THE BUILDING BLOCKS FOR OPEN ECOSYSTEMS OF ONLINE RESOURCES SERVING BUDDHIST COMMUNITIES

by Alex Amies*

ABSTRACT

The paper gives an overview of the state of the art of the software building blocks for development of online resources serving Buddhist communities and how those are driving new capabilities and broadening access. Possible choices of technologies that take advantage of the efficiencies denoted by economists as part of the Fourth Industrial Revolution are explained. The central theme described is the huge scale and rapid evolution of the open source movement and modular package management systems that are built on open source. Illustrative examples are given from the author's experience developing web applications for the study of Buddhist texts, including translation projects for Fo Guang Shan. The changes brought by these technologies in the last five years are great and further impact is still to come. The author hopes that the evolving technologies can bring more improvements to Buddhist resources, including large scale translation of the Chinese Buddhist canon and the collected works of Venerable Master Hsing Yun to English. Large scale translation of historic texts will not necessarily be based on machine translation but machine translation will be important nevertheless. An additional impact is the broadening of access to high quality scholarly resources beyond the academic community to the monastic and lay Buddhist communities.

* Google Inc., California, USA

1. INTRODUCTION

This paper describes building blocks of open systems at two levels: the level of user experience for people accessing online resources and the development of those resources. The software systems described include websites, dictionaries, and online text collections. The theme of this track of the conference[1] is the Fourth Industrial Revolution. In his book *The Fourth Industrial Revolution*, Schwab explains that the most revolutionary impact is not just the new technologies themselves but it is the amplification of the interconnectivity between these technologies (Schwab 2017, pp. 1-3). This amplification is most important at the level of development of software systems due to interconnectivity between both the developers of the systems and of the software components making up the ecosystems. In the now classic book on open source *The Cathedral & the Bazaar*, Raymond describes the how some of the best and brightest minds in the world were attracted to the computing industry in the 1960s-1990s and evolved an early "hacker culture" (Raymond 2001, p. 9-17). The open source community developed from this early computer science culture. However, it was not until the rise in popularity of cloud computing over the last ten years that open source became widely used by businesses. There are two ways that the amplification described by Schwab is seen in the open source community: 1. the opportunity to contribute visible artifacts in large communities and free communication drives a high level of engagement among skilled software developers; 2. the software modules themselves via the package management systems, which are continuously automatically downloaded, compiled, tested, repackaged, and deployed to create new software.

What is meant by an open ecosystem and why should project sponsors care? Software projects serving the Buddhist community are sponsored by Buddhist temples, universities, and lay people. At the level of end user experience a linked

[1] http://www.undv2019vietnam.com/

collection of simple websites is a great example of an open ecosystem. An important aspect of this open ecosystem is the links from external locations to specific content within a website. A project sponsor should care about enabling inbound links because it will enable more users to discover content and navigate directly to the most relevant location. Closed systems, in contrast, present user with a link to install new software or no link at all. Such closed systems lock up resources, inhibiting communication and innovation.

Buddhist teachings have had a large impact on many people's lives, from being able to think beyond ideas of self to dealing with impermanence of the loss of one's loved ones. However, most people in the 21st century have access to tertiary education and multiple religions and do not necessarily take Buddhist teachings at face value. Lay Buddhists and interested people from other belief systems often want access to authentic canonical texts and scholarly analysis of those. The use of open ecosystems is critical in extending open access to high quality, scholarly content beyond the academic community to communities of monastic, lay people, and the public in general. Barriers locking of academic publications in closed digital libraries and open alternatives will be discussed below.

Open discussion amongst end users is not the intended sense of 'open.' It is a possible interpretation but is outside the scope of this paper.

The Diamond Sutra states, "If a bodhisattva gives without abiding in any notion whatsoever, then his merit will be immeasurable" (Fo Guang Shan International Translation Center 2016, p. 17). The goals of open source movement in freely publishing software to the global community are mostly compatible with Buddhist ethics. While the individuals and business publishing the software have not necessarily been free of expectations of gaining something in return, the impact in many cases has been extremely large.

At the level of software development an open ecosystem is one that is based on publicly available software, such as open source, and also publishing the software developed under an open source license. Schwab writes, “when firms share resources through collaborative innovation, significant value can be created for both parties as well as for the economies in which such collaborations take place” (Schwab 2017, p. 60). The open source community is the largest such collaborative community today. As an illustration of the unprecedented size that successful open source ecosystems have become, the open source repository management platform GitHub currently reports 31 million developers with 96 million repositories.[2] Even the resources of the largest corporations and governments cannot match this. In contrast to the opaque silos created by closed ecosystems the transparency of open source drives massive, unplanned innovation, and often disruption. The disruptive aspect should also be considered. The disruption has mainly been to businesses that depend on sales of packaged software for revenue since their products are now in competition with freely available software.

The transformation that is occurring today with open source ecosystems is a combination of three things: 1. the transparency and free access of the open source model; 2. the availability of free services to distribute digital assets via open source repositories; and, 3. package management systems for the automated download, deployment, and use of software modules. It is not just the existence of open source software, it is the convenience and version control structure of free hosted services for open source repositories, such as GitHub and GitLab, that makes the difference. The version control structures imposed are typically rigorous. These consist of an ownership model enabling participation from the public in general, a review and approval process, revision history and rollback, and optionally format checking and testing. There has also been an explosion of free package management systems including the Debian package management system for Debian Linux, npmjs for JavaScript modules, Python Package Index, iPython (interactive Python) scripts, and Docker containers.

[2] https://blog.github.com/2018-10-16-state-of-the-octoverse/

Examples of modular package management systems that leverage these hosted services are the Debian apt-get install command, pip command for Python packages, and Colab for iPython. A recent change is building access to open source repositories directly into tools provided with programming languages, such as the 'go get' command for the Go language, which automatically downloads Go packages and their dependencies from GitHub[3].

The NTI Reader[4] website developed hosts the text of the Taishō canon and the Humanistic Buddhism Reader (HB Reader)[5] web site hosts the text of Venerable Master Hsing Yun's collected works. Both websites include an integrated Chinese-English Buddhist dictionary. These websites, developed by the author, use many of the open source and modular systems listed above. The code for the projects are hosted in GitHub published with an Apache 2 open software license and and text assets with a Creative Commons license. The NTI Reader and HB Reader depend on open source at every level. This includes software that is written in the open source programming language Go that runs in Docker containers. Container images for the web application are stored in a container repository that can be downloaded and deployed on demand. The text search application is based on a machine learning application is trained for document relevance using Colab. The text assets would not have been possible to develop without the use of text assets from other institutions published with a Creative Commons license.

These websites are among the resources used by teams of translators at Fo Guang Shan for translation of Chinese Buddhist texts to English. These teams of translators use other collaborative tools, such Google Docs, that allow multiple authors to collaborate and write publications in real time. Besides the NTI Reader and HB Reader, the group makes use of new resources developed by other Buddhist groups, from

[3] https://golang.org/cmd/go/

[4] http://ntireader.org

[5] http://hbreader.org

both temples and academic institutions. Some of the popular tools include the SAT Taishō Shinshū Daizōkyō Text Database 大正新脩大蔵経 from the University of Tokyo,[6] the Chinese Buddhist Electronic Text Association CBETA 中華電子佛典協會,[7] the Digital Dictionary of Buddhism,[8] and the Fo Guang Dictionary of Buddhism.[9] These new tools and online resources are enabling faster translation of a Chinese texts, which is very important considering the enormous volume to be translated. It now finally seems like there is hope for a complete translation to English of the historic body of Chinese Buddhist literature with or without the aid of machine translation.

There are several points that make Chinese Buddhist literature unique.

The long literary history of Chinese Buddhism provides opportunities and challenges. One opportunity is that the expired copyright enables free reuse of the text. A challenge is in processing and understanding the archaic language, some of which dates back two thousand years. We have lost much of the historic context and understanding of the languages, which needs to be reconstructed by a limited number of expert scholars.

Chinese has no spaces for delimiting words. This leads to more challenges in NLTP for text segmentation. Historic Chinese texts have no punctuation, which is even more difficult.

The use of traditional versus simplified Chinese characters and historic character variants provides even more challenges in text processing.

The religious nature of Chinese Buddhist texts fosters a large lay community

The implication of these unique points is that Buddhist studies has its own unique challenges to solve and cannot completely rely on the tools and resources of other communities.

2. ECOSYSTEMS AND COMMUNITIES

[6] http://21dzk.l.u-tokyo.ac.jp/SAT/ddb-bdk-sat2.php

[7] http://tripitaka.cbeta.org/

[8] http://www.buddhism-dict.net/ddb/

[9] http://etext.fgs.org.tw/search02.aspx

Buddhist and humanities academic communities have benefited from both open and closed ecosystems of software and online resources. Some of the building blocks of open ecosystems that are central to the online resources used by the Buddhist are described.

3. OPEN ECOSYSTEMS

There are a number of key elements that enable the openness of the World Wide Web but are easy to take for granted. Open access on the world wide web depends on the use of hyperlinks as defined in the Hypertext Markup Language (HTML) standard from the World Wide Web Consortium (W3C).[10] Users can navigate from one site to another by following these hyperlinks, crossing site ownership boundaries in the process. A development that may surprise users is that many large applications are now 'single page applications' (Mansilla 2018, pp. 162-163) written entirely in JavaScript. In these applications there is only one HTML page with JavaScript that dynamically manipulates the Document Object Model (DOM) to display new content to users and to handle incoming requests to different links. These applications are most frequently powered by web application frameworks like Angular, React, and Vuejs. These web application frameworks can help provide better user experience by minimizing the time spent going back and forth to the remote servers and still enable hyperlinks in the traditional way, if considered in the design of the application. With HTML5 local storage they can act like installed native applications. Careful thought is required in the design to retain the original flavor of the Web as a network of interlinked web pages.

At the other end of the spectrum are closed monolithic systems, including many digital libraries where access to the digital library requires login via a user account that is granted via membership in an organization, such as a university, or via a credit card. In many of these closed systems the links terminate at site ownership boundaries, such as a login screen. Some of this variety of ecosystems are not fully closed, for example incoming links to books in online bookstores to a specific book may be supported. However, the ebooks do not contain any outbound links or prevent export of

[10] https://www.w3.org/html/

data, say via disabling cut-and-paste. Other examples of closed ecosystems are mobile applications or single page web applications that do not handle inbound requests linking to specific resources. The closed nature of these systems can be avoided if the project owners take appropriate design and care in development. Some digital libraries are exemptions. For example the arXiv digital library developed by Cornell University and the Internet Archive digital libraries, allow inbound links to ebooks without login required and allow exporting of data out, via downloading in unlocks formats.

The academic community is now encouraged to move to a more open model for publications by the open access movement. The concentration of publication of academic journals in control of small number of publishers has created an economic barrier for readers of academic literature outside of large universities (Eger and Scheufen 2018, ch. 1.1). This has especially been a problem for Buddhist monastics and lay people seeking to access academic literature on Buddhist subjects. Open access refers to making academic publications available free of cost online. An example of an open access publisher is Frontiers Media, which is an Open Access publisher of peer reviewed academic journals.[11] Frontiers Media also enforces a Creative Commons license to allow free copying of publications. The free availability of open source software is one of the enablers of this. A compromise that allows authors to choose to self-archive their own articles with free public access is called 'green open access' (Eger and Scheufen 2018, ch. 1.3). In summary, open access is one of the building blocks of open ecosystems.

In the past and still continuing into the present, support of software by vendors was considered a major cost. Today the cost of that support is reduced by many vendors via online forums, such as Stack Overflow. According to their own survey, in 2018 about 50 million people visit Stack Overflow each month to seek and give answers to technical questions.[12] One of the great benefits of participating in a technical community like Stack Overflow is in avoiding antipatterns. An antipattern is a solution to a problem

[11] https://www.frontiersin.org/

[12] https://insights.stackoverflow.com/survey/2018/

that has negative consequences (Brown 1998, pp. 7-8). A common antipattern found in web applications is serving of dynamically generated web pages of data with in a way that prevents indexing by search engines. For example, if a web application framework, such as .NET or Java Servlets, retrieves data from a database and uses that to generate web pages behind a small number of URLs then search engines and their users may not be able to discover the content, even if login is not required. This related to prevention of linking described above. Therefore, public technical forums like Stack Overflow are another of the building blocks of open ecosystems.

4. BUDDHIST RESOURCES AND COMMUNITIES SERVED

The Buddhist online community is a union of the Buddhist monastic community, the lay community, the Buddhist academic community, and the interested public at large. The University of Tokyo has sustained a tradition of the study of East Asian Buddhist literature for over one hundred years. The University of Tokyo created the Taishō Shinshū Daizōkyō 大正新修大藏經 (Taishō canon), the main version of the East Asian Buddhist canon used by scholars today, over the period 1912-1926. Nearly a hundred years after that effort began in 2008, the University of Tokyo released the SAT digital version of the Taishō canon (Muller, Shimoda, and Nagasaki 2017, pp. 175-179). This effort was driven almost entirely by scholars. The Buddhist community in Taiwan has included more participation from monastics and lay people. Wilkinson describes the community of monastics, scholars, and lay believers who joined forces in the large effort for the development of the CBETA project for the initial scanning of the Taishō Tripitaka and development of the online canonical texts in Taiwan (Wilkinson 2017, pp. 160-162). Digital versions of the Korean Buddhist canon have been published as well (Lancaster 2010).

The initial digital versions of the Taishō canon took over ten years to develop, chiefly because the creators had to overcome challenges with coverage by the Unicode standard and also lack of optical character recognition (OCR) capabilities for Chinese characters at the time. Now that those initial foundational standards and capabilities are established subsequent capabilities are taking place more quickly. Both the SAT and CBETA allow integrated dictionaries and both inbound and outbound links to encyclopedic and bibliographic resources. The Bukkyo Dendo Kyokai (BDK) or 'Society for the Promotion of Buddhism,' founded by Yehan Numata (1897-1994) sponsors the translation to English of many of the texts in the Taishō.[13] These are integrated into the SAT website in the form of a parallel corpus.

Fo Guang Shan has published a number of online resources as well. These included print and online versions of the collected works of Venerable Master Hsing Yun, the founder of Fo Guang Shan; the Fo Guang Dictionary of Buddhism, a dictionary of over 32,000 terms with both print and online versions, and a set of canonical writings including later period writings from the Ming and Qing not available elsewhere. Fo Guang Shan has assembled a large team of translators that includes lay volunteers, monastics, full time staff, university faculty, and graduate students. General online tools that have recently become available, such as Google Docs and video conferencing, are enabling teams to scale out. The online resources mentioned are enabling more rapid progress.

5. PLATFORMS, STANDARDS, AND BUILDING BLOCKS

Free hosting services for open source repositories have been one of the most important building blocks for the open source movement. GitHub is based on the open source version control system git, first released by Linus Torvalds in 2005 in order to host the source code for the Linux kernel (Loeliger and McCullough 2012, loc. 254). GitHub was founded in 2008 and acquired by Microsoft in 2018 (Microsoft 2018). The unit of hosting on GitHub is the repository. A repository can be freely and instantly created by anyone with Internet connection. Changes to source code can also be pushed freely and instantly. However, the highly structured process defined by the git is critical in maintaining software quality.

[13] http://www.bdk.or.jp/english/english_tripitaka/publication_project.html

An ownership and review process enforced and the record of changes maintained are central to the release process. The software and base digital assets for the NTI Buddhist Text Reader is an example of a Buddhist project is maintained on GitHub.[14]

5.1 Standards

Developers of web platforms have been aware of the importance of standards since the beginning of the web. Standards for basic web development include HTML, extensible markup language (XML), and JavaScript Object Notation (JSON) emerged and evolved with the development of the web. At a higher level standards like the Text Encoding Initiative (TEI),[15] and Resource Description Framework (RDF) where developed for needs closely related to digital libraries. TEI is a standard that has been adopted the digital humanities community. TEI gives recommended structures for text corpora, bibliographies, and dictionaries. The Digital Dictionary of Buddhism (DDB) uses TEI for storage of the dictionary terms (Muller, Nagasaki, and Soulat 2012). The DDB benefited from a number of other standards, including XML and Unicode, since its initial creation in 1986.

The standards mentioned have been critical enablers for projects, such as DDB. However, standards and the committees that lead them move relatively slowly compared with the rapid movement of the open source community at large. Nevertheless, standards have been critical for the development of open ecosystems of modules.

5.2 Breaking Down Monolithic Systems

Software modules, also called components, are not easy to design. According to Bevacqua, the most important principle in module design is that a module should have a single responsibility (Bevacqua 2018, pp. 38-43). In addition, a module should be accessed via a well defined interface that is not coupled to its implementation. Also module should also be testable. When modules have these simple but hard to achieve properties and they are stored in publicly accessible package management systems then

[14] https://github.com/alexamies/buddhist-dictionary

[15] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

a high degree of automation is possible in a continuous integration / continuous delivery (CI/CD) pipeline. That is, software developers are continuously pushing software that is automatically downloaded, integrated, tested, and deployed in stark contrast to previous generations with waterfall processes where release cycles often took years.

In thinking about the requirements for various projects that the author has worked on for Fo Guang Shan, the author proposed building a digital library. A digital library would be able to combine many requirements into a single consistent home for users to discover and access everything necessary for their work. However, the problem with the proposal was its monolithic nature. Fo Guang Shan has many projects, which made a large project like a digital library a distraction. Digital libraries include a large set of requirements for submitting, cataloging, searching, and accessing collections of books and other digital assets (Xie and Matusiak 2016, loc. 569-700). However, modularity and participation by the software development community is not prominently discussed in the digital library literature. Today is not common for ebooks to link to specific pages in other ebooks. Rather ebooks are still using traditional citations. These points are in contrast to open systems of websites, many of which today are implementing features overlapping with digital libraries. However, digital libraries maintained by other parties are important tools for translators of canonical Buddhist texts especially for research of historic context. In summary, a digital library may be too big of a building block for practical development.

The principles of modular design can be illustrated using the author's experience with Buddhist dictionaries. There are many different types of dictionary: monolingual dictionary, bilingual dictionary, historic dictionary, specialist dictionary, and others (Atkins and Rundell 2008, pp. 24-26). The goal of the NTI Reader project is to aid Chinese users to read and translate Chinese Buddhist texts. Therefore the type of dictionary selected is a Chinese-English bilingual dictionary.

Human language is complex and historic texts are more complex but coping with this can be made efficient if the user tasks can be modelled with software components. An entry in a dictionary can have multiple word senses, which are nearly the same as lexical units as defined lexicography (Atkins and Rundell 2008, pp. 130-131). One requirement for the online dictionary software is also to underline or otherwise highlight certain terms, such as Buddhist terminology or proper nouns, in a passage of Chinese text to let the reader know the key terms. Another requirement from the dictionary software stakeholders is to display entries from multiple dictionary sources for a given word. Combining these requirements into a model, the author developed the object model for the chinesedict JavaScript component described in Table 1.

Table 1: Object Model for the chinesedict-js JavaScript Component

DictionaryBuilder	Retrieves dictionaries from the server, parses them, and makes the data available to the browser
DictionarySource	The source of a dictionary, including file location, the name of the creator, and a title
DictionaryView	Presents the dictionary to the user, such as for looking up a term
DictionaryEntry	The term to be looked up and the data in one of the dictionary sources
WordSense	One word sense for the term, including the pronunciation, part of speech, English equivalent, and notes

The developer of an application that uses the dictionary will supply multiple DictionarySource objects to the DictionaryBuilder, which build the dictionary and return a DictionaryView showing the highlighted words. When the user clicks on one of the words the DictionaryView will present a dialog showing the Term with one of more DictionaryEntry objects, each of which will have one of more WordSense objects. Although this appears complex, it is simpler that the TEI recommendation for dictionaries, for which a dictionary entry can include orthography, pronunciation, part of speech, senses, quotation, usage, etymology, related entries, and notes ("P5: Guidelines for Electronic Text Encoding and Interchange" 2018, ch. 9). This is somewhat analogous to a person reading a printed text. The person would have a collection of dictionaries: specialist Buddhist dictionaries, bilingual Chinese dictionaries, monolingual literary Chinese dictionaries, and specialist dictionaries of historic people and places. For interesting words encountered in the text, the person would consult the various dictionaries to decipher the meaning intended in this context. In summary, if user actions are modelled and developed as components, reading and translation of difficult texts can be made very efficient.

5.3 Modular components

Modular software development allows for convenient re-use of software in the form of libraries. The value of this has been recognized for many decades. What is new in the last several years is the combination of modular software systems with open source software and the emergence of platforms for hosting the modules. Various languages and platforms use terms other than 'module', often 'package' or 'container.'

Besides enabling source code revision control for Linux and related projects, git has been also been an enabler for modular software development in general. For example, the Go language, which was initially released in 2012, uses a package management system allows importing of third party packages retrieved from remote sources using git. This can be done using the 'go get [URL]' command.[16] This makes it very convenient to reuse open source Go packages hosted on GitHub. The Chinese Notes software[17] that powers the NTI Reader uses Go and is hosted on GitHub. However, the software is not organized in a way that allows other developers to access the packages with the go get command. There is an opportunity to refactor the code in a way that is more reusable by others. The step that is missing here is illustrative of one challenge of software reuse: it takes extra effort and careful planning to make your software reusable by other developers.

[16] https://golang.org/cmd/go/#hdr-Module_aware_go_get

[17] https://github.com/alexamies/chinesenotes.com

Container systems are another technology that has exploded in the last several years. The most prominent of these is Docker, first released in 2013. Containers have been revolutionary in enabling dependency management and efficiency in deployment and running of systems with open source software. The NTI Reader makes use of Docker containers but does not provide a container for other developers to conveniently reuse. The benefit to the NTI Reader of using contains are that they allow reliable operation of the web application in a cluster and easy rollback in case a code change is rolled out that breaks the application.

JavaScript module systems are one of the newest and most important developments impacting web application development. There has been rapid evolution of JavaScript or ECMAScript in the last few years via the ECMA standards body.[18] In 2015 the ES2015 release included major changes, such as changes in scoping, arrow syntax, classes, template strings, Maps, Promises and modules. These new features have been a substantial enabler of more complex and more powerful JavaScript applications and driver of change in the JavaScript ecosystem. Modern browsers have now adopted most of the recommendations in ES2015. However, the most startling phenomenon has been the rapid rise of the ecosystem of JavaScript modules. At the time of writing this paper, there were 883,140 modules hosted on the Node Package Management service npmjs.org with 8,897,268,546 downloads in the previous week.[19] The number of downloads is startling as an indication of the degree of automation in downloading by CI/CD pipelines.

In addition to ES2015 modules there are other JavaScript module systems, including CommonJS and the AngularJS module systems (Bevacqua 2017, pp. 296-297). NPM leverages the package.json[20] format that is used to describe publication of JavaScript modules. This greatly helps structure the download and use of the modules. The UNPKG[21] system freely distributes JavaScript packages to make them directly accessible to websites via a content distribution

[18] https://www.ecma-international.org/memento/tc39.htm

[19] https://www.npmjs.com

[20] https://docs.npmjs.com/files/package.json

[21] https://unpkg.com/#/

network (CDN). This is exactly the kind of amplification described Schwab above by enabling low cost, universally available, structured automated download and distribution of software.

5.4 Artificial Intelligence

Artificial intelligence, specifically machine learning, is becoming an important tool that can be leveraged for digital humanities projects, especially with natural language text processing. Three recent developments in machine learning are having a large impact on the development of Buddhist resources and digital humanities in general: (1) the improvement of deep learning methods for processing natural language, (2) the release of open source machine learning frameworks, and (3) the 'democratization' of machine learning through application programming interfaces (APIs) and services that do not require deep specialization in the field by the developers using them. This is really the same pattern as described for open source software in general as described above.

TensorFlow is an deep learning (artificial neural network) library released to open source by Google in 2015 and in 2016 became the most popular machine learning project on GitHub (Dean 2017). Keras is an open source project that wraps TensorFlow and other deep learning libraries to make them easier to use by software engineers not specializing in machine learning (Chollet 2018, pp. 29-30). The Colab[22] service hosted by Google provides a free hosted solution for running iPython programs that leverage machine learning software, particularly TensorFlow.[23] The iPython programming model encourages an iterative programming style where code and data can be viewed and save together. Colab sheets can be saved to GitHub and the output viewed directly by other users. Together TensorFlow, Keras, and Colab enable software engineers and data scientists to conveniently and cheaply develop machine learning applications and collaborate in a global community.

The NTI Reader uses machine learning for document search. The NTI Reader text search feature classifies documents as relevant

[22] https://colab.research.google.com/

[23] https://github.com/tensorflow

or not-relevant with a combination of vector space model document similarity scores (Zhai and Massung 2016, pp. 90-92). The scores are based on word and bigram frequencies. The scores are combined using machine learning with a technique called logistic regression by the open source library Scikit-learn (Pedregosa et al. 2011). The resulting Colab sheet was saved to the author's GitHub project to allow anybody to review the data re-run the code.[24]

The Buddhist academic community is investigating the use of machine learning to scale analysis of historic data beyond individual scholars manually examining data. For example, Bingenheimer discusses the use of named entity recognition for the discovery of the identities and references to people and places from corpora of historical East Asian texts, in particular the Digital Archive of Buddhist Temple Gazetteers[25] (Bingenheimer 2015). There are many obstacles to machine learning with historic Chinese text sources, including lack of digitization and lack of punctuation in the stream of text. An example of machine learning in Buddhist studies to overcome this is the automatic insertion of punctuation and optical character recognition in historic Chinese texts by researchers working with Longquan temple (Liu 2018).

While the early digital versions of the canon were manually typed from print editions (Muller, Shimoda, and Nagasaki 2017, pp. 177) later versions have been able to be digitized using OCR. Early OCR techniques worked adequately for printed Chinese texts enabling scanning of the Taishō, Yongle Northern, and the Qing Dragon canons (Fang, p. 210). Correction of scanning errors was a large task for these scanning projects. Recently, OCR techniques based on machine learning have been used to scan handwritten historic texts that were not able to be scanned with early generations of OCR technology.

Although important, the use of machine learning by the Buddhist academic community is trivial in comparison to the highly dynamic community of data scientists in competitions, such as those hosted

[24] https://github.com/alexamies/chinesenotes.com/blob/master/colab/querying_cnotes.ipynb

[25] http://buddhistinformatics.ddbc.edu.tw/fosizhi/

on Kaggle.[26] 600,000 new users joined Kaggle in 2017 for a total of 1.3 million members.[27] Such a large community will certainly drive improvements in machine learning that will have an impact on Buddhist studies.

To date machine translation has not been useful for translation of canonical and other historical texts. However, it is not inconceivable that this will change in the near future. Today, machine translation is already having an impact in the translation of academic publications in modern languages other than English. Medieval Chinese literary scholar Knechtges describes the challenges of translating historical texts in his essay on the translation of the important and large historic text *Wen Xuan*, which took over 14 years (Knechtges 1995, pp. 41-42). One problem was the variety of styles encountered considering the long historic period that the work covered. The sources for understanding the background to these styles themselves are untranslated historic texts. This resulted in the need to establish translations for large number of new terms. However, much of the historical context of these works is also discussed in modern Chinese sources, which can be accessed via machine translation.

In Buddhist studies, many Western Buddhist scholars have learned Japanese to access the large body of modern Japanese academic literature on Buddhism. Translation of this body of literature from Japanese to English and other languages is also possible with machine translation. Starting in 1984, Japanese journal articles on Buddhism have been indexed in INBUDS and these are linked from SAT (Muller, Shimoda, and Nagasaki 2017, p. 176).

Some Fo Guang Shan translation team members have experimented with machine translation for modern text. While the results is not directly suitable for publication, it can be helpful as a first step.

5.5 Mobile Accessibility

For online resources to be accessible to a wider community they

[26] https://www.kaggle.com/

[27] http://blog.kaggle.com/2017/12/26/your-year-on-kaggle-most-memorable-community-stats-from-2017/

should be accessible on mobile devices.One issue that has arisen is the accessibility of web sites to mobile devices. For example, some websites display very small text on mobile devices or rely on behavior, such as mouseover, with no mobile equivalent. This led to an explosion of mobile applications. However, native mobile applications behave as islands and do not work well in open ecosystems. Responsive web applications are websites that can be accessed on both workstations and on mobile devices without substantial degradation of experience. In the last several years responsive web application frameworks, such as Material Web from Google,[28] have been developed and released to open source. A best practice is to develop web applications rather than native mobile device applications and leverage web component frameworks to enable use on mobile devices. In that way, the websites can function as building blocks for open ecosystems on mobile devices as well as workstations.

ACKNOWLEDGEMENTS

The author is employed by Google in Mountain View, United States and volunteers at Fo Guang Shan. None of the views in this paper represents either organization. The author wishes to thank Fo Guang Shan for encouragement and including him in the translation team. All of the costs for the projects mentioned as created by the author were provided from the author's own financial resources.

ABBREVIATIONS

API: application programming interfaces

CBETA: Chinese Buddhist Electronic Text Association 中華電子佛典協會

CI/CD: Continuous integration / continuous delivery

DDB: Digital Dictionary of Buddhism

ECMA: European Computer Manufacturers Association

ES: ECMAScript

HB: Humanistic Buddhism

[28] https://material.io/develop/web/

NPM: Node Package Manager

NTI: Nan Tien Temple

OCR: Optical character recognition

SAT: Taishō Shinshū Daizōkyō Text Database

TEI: Text Encoding Initiative

References

Atkins, B. T. Sue, and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bevacqua, Nicolas 2017, *Practical Modern JavaScript: Dive into ES6 and the Future of JavaScript*, Sebastopol, CA: O'Reilly Media, Inc.

Bevacqua, Nicolas. 2018. *Mastering Modular JavaScript*. Sebastopol, CA: O'Reilly Media, Inc.

Bingenheimer, Marcus. 2015. "The Digital Archive of Buddhist Temple Gazetteers and Named Entity Recognition (NER) in Classical Chinese." *Lingua Sinica* 1 (1): 8.

Brown, William J. 1998. *AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis*. Wiley.

Chollet, François. 2018. *Deep Learning with Python*. Manning Publications Company.

Dean, Jeff. 2017. "The Google Brain Team — Looking Back on 2016." *Google AI Blog* (blog). January 12, 2017. http://ai.googleblog.com/2017/01/the-google-brain-team-looking-back-on.html.

Eger, Thomas, and Marc Scheufen. 2018. *The Economics of Open Access: On the Future of Academic Publishing*. Cheltenham, England: Edward Elgar Publishing.

Fang, Guangchang. 2017. "Appendix Defining the Chinese Buddhist Canon Its Origin, Periodization, and Future." In *Reinventing the Tripitaka: Transformation of the Buddhist Canon in Modern East Asia, Edited by Jiang Wu and Greg Wilkinson*, translated by Zi Xin and Wu Jiang, 187–215. London: Lexington Books.

Knechtges, David R. 1995. "Problems of Translation: The Wen Husan in English." In *Translating Chinese Literature,* edited by Eugene Chen Eoyang and Yaofu Lin, 41–56. Bloomington: Indiana University Press.

Lancaster, Lewis. 2010. "From Text to Image to Analysis: Visualization of Chinese Buddhist Canon." *Digital Humanities 2010,* 184.

Liu, Jiefei. 2018. "Longquan Temple Is Using Artificial Intelligence to Organize and Spread Buddhist Scriptures." *Technode,* 2018. https://technode.com/2018/07/09/longquan-temple-techcrunch-hangzhou/.

Loeliger, Jon and Matthew McCullough, 2012. *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. 2nd ed. O'Reilly Media, Inc.

Microsoft. 2018. "Microsoft Completes GitHub Acquisition." The Official Microsoft Blog. October 26, 2018. https://blogs.microsoft.com/blog/2018/10/26/microsoft-completes-github-acquisition/.

Muller, Charles, Kiyonori Nagasaki, and Jean Soulat. 2012. "The XML-Based DDB: The DDB Document Structure and the P5 Dictionary Module; New Developments of DDB Interoperation and Access." *Chung-Hwa Buddhist Journal* 25: 105–28.

Muller, Charles, Masahiro Shimoda, and Kiyonori Nagasaki 2017, "Chapter 7: The SAT Taishō Text Database A Brief History," In *Reinventing the Tripitaka: Transformation of the Buddhist Canon in Modern East Asia, Edited by Jiang Wu and Greg Wilkinson,* 175–185. London: Lexington Books.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Raymond, Eric S. 2001. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, CA: O'Reilly Media, Inc.

Schwab, Klaus 2017, *The Fourth Industrial Revolution*, New York:

Crown Publishing Group.

"P5: Guidelines for Electronic Text Encoding and Interchange." 2018. TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/.

Wilkinson, Greg 2017, "Chapter 6: The Digital Tripitaka and the Modern World," In *Reinventing the Tripitaka: Transformation of the Buddhist Canon in Modern East Asia, Edited by Jiang Wu and Greg Wilkinson,* 155–74. London: Lexington Books.

Xie, Iris, and Krystyna Matusiak. 2016. *Discover Digital Libraries: Theory and Practice*. Amsterdam, Oxford and Cambridge: Elsevier.

Zhai, Chengxiang, and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool.