Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Personal View:

Challenges in sustaining the Million Book Project, a project supported by the National Science Foundation

Gloriana St. CLAIR
Director, Universal Library Project
Dean, Carnegie Mellon University Libraries, Pittsburgh,
Pennsylvania, USA
E-mail: gstclair@andrew.cmu.edu

doi:10.1631/jzus.C1001011

One of the main roles I have played as a director of the Universal Digital Library has been to write grant proposals to support our work. Both for this project and for another project, Olive.org, an archive of executable content, how to sustain the final product is the most difficult challenge. This paper discusses the various models that might be adopted to sustain a large corpus of digital material, such as that of the Million Book Project. Methods discussed here include government funding, foundations and nonprofits, university homes, and joining existing projects. All individuals working with large digital projects should be concerned about how their work will be kept available to the public.

Government funding. Many of the partners in this project have benefited liberally from government funding. The Chinese partners have had significant government support through several successive Ministry of Education five-year plans. The Indian government has supported the project with funding for language translation research projects. The Egyptian government funded the creation of the Bibliotheca Alexandrina and continues to contribute to it. In the U.S., the National Science Foundation supported equipment, travel, and meetings.

This support has been essential to the creation of this large corpus of material. The governments very wisely invested in bringing educational and cultural resources to a large segment of their constituents. The disadvantage is that, as government budgets tighten, the funding necessary to sustain a project can be lost.

One great advantage of government funding is that the government wants to serve the whole public. Beginning this year, in the U.S., the National Science Foundation now requires principal investigators to explain how the data they have collected will be made available to the larger research community and how it will be sustained. In the U.S., the government also wants free-to-read access and at the same time allows creators to charge for enhanced versions.

Foundations and other not-for-profit organizations. Foundations, like the government, are excellent sources of support for the initiation of a large digital project. They have the vision to see what could be accomplished by increasing progress in selected disciplines, such as high-energy physics and astrophysics, and broadening the availability of educational resources. JSTOR and ArtStor are two resources initially supported by the A.W. Mellon Foundation.

The Qatar Foundation gave funding to create the Qatar Arabic and Islamic Heritage digital collections. Because that collection so actively reflects the country and region's culture and because the Qatar Foundation is so focused on educational goals, they are more likely than other foundations to sustain it. Other foundations, such as A.W. Mellon, require that sustainability models be explained before they will fund the initial project. Mellon has been particularly focused on the issue of sustainability.

Some electronic products and services found in U.S. academic libraries are licensed through consortia and some come from not-for-profit organizations. One of the more popular ones is JSTOR, a database of articles in journals in a wide variety of fields.

Originally, all the articles in this database were five years old or older, but this year, some publishers have begun putting more current material into JSTOR. The Online Computer Library Center (OCLC) is another prominent not-for-profit organization. Each of these organizations does realize enough ‘profit’ to grow and to maintain a significant reserve.

These organizations fund themselves by selling subscriptions, services, and products. In OCLC’s case, a membership fee also exists. This approach has been most successful because libraries need the content provided and can pay the fees necessary. Chinese partners have created a licensed resource and the inclusion of the Million Book Project books in that resource provides a good sustainability plan for that part of the corpus. Of course, when materials become licensed, they are often no longer free to read. The challenge of a licensed database is that a significant organization may be required to select and administer the resource, unless the corpus can be placed with an existing organization.

University homes. The initial vision we had for sustaining the Million Book Project was that it would have a permanent home in the School of Computer Science. The Universal Digital Library (UDL) directors observed that the price of storage was falling steeply and thought that, even though the corpus was large, funding would be available to purchase storage. However, storage was not the only resource needed to sustain the corpus. A system manager to curate the data—to ingest, backup, regularly review, and respond to queries—was also needed. When that position was lost, graduate students began to fill in, but their primary attention is elsewhere. The result did not meet standards for persistent access. To date, the libraries, which are committed to long term, 24/7 access, have not had the resources to be able to step up to this challenge.

One particularly successful example of a large, extremely popular digital resource is arXiv, a repository of preprint articles in high-energy physics and related fields. With the leadership of Paul Ginsparg, the repository was originally created at Los Alamos with government funding. The free-to-read nature of this article repository does foster efficient progress in the field. Librarians who were concerned for its sustainability were relieved when Cornell University gave arXiv a more permanent home. However, this year, arXiv has begun aggressively asking for academic libraries to contribute to arXiv’s upkeep. Thus, this project, initially funded by the government, then hosted by a university, now appears to be moving towards a subscription-like model.

Universities have much to offer as homes for digital projects because historically they have been stable. As a creator of new knowledge, which inevitably is related to and derives from older learning, universities, and especially their libraries, care about the preservation of knowledge. Nevertheless, resources for funding are scarce and are expected to continue to be scarce.

Joining existing projects. Another option is to join an existing digital project that has already solved the sustainability problem. Three alternatives are Wikibooks, Open Content Alliance, and the Google Books Project.

1. Wikibooks. According to its Web page, Wikibooks is a collection of textbooks. If they are ingesting only textbooks as content, then only a small fraction of the existing million book corpus would be ingested. As part of Wikipedia, Wikibooks is a nonprofit and appears to rely on contributions to sustain it. As long as it remains the preeminent online ‘pedia’, it may be sustainable. The free-to-read model is characteristic of Wiki resources.

2. Open Content Alliance (OCA). OCA is also a nonprofit, associated with the Internet Archive. Brewster Kahle has long been a partner and fellow traveler with the Million Book Project. At the founding of OCA, he ingested materials collected from India and those materials are still part of OCA. At our 2007 Pittsburgh meeting, the partners agreed to become a part of OCA, but OCA has not actively followed up on that decision. Certainly, the Internet Archive does plan on sustaining itself long term.

3. Google Books. The U.S. directors of the UDL project all believe that giving Google non-exclusive access to our corpus is the best alternative. We believe that not only would the corpus be maintained long term but also that the materials would receive maximum use because of the popularity of the Google search engine. Many research studies show that U.S. students and faculty both go directly to the Web and a majority of them directly to the Google search engine as their first source of information. Placing our content where it can be most easily found and used will

be the most successful means of achieving our original goal.

Google is an extremely successful for-profit company whose corporate philosophy mirrors that of the Million Book Project. Their aim is “to organize the world’s information and make it universally accessible and useful” (Google Books Mission, available from http://books.google.com/googlebooks/agreement/#6).

They do make money through advertising from the over five million volumes they have already digitized. This revenue stream provides both an incentive and a practical resource for the sustenance of Google Books. The Google collection grew from digitizing books at their partner libraries, with the University of Michigan contributing the most volumes. Our Chinese and Indian language books would complement the existing western-focused collection.

The consensus is that there are perhaps 100 million books in the world. This figure is based on the size of the OCLC Worldcat database and a perception that worldwide, many, many books exist which are not in U.S. and European libraries, whose collections are Worldcat. A non-exclusive partnering with Google Books at this time would be significant as compared with a later time when the database is much larger.

Google and the Google Books Project have many critics. Three major issues that are discussed briefly below, are privacy, net neutrality, and the Google Book Settlement.

Privacy is a valid major concern because the success of Google’s marketing of advertising revolves around the company’s ability to target ads to those who have demonstrated interest in the product. Google does track, analyze, and profit from its knowledge of individual interests. One of the values of U.S. libraries has been to protect individual privacy. U.S. libraries typically do not reveal the searching and check-out behavior of their constituents; we are even careful to erase check-out records for our integrated library system so that we cannot be forced to divulge this private information. Perhaps many would find these values and practices old fashioned.

Certainly, the advent of social networking tools, such as Facebook, Twitter, and YouTube, represents a different approach to personal information. These tools encourage the sharing of private information at a level some would consider both profligate and tedious. Societal norms around privacy issues are changing, and in that changed environment, individuals seem willing to exchange personal information for focused information, including advertising, on areas of interest.

Net neutrality is a stance that libraries and computing organizations have taken vis-a-vis the governance of the Web. These organizations argue that research libraries and higher education institutions are enormous providers of content and applications. The information thus provided fosters research, creativity, and education, and should be allowed to flow freely. They believe that Verizon and Google would like to prioritize content from their affiliated and fee-paying sources relegating other content to a slower delivery system. They believe that Google-Verizon want to establish a second Internet with expensive, discriminatory wireless services to those who can pay primarily deriving from their paying sources.

Conversely, an Engadget article by Nilay Patel (August 5, 2010) reports that Google CEO Eric Schmidt said repeatedly on the call that “Google will never pay for prioritized access and Google products would remain on the public Internet” (http://www.engadget.com/1020/08/09/google-and-verizons-net-neutrality-proposal-explained). Schmidt portrayed Google as a watchdog on Verizon to make sure that nothing untoward happens to the public Internet.

Google Book Settlement is still unresolved as of September, 2010. Several issues continue to concern those engaged around the creation of digital libraries. My colleague, Denise Troll Covey, identified these concerns about the settlement:

(1) Library partners signed on to pursue the legality of snippets as fair use, yet Google now proposes a different schema. Fair use may be weakened.

(2) Machine (Non-consumptive) use is restricted to research and researchers that Google approves.

(3) Google continues to make machine (non-consumptive) use of content they scanned but do not include in the Google Books database because the copyright owners opted out or brought law suits (e.g., France, Germany).

(4) Google’s proposed solution to the orphan works problem makes it unlikely that Congress will pass orphan works legislation that will be equitable; the proposed orphan works fiduciary will NOT have

the power to make orphan works available open access.

(5) Proposed settlement gives Google—and only Google—a license to break copyright laws, in effect creating a copyright regime for Google and another copyright regime for everyone else.

(6) Academics were not adequately represented in the class action settlement (email between St. Clair and Troll Covey, August, 2010).

Each of these points has validity and should be considered carefully.

Perhaps the larger issue around sustainability is the long-term prospects of Google. That issue comes to the center of Google as a recommended solution to million book project sustainability issues. The financial health of Google may revolve around the outcome of the book settlement with the publishers. If the ruling were to be adverse, Google's current robust financial situation could be effected. The company might no longer be a favored choice for sustaining the corpus of the million book project.

Conclusions

This paper offers several choices of types of institutions as long-term sustainers—governments, foundations and nonprofits, foundations, universities, and commercial companies. Some of these institutional types have established records of being able to accommodate societal change. Governments have a mixed record around longevity. Nonprofits and foundations also have less robust track records, although the Catholic Church is over 2000 years old. Universities, for instance, have existed in a recognizable form since the founding of the University of Bologna in 1088. Similarly commercial entities can also demonstrate longevity. According to Wikipedia, several firms in Japan began in the 700s and one construction company claims 578 as its founding date. Yet, many, many companies, and especially technology companies, such as Google, are short-lived. Some fail, some merge, and some are bought out. These casual historical examples would suggest that either a company or a university institution type would be suitable for long-term sustainability.

The Million Book Project represents the cooperative work of hundreds of people in several different countries. Our vision was to demonstrate to the world that large-scale digitization could increase the amount of useful knowledge available on the Web free to read for students and scholars around the world. If our vision is to continue, then we must select a good model to sustain our work.