2010

Open Access and Digital Libraries: A Case Study of the Text Creation Partnership

Shawn Martin
University of Pennsylvania, shjmarti@indiana.edu

Open Access and Digital Libraries: A Case Study of the Text Creation Partnership

Abstract
Many people operate under the assumption that Open/Closed access is a binary proposition. Either the material is available to everyone on the web or it is closed to a limited number of subscribers. The reality, however, is much more complicated. What is the use of a digital library, no matter how open, if it is unable to sustain and maintain itself over time? What is the point of a well funded collection that is closed to the people who need it most? There are in fact many models for maintaining both open and closed access digital libraries. Though the conversation often focuses on the furthest ends of the spectrum (greedy publishers extorting money to content, or, conversely, benevolent academics making knowledge freely available to the world via grants), there are in fact many models that are in between these extremes that exhibit characteristics of both closed and open access models. In particular, the Text Creation Partnership (TCP) tries to work with commercial publishers to create a middle road between these extremes. By investigating the many types of open and closed access models, and seeing how models like the TCP fit in this landscape, it is possible to make better determinations on how to build digital libraries in the future. How should the community come together to find a more moderate path, and what will that road look like?

Comments
This paper is also available in a Bulgarian translation at http://www.fatcow.com/edu/library-papers-bl/

Open access and digital libraries: a case study of the Text Creation Partnership

Abstract:

Many people operate under the assumption that Open/Closed access is a binary proposition. Either the material is available to everyone on the web or it is closed to a limited number of subscribers. The reality, however, is much more complicated. What is the use of a digital library, no matter how open, if it is unable to sustain and maintain itself over time? What is the point of a well funded collection that is closed to the people who need it most?

There are in fact many models for maintaining both open and closed access digital libraries. Though the conversation often focuses on the furthest ends of the spectrum (greedy publishers extorting money to content, or, conversely, benevolent academics making knowledge freely available to the world via grants), there are in fact many models that are in between these extremes that exhibit characteristics of both closed and open access models. In particular, the Text Creation Partnership (TCP) tries to work with commercial publishers to create a middle road between these extremes.

By investigating the many types of open and closed access models, and seeing how models like the TCP fit in this landscape, it is possible to make better determinations on how to build digital libraries in the future. How should the community come together to find a more moderate path, and what will that road look like?

Introduction

"In considering how best to organize the publishing side of scholarly communication, it will also be important to be open to new business models" (Unsworth et. al 2006 32). Most recently, many new business models being discussed revolve around Open Access which, according to Peter Suber, means that the resource is "digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber). Recently, the discussion about Open Access has revolved around how open a resource is. On the one hand, most academic grant funded projects are open and freely available to the world. Yet such projects often tend to be small in scope and dependent on the dedication of one faculty member or research group at a particular university. On the other hand, commercially funded databases like Early English Books Online (EEBO) from ProQuest Information and Learning, among many others, tend to be extremely large in scope and less dependent on the commitment of faculty members. Nevertheless, commercially produced databases also tend to be very expensive and limited only to small numbers of people (those who belong to research institutions able to pay the large subscription fees required). So, ideally the academic community would like to have the best of both worlds, a large comprehensive database of research material that is open and freely available to the world. How is it possible to create such a thing?

The answer can be found in one word, sustainability. Any electronic resource, regardless of whether it is Open Access, cannot survive without monetary and community support. These two things are essential to sustainability, yet they are extremely elusive. The key is to create a business model that captures all of the desirable features and that, rather than falling toward one extreme (closed access commercial model) or the other (Open Access academic model), finds a middle ground in which the resource is mostly freely available and is sustainable over time.

Many models have attempted this in various ways. Of course, the largest and most well known is Google, a massive digitization project that is now in the midst of legal settlements with the Authors' Guild, the Association of American Publishers and many others about the exact nature of what can and cannot be scanned and given away freely. Additionally, national governments like France have sponsored national digital library programs like Gallica. Finally, there are projects driven by volunteer labor such as

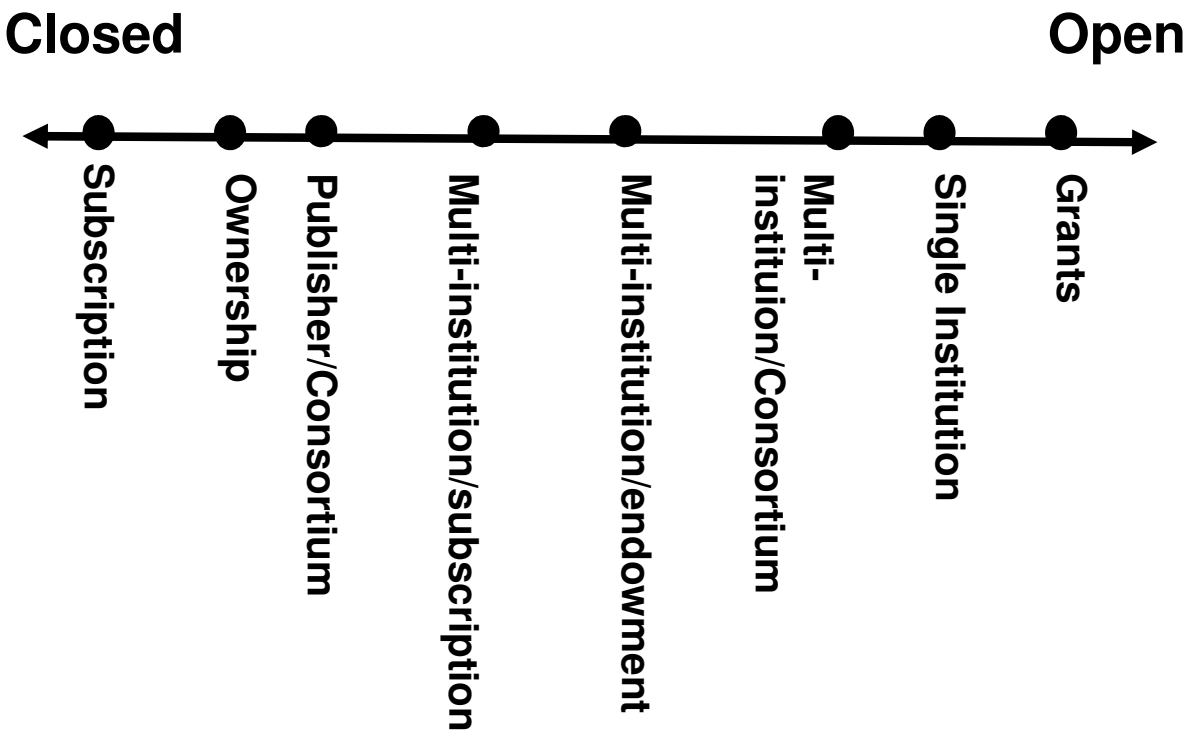
the Gutenburg project and its distributed proofreaders. All of these projects have their strong points. Google especially may fundamentally change the way that scholars and librarians think about digital content creation. Nevertheless, much of the legal work is still in negotiation and it remains to be seen how Google's digitization will affect electronic collections. Gallica is also intriguing, but not entirely applicable to countries like the United States where government funding of that magnitude seems unlikely. Gutenburg is a tremendous asset to the world, but, it could be argued, has limited utility for advanced textual scholarship where it is paramount that the text be completely accurate and editions be verifiable. So, there is definitely more work to be done to discuss how projects like Google, Gallica, and Gutenburg are useful to scholarship, but largely outside the scope of this article. Rather, this paper will focus on the myriad of digitization projects now being undertaken by scholarly projects and libraries at universities primarily in the US (and to an extent in the UK and Canada) and how to make such projects sustainable over time.

In the United States, as in other countries, there has been much talk recently about how to create such a model. Recently, Ithaka (an organization funded by the Mellon Foundation and dedicated to researching organization and accelerating the productive uses of information technology for the benefit of higher education) released a report on sustainability and revenue models which provide some guidance on this issue. It identified two large categories and several subcategories of revenue models (Guthrie et. al. 2008). Though Ithaka's report is helpful in thinking about revenue models, it does not fully capture all of the arrangements that universities have made for scholarly resources, particularly in the humanities. Currently there seem to be at least eight broad types of model that universities have used for digital libraries in the humanities:

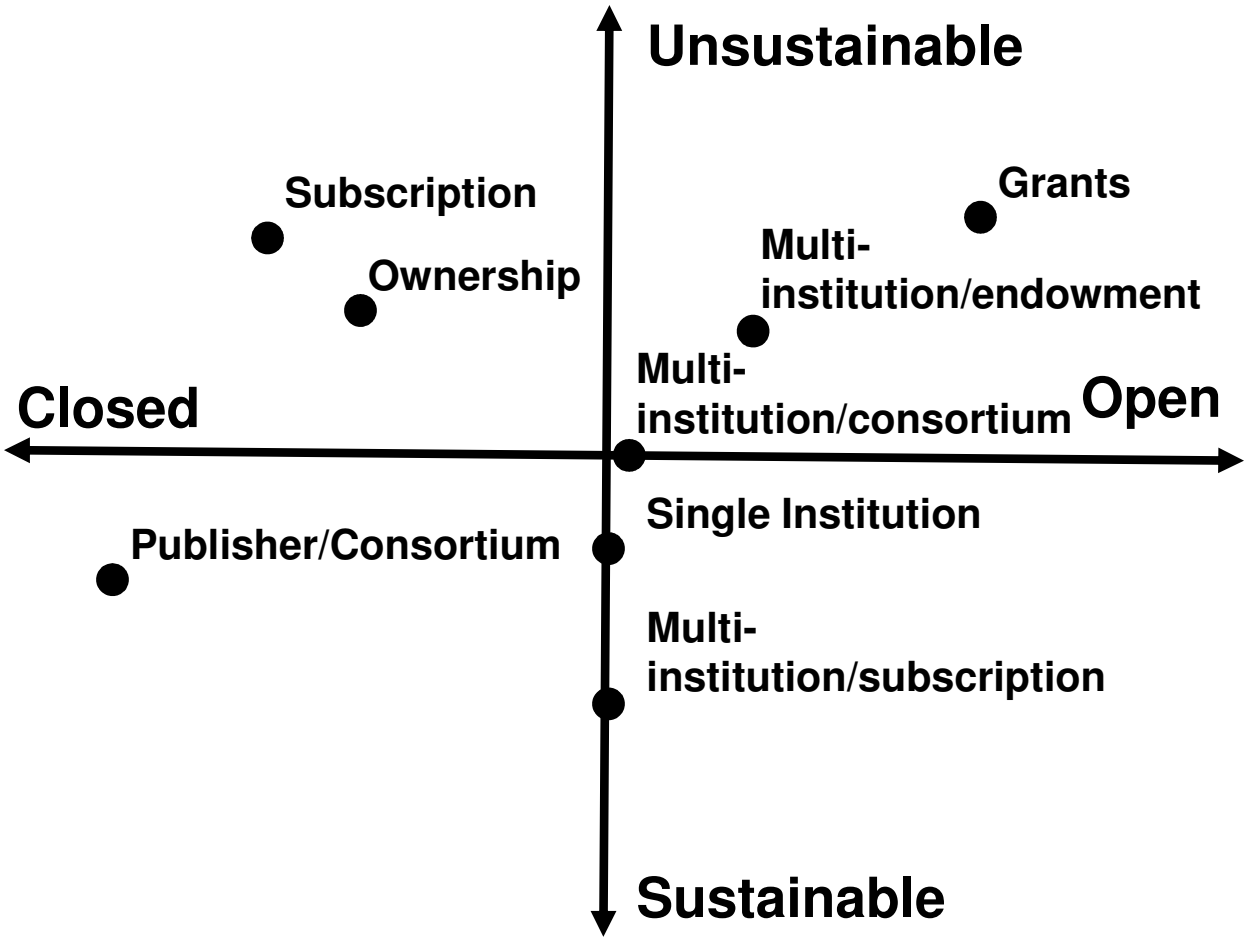
1. *Subscription* – the university pays the publisher for access to a resource it has created
2. *Ownership* – the university pays the publisher for rights broader than just access to the resource it has created
3. *Grants* – the university gets a grant to pay for the creation of the resource it wants to use
4. *Single Institution* – the university supports creation of the resource it wants to use with internal funds
5. *Multi-institution/Consortium* – multiple universities cooperate to build the resource they want to use
6. *Publisher/Consortium* – universities cooperate with the publisher (usually a university press) to create a resource they want to use
7. *Multi-institution/Endowment* – universities contribute to a common endowment for access to the resource they want to create and use
8. *Multi-institution/Subscription* – universities pay the publisher a subscription to a resource they want to use and that eventually becomes Open Access

By discussing the strengths and weaknesses of each of these models, particularly in the United States, and highlighting one particular model, the Text Creation Partnership (TCP) that employs aspects of all of these techniques, I hope to suggest how the community can arrive at a more complex picture of the ways in which the Open Access environment works.

Generally, arguments tend to discuss such models in a binary way like this:



In reality, we need to adopt a more complex graph that would recognize that in addition to a resource being open or closed, it can also be sustainable or unsustainable. Therefore, a resource that is open might not be sustainable and a resource that is closed might be unsustainable. Such a graph might look something like this:



Ideally, any new project should try not to go to any of the extremes, but rather remain in the gray area between them. TCP is just one of the projects that attempts to do this, and, arguably could be a model for future digital library development.

TCP Background

In 1998, ProQuest Information and Learning published the Early English Books Online (EEBO) database containing images of every book printed in England or in English between 1470 and 1700, amounting to roughly 125,000 titles, or in essence the works listed in the Short Title Catalogs I (Pollard and Redgrave) and II (Wing), the Thomason Tracts, and the English Tract Supplement. These were scanned images of their already existing microfilm collection. The publication of EEBO was naturally a major step forward in electronic availability of primary source titles. Nevertheless, at least for the University of Michigan library, it did not present a major step forward in holdings. The university library already held all of these titles in microfilm, and, although the ability to view individual titles from one's home computer, have multiple views of the same book, and so forth, were great access advantages, they did not add greatly to the existing collection.

Librarians at Michigan felt that the true value for this collection lay not in digital facsimiles, but in the possibility of full text searching. That way, researchers and students could search individual words or concepts across titles and engage with the sources in ways previously unimaginable. ProQuest saw the

advantage of adding full text, but felt that the cost of converting these images into full text would be millions of additional dollars and add so tremendously to the price of the product that libraries would be unable to afford it. Thus the TCP project was born. Rather than taking no for an answer, the University of Michigan felt that it could get support for the creation of full text versions for at least a subset of the EEBO collection. Under the leadership of Mark Sandler, then collection development officer at the University of Michigan Library, ProQuest agreed to partner with the University Library for the purpose of creating full text for a subset of 25,000 titles in the initial phase, with the understanding that the project might continue depending on the support it got.(Sandler 2004, 4-6)

Another important point to add about the TCP project is that it allowed both commercial publishers and academic libraries to compromise and get something they wanted. For libraries, it is important that these texts become publicly available to the world, not just paying subscribers. So, all of the texts that TCP creates will enter the public domain after a period of five years. During those five years, the commercial publishers have exclusive sales rights and the ability to develop specific tools to search the text that TCP creates. This creates a great opportunity for them to recoup their investment and generate new sales because of increased functionality.

What is unique about the TCP initiative though is not so much the partnership between private and public enterprises; rather, its unique structure and new prototype for cooperation between university libraries, the academic community, and commercial publishing is the most important aspect of the project. The TCP is not a traditional grant-funded project but a partner funded initiative that seeks library contributions for the creation of full texts. Additionally, a full text that TCP creates is not just another product, but a benefit to the academic community. All texts created by the TCP also enter the public domain. So, in essence, universities are paying for texts which they own and will have the ability to distribute beyond their own campus communities rather than just having ownership of the file for their own local and restricted use, as they would for any other commercial product. This model has been largely successful with (at time of writing) nearly 20,000 texts available. In fact, it has been successful enough to be extended to two other similar commercial databases namely, Evans Early American Imprints (Evans), a collection of every work printed in Britain's American colonies and later the United States between 1639 and 1800 (based on the Evans Bibliography) available from Readex Incorporated and Eighteenth Century Collections Online (ECCO), a database containing over 150,000 titles printed in Britain between 1700 and 1800 available from Thomson-Gale. EEBO-TCP has begun a second round of funding to complete the remaining 25,000 unique titles.

Essentially, how the model works is each partner institution according to its size contributes a set amount of money to the TCP and that contribution is matched by the commercial publisher. All of the money that is collected goes directly to creating texts. So the more money that is collected, the more texts the project is able to create, and the less expensive each text becomes for each contributing partner. In all, this model has allowed TCP to sustain a budget of about $1.4 million per year over approximately 7 years, far more than any grant funded or institutional model would be able to do. Also, these funds allow the project to create text to a fairly high standard, each text is transcribed twice by two different people, then any discrepancies are corrected by a third person, and they are then reviewed by a group of experts from the University of Michigan, Oxford University, the University of Toronto, and the National Library of Wales. Though there are certainly some mistakes that slip through, the important thing is that each text is as accurate as it can be given the constraints that TCP faces.

It is hoped that the EEBO-TCP model can be extended even further to other similar collections. There are a few important caveats to this. Firstly, ECCO actually does have some full text searchability. Gale used Optical Character Recognition (OCR) software to generated text files from the images of books in its collection. Necessarily, whenever a printed page or the digital image of it was blurred, or either of them had the printed lines not quite horizontal, or contained unusual (such as non-Latin alphabet)

characters, the OCR process produced imperfect results. Thus there is an argument that at least a portion of ECCO needs to be accurately transcribed for scholarly use. Secondly, only around 75,000 out of roughly 125,000 texts in EEBO will be transcribed by TCP. The remaining 50,000 titles are largely reprints or later editions of the same titles. Though some would argue that those titles need to be transcribed as well, in the majority of cases we would simply be reproducing an existing transcript, since in this period most reprints of books simply copied their predecessor edition without introducing significant changes. Since images are already available within EEBO to those scholars who need to see the small differences that were introduced in reprinting, transcriptions of such reprints are not a top priority. In all, these works comprise a seminal corpus of primarily English-language materials, although other languages are also included, and represents a wealth of primary source material for every avenue of scholarly endeavor. The TCP hopes to create a cross searchable, public domain collection from these (and perhaps other) databases that will form a vast base of material for digital library development for many years to come.

Strengths and Weaknesses of Various Models

Generally all of the above models fit into two larger categories. First there is the commercial model, meaning products usually produced by publishers looking to sell access to the content. Second is the institutional model, meaning projects usually produced by faculty at a particular university and not seeking to raise income. Both of these larger models have subcategories each with their own advantages and disadvantages. Nevertheless, TCP is a model that really does not fit into either category. Rather, it is a kind of hybrid model of which there are few examples. Though this model is complex, it is also more likely to be the kind of model which is successful in an increasingly complicated electronic environment.

Commercial models have the advantage of being able to produce large amounts of material in a relatively short amount of time. Because of the relatively large amount of money publishers have (compared to universities), the interface and database system usually have better functionality than those produced at universities. Also, commercial products have better marketing and outreach infrastructures behind them. So it is easier to get the word out about new developments. Yet, commercially produced databases also tend to restrict access, and sometimes very heavily. There tends to be less scholarly input in them, because of the amount of time it would take to involve multiple scholars and librarians, and commercial databases do not always improve their systems as technology changes because once a publisher has sold the content to a university and made a sufficient profit, it is in the publisher's interest to move on to new revenue-generating projects. Examples of such commercial models would include:

> *Subscription* – This is a model parallel in many ways to what libraries and publishers did in the print world. In essence, the library pays a subscription and the publisher then provides access to the campus community to a particular set of titles. This has the advantage of being a straightforward and familiar arrangement. However, as opposed to the print world, libraries actually do not own any of the content and are in essence "renting" it from a publisher. Therefore, librarians and the researchers that use them do not have the same abilities to copy, loan, or use electronic materials in the same way that they do with printed books.
>
> *Ownership* – This model is similar to the subscription model. However, it allows libraries broader rights over particular materials. For instance under this kind of arrangement, a library might maintain rights to copy or print materials outside of the library or to make backup copies for the use of scholars and students.

Institutional models are familiar in humanities departments. Scholars are used to getting grants to complete a book or a project which results in an exhibition or special issue of a journal. In the digital world, these models have been migrated into funding online databases of primary resource materials,

similar in some ways to commercial products like EEBO. These models have the advantage of being open to the world, have significant input from the scholarly community and are therefore of greater utility for scholars and students. Yet, they also tend to be fairly small and narrowly focused. So they may be quite useful to scholars in particular areas of study, but not so useful to researchers outside of the field. Examples of institutional models would be:

> *Grants* – This is probably the most common model within the humanities. A funding agency such as the US National Endowment for the Humanities (NEH), the Canadian Social Sciences and Humanities Research Council of Canada (SSHRC) and the UK Joint Information Systems Committee (JISC) gives a set amount of money to a faculty member or group of faculty members to create a database of materials for their area. When that money has run out or the faculty member leaves, however, these projects often die or are unable to garner the same amount of interest that they once did.
>
> *Single Institution* – This model is similar to grants in that an institution is giving money to a center or program that in turn creates digital resource material. It is more stable than a grant because the institution generally makes a commitment to maintain the program over a long period of time. Nonetheless, institutional models still tend to be small and narrowly focused because institutions do not have the same amount of purchasing power as a large commercial publisher.

These four models tend toward the extremes of the Open/Closed Access spectrum with institutional models on the open side and commercial ones toward the closed. Recently though, both publishers and universities have been experimenting with other kinds of models that attempt to combine aspects of both. Such experimenting allows economies of scale so that universities can create larger resources in a way similar to commercial publishers and allow these resources to remain open. Publishers, realizing the trends toward Open Access, have also collaborated with universities to create publishing models that they hope will allow them to make a profit while at the same time adhering to the demands of their customers. Examples of these kinds of hybrid models include:

> *Multi-institution/Consortium* – In this model, institutions come together to produce a large database of material. By pooling their resources, it is hoped that they will be able to rival large commercial publishers and to be able to maintain open access. The Open Content Alliance and the Internet Archive demonstrate some ways in which consortia of libraries are coming together to create electronic resources. These projects do have the advantage of creating larger collections but still tend not to have the functionality of commercial databases because of their relative lack of expertise in interface design and back-end production.
>
> *Publisher/Consortium* – Most often, universities have collaborated with university presses and scholarly societies on projects like the Humanities E-Book project. Again these projects are larger than grant and institutional electronic resources and often will have better interfaces, but are still significantly smaller than large commercial databases.
>
> *Multi-institution/Endowment* – The most well known example of this type of model is the Stanford Encyclopedia of Philsophy which brings multiple institutions together to create an endowment which will then allow the encyclopedia to be maintained in perpetuity. Though this model may work for some seminal resources in the humanities, it is also expensive and probably not sustainable. It would be impossible for the community to come together to support a similar endowment for every electronic resource needed for humanities research.
>
> *Multi-institution/Subscription* – This is the most common type of model for commercial publishers to come together. In this model, commercial publishers will charge a subscription for a

certain amount of time and then release their materials into the public domain. Highwire Press at Stanford has tried this model. Though it remains to be seen how successful that particular project is, it has the potential to bring together the best of all possible worlds marrying the large content capabilities of commercial publishers with the academic and open access needs of universities.

How does the TCP model fit in with these models? It most closely resembles the *Multi-institution/Subscription* model, but it is much broader than that. TCP tries to negotiate between the philosophical differences between commercial publishers and the academic community in order to achieve its goal. In general, the TCP tries to find a middle ground between all of these approaches. One of the more intriguing aspects of the model is that it gets the money to create the product. Rather than grant funding which often is not sustainable over thousands of texts like this, the TCP has opted to let academic institutions, normally libraries but also departments and grant funds as well, contribute funds to the project which are then matched by the commercial publishers. That money is used to fund text production, and the more institutions that join, the more text can be created, thus making each text less expensive. Also, TCP is doing much more than creating another product. The University of Michigan is committed to university ownership, public domain access, and scholarly communication. Universities that are part of the TCP project actually own the texts and will eventually be able to distribute them beyond their own campus community. TCP is committed to working with scholars, librarians, publishers and members of the community to ensure that the needs of all three are met whether that means enhancing the interface, soliciting help for selection, or partnering with scholarly projects (Garrett 2002 117-119).

Sustainability

In essence, all of these models come down to one thing, sustainability. Institutional models often are highly useful to particular communities, but are not sustainable because they are very expensive compared to commercial products of similar scope, and such projects are usually dependent on the energy of a single faculty member. It would be impossible to build a universal digital library using multi-million dollar grants to create scholarly editions of every single book. Likewise, commercially produced databases are unsustainable because the subscription fees they would are unaffordable by the universities that require them. As a result, projects like Google Books are filling the void. Models like the TCP could certainly work with a Google library (Martin 2008). The key is making any model of electronic resource sustainable. In order for a project to be sustainable it has to have enough money; in order to have enough money there needs to be a large enough audience to support a project monetarily; in order for there to be a large enough audience, there has to be a broad enough range of material to support such an audience; in order to create so broad a range of material, there needs to be a standardized procedure for creating it; in order to create a standardized procedure, one has to sacrifice some of the editorial work available in many scholarly projects. TCP has been very successful in most of these factors, though it still struggles with how to get greater support among libraries and the scholarly community.

One of the main problems all digital libraries face is money. It is important that all projects realize how expensive a digital library is to create. The figures from TCP show that to complete 41,000 texts (approximately 20% of the entire collection in EEBO, Evans, and ECCO) will cost approximately $13,000,000. For TCP to create full texts for the entire collection of roughly 300,000 texts would cost over $100,000,000. These figures also do not count the ongoing costs of maintenance and preservation which will likely need to be borne by institutions in the future. Though TCP's costs are perhaps not applicable to all types of digital libraries, they indicate that the cost of creating an electronic collection is higher than any grant, single institution, or combination thereof is likely to be able to generate. Digital libraries of the future will need to generate large amounts of capital and will probably need to seek it among multiple institutions and from the commercial sector. TCP is just one way of doing this.

The EEBO collection is unique in that it contains nearly every book published between 1470 and 1700. Therefore it is of use not just to literary scholars but also to historians, legal scholars, and many others in all disciplines of the humanities. As a community we need to look for broad ranges of material that will be useful to large numbers of people. Google has already done this for a large number of books. Yet, it seems unlikely that they will be able to digitize the vast amount of rare book and manuscript material available. What are some efficient ways of digitizing this material collaboratively and in large enough quantities to create a sustainable audience? TCP may provide some answers here, for one of its successes has been the standardized workflows it uses. All of the staff at the Universities of Michigan, Oxford, and Toronto, the National Library of Wales, and other projects that cooperate with TCP, adhere to the same principles of text creation. There is a standard way TCP creates text, a common philosophy under which we operate, and a standard editorial policy used for all texts These are working documents not meant to produce a standard per se. Rather, they are constantly evolving ways of thinking as staff at the TCP find new problems.

The key to these constant struggles between collaboration and centralization, or standardization and meeting individual needs, has been TCP's desire to seek a middle ground between seemingly opposed and entrenched positions. Many scholarly projects seek to create a highly edited and tagged corpus of material for a specific group of scholarly users. Though these projects unquestionably offer the best for scholars in those disciplines, they cost a great deal of money and produce a very small number of titles. TCP on the other hand produces many more texts than smaller projects like have done. Though it is true that TCP texts are produced to a much lower standard of metadata and, therefore, are less useful to scholars than a highly tagged text would be, TCP does not aim to be a project useful only to particular groups within the humanities community. Rather the project seeks to provide a foundation for other groups to build upon. The foundation is the basic structural tagging TCP provides, the largely accurate transcriptions, and the standards-based text all done to a particular and overt philosophy.

Conclusion

In the digital world, there seems to be a divergence of opinion between the commercial and non-profit worlds about how to create content and how to create sustainable publishing. Large electronic publishing operations like Google are digitizing content on a massive scale with the hope of making money from advertising, selling chapters in print, or otherwise commercializing small pieces of content for niche markets. For a large corporation, the considerable investment needed for mass digitization would seem to be returned by the potential "long-tail" income from selling advertising and print on demand like services to a large number of niche markets. For non-profit organizations, however, and particularly universities, publishing models tend to focus on small niche markets and make investments in digitizing small amounts of material (manuscripts, collections of books); the cost of doing this far outweighs the potential income they may generate . Since all universities are grappling with economic downturn, shrinking budgets, and increasing costs, it no longer seems likely that they will be able to sustain investment on this scale. Nonetheless, universities, unlike large corporations, have an important mandate to disseminate information which by its very nature is of negligible market value, though it may be of extremely high intellectual value. How do we deal with these issues which seem to be pulling universities in particular into two directions?

The Text Creation Partnership is just one possible model that attempts to balance these competing forces. It has sought to maintain a middle way by which Open and Closed Access can work together and in which commercial and academic interests can be promoted side by side. If nothing else, it serves as an example of how the library, scholarly, and publishing community can come together in order to find common solutions. No scholarly project will ever match the size of a commercial database, and no commercial database will ever have the editorial apparatus of a scholarly project. Grant funded scholarly projects and other similar Open Access projects serve their purposes, and commercial databases serve

theirs. In the wake of increasing pressure from the commercial world, it is essential that the academic community comes together to create models that satisfy the needs of as many constituents as possible. In many ways, what we are discussing is how to create an entirely new infrastructure for scholarship in the electronic world. Though that is too broad a question for just one essay, it is hoped that by looking at one particular project, it will be possible to contribute in the creation of a solution.

Reference List

Jeffrey Garrett "The Early English Books Project meets at Northwestern" *College and Research Libraries News* 63, no. 2 (February 2002)

Kevin Guthrie, Rebecca Griffiths, Nancy Maron *Sustainability and Revenue Models for Online Academic Resources* Ithaka (May 2008)
http://www.jisc.ac.uk/media/documents/themes/eresources/sca_ithaka_sustainability_report-final.pdf

Shawn Martin, "To Google or not to Google, that is the question: Supplementing Google Books to make it more useful for scholarship." *Journal of Library Administration* 47.1/2 2008

Mark Sandler, "New Uses for the World's Oldest Books: Democratizing Access to Historic Corpora," *Association of Research Libraries* (*ARL) Bimonthly Report* 232 (Febuary 2004).
http://www.arl.org/resources/pubs/br/br232.shtml

Peter Suber, "Open Access Overview" http://www.earlham.edu/~peters/fos/overview.htm

John Unsworth et. al., "Our Cultural Commonwealth: The report of the American Council of Learned Societies'Commission on Cyberinfrastructure for Humanities and Social Sciences",
http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf