time.

The rich base of open source in library applications is paralleled by a strong base of applications in open source for AI development. There is clear synergy among these systems and their use of a large corpus of text, image, and video to provide knowledge discovery and understanding. The integration of AI in library systems will extend the capability of library systems to foster new knowledge creation. The open source platforms in both AI and library systems will accelerate this integration.

7 Prospects for AI in libraries leveraging on open source

The open source movement has been a pivotal force in the evolution of AI systems, fostering an environment of collaboration, transparency and innovation. Open source AI models and frameworks, such as TensorFlow, PyTorch, Keras and scikit learn, have democratised access to cutting-edge technology, enabling researchers, developers and organisations of all sizes to participate in the development and application of AI (Vidal 2023). Open source AI tools have accelerated the pace of research and discovery by allowing the community to build on each other's work. This collaborative approach has led to rapid advancements in fields like natural language processing, computer vision and machine learning. By sharing knowledge and code, the AI community has been able to tackle complex problems more efficiently and push the boundaries of what is possible (Vidal 2023).

Moreover, open source AI contributes to the ethical and responsible development of technology. With access to source code, researchers and practitioners can scrutinise AI models for biases, errors and ethical concerns, ensuring that the systems we rely on are fair, accountable and transparent. This level of scrutiny is essential as AI becomes increasingly integrated into critical aspects of society (Gibney 2024).

In recent years, the emergence of large language models (LLMs) has highlighted the importance of open source principles. While some early models like BERT and GPT-2 were commercial friendly and proprietary, the research community has built several important open source models, such as EleutherAI's GPT-Neo and BigScience's BLOOM. These models have not only advanced the state of AI, but also ensured that the benefits of these technologies are accessible to a wider audience (Vidal 2023).

Many toolkits are in Python programming language or C#. They distil a range of complex pattern recognition algorithm and data management tools into simple and elegant processing pipelines to "train" patterns in a large body of data to create a "model" and apply real-life data against the model to discern matches.

The term "algorithm" is used to encompass the set of machine learning techniques that ca be deployed to glean "intelligence" from "data". The term probably more accurately refers to the process of model creation. The difficulty with the term "algorithm" is that it lends itself to an anthropomorphic interpretation of AI that is not quite accurate. Machine learning uses a range of techniques to harness the speed of computers in image, textual, location and other data points to derive an interpretive model out of the data. This model can then be used to suggest a range of things: a face matched to a name, an emotional state or reaction, a set of structured terms to describe a document or image, a location based on an image and many other similar activities.

Open source library systems are well positioned to adopt the emerging platforms for AI. Many AI toolkits have their origins in open source, either in Python or C++. The use of open source allows inspection of the techniques used and the nature of the community of developers and allows an agile approach to the development and testing of their "fit" to the tasks for which they are being employed. Most toolkits for AI have "wrappers" that allow them to be embedded using popular programming languages, including Python, Java and C++ Examples and the following:

- OpenCV (https://opencv.org) – A popular computer vision library that includes tools for face detection and recognition.
- dlib (http://dlib.net/) – A C++ library that includes Python bindings for facial landmark detection, face detection and face recognition.
- Tensorflow (https://www.tensorflow.org/) – A machine learning framework that includes tools for face recognition.
- Face recognition (https://github.com/ageitgey/face_recognition) – A simple Python face recognition library that wraps around dlib.
- PyTorch (https://pytorch.org/) – Another Python machine learning framework that includes tools for face recognition.
- Keras (https://keras.io/) – A high-level deep learning API that includes tools for face recognition.
- MXNet (https://mxnet.apache.org/) – A deep learning framework that includes tools for face recognition.
- Scikit-learn (https://scikit-learn.org/stable/index.html) – A machine learning library that includes tools for facial recognition using SVM and other algorithms.
- Caffe (https://caffe.berkeleyvision.org) – A deep learning framework that includes tools for face recognition.
- CodeProjectAI (https://www.codeproject.com/Articles/5322557/CodeProject-AI-Server-AI-the-easy-way).

Some of these are projects that make use of other projects, a good example of the evolutionary strength of open source. An open source implementation of AI tools has a range of privacy benefits to the library. Image and digital content do not need to be sent to external services where the national hosting and privacy of this content may be uncertain.

Scikit-learn is a highly regarded open source Python library that provides simple and efficient tools for data mining and data analysis, particularly in the realm of machine learning. It is built upon the foundations of NumPy, SciPy and matplotlib, and it offers a consistent interface for fitting, predicting, and evaluating various types of machine learning models. Scikit-learn is widely used for a range of applications including classification, regression, clustering, and dimensionality reduction. Its versatility allows it to be applied in numerous contexts, from spam detection and image recognition to predicting stock prices and customer segmentation.

Scikit learn's capabilities extend to the classification of library documents, a task that involves organising and categorising text-based content for easier retrieval and analysis. In the context of library applications, scikit-learn can be employed to classify documents into predefined categories such as genres, topics, or authorship. This process typically begins with the conversion of text documents into a numerical format using techniques like Term Frequency-Inverse Document Frequency (TF-IDF), which reflects the importance of words within the documents relative to a collection. Examples in open source are readily available – another benefit of open source in this context (scikit learn).

Once the text is vectorised, scikit learn's various classification algorithms, such as Naive Bayes, Support Vector Machines (SVM) or Decision Trees, can be trained on a labelled dataset comprising documents whose categories are already known. The trained model can then predict the categories of new, unseen documents with a high degree of accuracy. This automated classification is particularly useful in large libraries where manual categorisation would be time-consuming and prone to inconsistency. Once again, examples are available in open source, illustrating the accessibility of AI development for libraries in the context of open source (QuantStart 2019).

PyTorch, an open source machine learning library, is used in book libraries for tasks such as building recommendation systems. For example, a project detailed on Medium describes the creation of a book recommendation neural network using PyTorch. The system employs collaborative filtering, a method that makes automatic predictions about the interests of a user by collecting preferences from many users.

In this project, the developer used the Amazon Book Reviews dataset from Kaggle, which contains millions of reviews, to train their model. They used PyTorch's neural network library to create a collaborative filter model class to generate user and item vectors. The model takes user IDs and book titles, encoded into numerical format, and predicts the likelihood of a user liking a particular book helping to personalise recommendations in a library setting (Dickens 2024).

More recently, LLMs have dramatically changed the scene in all aspects of text, image, and video generation. Models such as ChatGPT have increased greatly in capability. Library developers can now use local language models through the use of locally installable pre-built models such as Ollama (Mahapatra 2024). The recently released Ollama 3 is a substantial model. A locally run large language model can be enhanced with local library resources using open source toolkits and LLMs such as Ollama using retrieval augmented generation (RAG) (Born Digital 2024). This gives the library user greater control over the use and quality of content generated by AI.

One of the ethical considerations in the adoption of AI is the degree to which the algorithmic elements of the solution used can be scrutinised, tested, and understood. In the case of open source, the library has agency over the algorithmic design of the AI implementation. The example of RAG used with Ollama illustrates the rich interplay of AI toolsets for their use in library applications.

An open source implementation of AI tools has a range of privacy benefits to the library. Images and digital content are not sent to external services where the national hosting and privacy of content may be uncertain. One of the ethical considerations in the adoption of AI is the degree to which the algorithmic elements of the solution used can be scrutinised, tested, and understood.

With open source, the library has full agency over the algorithmic design of the AI implementation. The very nature of the “open” in open source addresses one of the key ethical dilemmas in AI integration with systems. The library has the opportunity to build its own large-language model that is culturally and linguistically sensitive to the context of the library. The methodology for AI implementation is open to scrutiny and criticism.

8 Conclusion

It is clear that open source is healthy and alive in the library field. Open source solutions in libraries developed from individual supporting components for open standards such as MARC into full-fledged traditional and digital library systems. From early collaborations in open source, there has been rapid evolution of stable and increasingly well-accepted open source solutions in libraries. AI also has its root in open source. This article explores the synergies between these two platforms as a vehicle for effective introduction of AI into library systems and highlights the opportunity open source presents for the rapid adoption of AI in an effective and ethical way by giving the library agency of the use and application of AI in library systems.

References

Balnaves, E. 2005. Content model for reuse: systems for enterprise content reuse. Unpublished.

Balnaves, E. 2008. Open source library management systems: a multidimensional evaluation. *Australian Academic and Research Libraries,* 39(1): 1-13.

Balnaves, E. 2011. *The integrated library vision – the open source/open access model*. ANE Books.

Balnaves, E. 2012. *Ubiquitous and open access: the NextGen library.*

Born Digital. 2024. *Retrieval augmented generation: what you need to know.* https://borndigital.ai/retrieval-augmented-generation-what-you-need-to-know/.

Breeding, M. 2008. Open source library automation: overview and perspective. *Library Technology Reports,* 44(8): 5-10.

Breeding, M. 2014. *The history and background of Koha*. https://librarytechnology.org/document/19403 April 2014

Butler, D. 1999. *The writing is on the web for science journals in print*. *Nature*.

Chalon, P. X., Alexandre-Joaquim, L., Naget, C. and Becquart, C. 2005. Open your mind! Selecting and implementing an integrated library system: the open-source opportunity. *10th European conference of medical and health libraries, Cluj-Napoca*, Romania, 11 – 16 September 2005.

Dickens, W. 2024. *Building a book recommendation Neural Network in PyTorch.* medium.com.

Dorman, D. 2004. The case for open source software in the library market. *Ubiquity,* 4(47).

EPrints Services. 2023. *A brief history of eprints*. [Online] https://wiki.eprints.org/w/History (3 May 2023).

Gibney, E. 2024. Not all 'open source' AI models are actually open: here's a ranking. *Nature*.

Had, R. 2023. *Keycloak*.

IFLA. 2023. *From open science to sustainable development: libraries as essential infrastructures*. [Online] https://www.ifla.org/news/from-open-science-to-sustainable-development-libraries-as-essential-infrastructures/.

Koha Community. 2023. *Support companies by country*. [Online] https://koha-community.org/support/paid-support/country/ (1 May 2023).

Mahapatra, S. 2024. *How to run open source LLMs locally using Ollama.* freeCodeCamp.org.

Open Library Foundation. 2023. *Folio: the future of libraries is open*. [Online] https://www.folio.org/ (4 May 2023).

perl4lib. i. https://perl4lib.perl.org/.

perl4lib. 1999. *MARC.pm: machine readable cataloging Perl module*. http://marcpm.sourceforge.net/writings/marc_and_perl.html.

QuantStart. 2019. *Supervised learning for document classification with Scikit-Learn*.

Randhawa, S. 2013. *Open source library management softwares*.

SciKit Learn. *Classification of text documents using sparse features*. scikit-learn.

Sharma, A. 2022. *Open source software host Fosshost shutting down as CEO unreachable*. [Online] https://www.bleepingcomputer.com/news/technology/open-source-software-host-fosshost-shutting-down-as-ceo-unreachable/ (3 June 2023).

The Evergreen Project. 2023. *Evergreen – open source library software*. [Online] https://evergreen-ils.org/ (3 May 2023).

Vidal, N. 2023. *The AI renaissance and why open source matters*. Open Source Initiative.