TBRC and its Model for Linking Text Images with a Bio-Bibliographical Finding Database

Fred Coulson
February 28, 2001

The Tibetan Buddhist Resource Center, or TBRC, is a not-for-profit venture dedicated to creating and maintaining a digital archive of Tibetan literature, sharing that archive with as large an audience as possible, and coordinating its efforts with other Tibetan data retrieval projects.

Introduction: History of TBRC

TBRC's origins go back at least 30 years, to the activities of the Library of Congress in India. There E. Gene Smith, in his capacity as Field Director for the Library of Congress, was able to secure for publication several thousand Tibetan texts that had been trickling into India along with the Tibetan refugee diaspora. This publication work was accomplished under the auspices of the U.S.–India tied aid program known as Public Law 480, or PL-480, whereby the proceeds from the sale of American grain to India was re-invested in Indian cultural projects. As one such investment, the Tibetan Text Publication Project of the PL-480, was able to recover over 5,000 volumes of Tibetan literature using a modern (for the day) offset process. These volumes — some of which were one-of-a-kind manuscripts, while others were limited-edition block prints with few extant copies — were provided to universities and in many cases formed the material basis for the academic study of Tibetan literature and culture in the West.

The PL-480 collection was the nucleus of what would eventually become the TBRC library. As the contents of Mr. Smith's library grew beyond the output of the PL-480 project, the need for a systematic catalog became pressing. Contemporary bibliographic metadata capture techniques were found to be inadequate for Tibetan works: a single Tibetan work is in many cases a compilation of many different works by different authors, and to merely catalog the top-level title and author is to miss the vast richness the work contains. Unfortunately, this is precisely how the widely-used MARC cataloging standard treats these compilations; this renders the standard card-catalog entries for many Tibetan texts all but useless.

In the Tibetan scholarly tradition, a text is cataloged and indexed using a human mind as the storage medium. The contents of the catalog are transmitted from one generation to the next by way of the student-teacher relationship. Because of the volatility of the human medium, entire lineages of commentary have been wiped out simply through the agency of war, pestilence and lack of interest. The Western scholarly tradition addressed this volatility centuries ago by developing methods of classifying and storing information about a literary tradition independent of the people who use it; such techniques, which led to the development of the modern library, essentially open up specialized fields to non-specialists and in general promote the dissemination of information. (One might argue that this is also a drawback — we have become a text-based culture, and it appears that we could destroy that culture merely by destroying our books). But these data-capture techniques are ill-designed to capture the subtlety of the Tibetan literary tradition, with its compilations containing complex internal structures and its heavy reliance on an oral tradition for explication of those structures.

Mr. Smith has spent a lifetime trying to remedy this deficiency in the Western metadata standards. In the beginning he cataloged the contents of each volume he laid hands on manually, using a typewriter and loose-leaf paper. Later, as personal computers became popular, he began cataloging his acquisitions using a flat-file database product called askSam©. However, this database format made it difficult to capture certain kinds of data, such as repeating title fields and biographical author data. As a result, much of this information wound up being stored in unstructured free-text memo fields. Thus digitalized, the askSam catalog was a great advance over the paper-based catalog, but it was still of limited use to anyone but the database's creator.

The Birth of TBRC

By the early 1990's, Mr. Smith's library had grown to over 10,000 volumes, and after having followed him around the world it was beginning to show some wear. As well, many of the works had been printed on poor quality paper, and some were beginning to show signs of the brittleness and decay that will overtake all books published in the era of acidic-paper publishing. Even the core PL-480 collection, which had been printed in several editions and existed in redundant locations around the world, was not being stored very well — many collections were missing volumes (or pages from volumes), and there were few complete collections still existing anywhere. And because of the inadequacies of the standard cataloging system mentioned above, it was (and is) in some cases impossible for libraries to determine which portions of the original collection they lacked. Thus a very significant portion of Tibetan literature seemed in danger of disappearing, not through ideological fervor, but through neglect.

With these two problems — the imminent decay of the books, and the possibility of never being able to find anything in them anyway — looming ever larger, the notion of digitalizing the library, and providing free (or next-to-free) access to it over the Internet, was born.

Access to the Data — Some Theoretical Concerns

Efforts to digitalize the Tibetan literary corpus are hardly new; quite a number of important projects have been working tirelessly for many years to input large collections as text. The fact that there is still no universal standard for Tibetan character encoding forces many of these projects to work in isolation: one project might use a plain-ASCII input scheme using a variant of the Wylie Tibetan romanization, while another might opt for input using native Tibetan script, using a proprietary encoding format. Although the work of these various projects is mutually exclusive at present, the time will come when there is a universal Tibetan character-encoding scheme, and converting the work that has already been done into whatever universal scheme arises should be a trivial programming task — in theory, at any rate.

The TBRC approach, however, is somewhat unique, to the best of our knowledge. Rather than inputting the literature as character data, we scan pages in as bitmaps. This technique is not being put forward as superior to the text-input projects that are operating, but rather is a necessary adjunct to them, one which will eventually help them in their work.

It may be stated as given that all digitalized textual material should be stored, ultimately, as text. Not only can such material be stored in a relatively compact medium, it also lends itself to searches and automated analysis in a way that is foreign to bitmap storage. However, bitmap storage does have some advantages:

- Time – In the time it takes a single operator to type in a 7-line folio page of Tibetan text, dozens of similar pages may be scanned as high-resolution bitmaps. This is quite an important point, given the volatility of the paper medium. Many works in the TBRC library may not last another decade, and the expense of prolonging their life by temperature and humidity controls is prohibitive.
- Editing – The nature of the traditional Tibetan printing process — the carving of a text backwards onto wooden blocks by sometimes poorly-educated monks — tends to introduce more ambiguities and inaccuracies into a text than many Westerners would believe acceptable. These textual inaccuracies and ambiguities are, often, resolved by means of an extensive tradition of oral commentary. For this reason, any textual input project requires editorial oversight by a learned authority. Ideally, this editorial process would go through several iterations before a definitive product were published. Even in the best case, however, there are two problems with this scenario: 1.) the time it takes to bring a text to market is increased in direct proportion to how much editing it requires, and 2.) any editing that takes place will necessarily come from an existing commentarial background, hence textual ambiguities — upon which divergent lineage interpretations might rest — will be driven out of the final product, at the expense of one or another interpretation. By scanning the text as a bitmap, we do not make any impact on the commentarial tradition, and leave all ambiguity intact for future scholarly analysis.
- Universal access – By scanning pages as bitmaps, we immediately make a text available to the client using platform-independent, easily-available image viewing software; there is no need for special fonts or other proprietary considerations.

Thus it should be apparent that the TBRC's plan does not compete with text-based input projects, and can be an important partner in their work. A concerted scanning effort could permanently preserve a significant portion of the Tibetan corpus in a few years; text input and translation projects could then use these high-quality scans for years to come, without having to worry about losing pieces of the deteriorating paper-based collection. In a sense, we are faced with a situation similar to the European Renaissance work of preserving the Western Classical corpus: parchment manuscripts that made the transition to moveable type have come down to us, while those that did not are lost, probably forever. We have an advantage that Gutenberg did not possess, though: we can take a snapshot of our vanishing corpus, and then parse it, analyze it, and publish it at our leisure, long after the original books have turned to dust.

Access to the Data — The Practical Application

In practical terms, the TBRC work is still very much in its infancy. We have scanned in fewer than 100 works to date; the scanning is done by a small but dedicated group of volunteers who can only work on a part-time basis. Setting up a large-scale scanning operation is a TBRC priority that is contingent on funding.

At present, the area that receives our most concerted effort is the development of the catalog for tracking the TBRC library. An important feature of the design of this catalog is its backward-compatibility with existing library metadata standards: we feel it to be of the utmost importance to be able to construct valid MARC-compatible records out of the contents of our catalog. Having said this, it must be stated that we feel in no way constrained by the inherent limitations of the MARC standard. Furthermore, we do not feel constrained to track only our books in this catalog. People, places, lineages, and family structures are all valid objects for our database to track.

This database, which we call **TBRCDat**, loosely follows an object-oriented model, although it is implemented as a relational database. Top-level objects include Works, People, Places, Lineages, Families. Each object, in addition to possessing certain attributes, can be an attribute for another object. So, for example, a Work object may possess one or more Person objects as attributes (as author, translator, etc.). A Person object may possess any number of Work objects (as written works), Place objects (as seats, residences, places of activity), and even other Person objects (teachers, students). A Family object — this includes aristocratic families, nomadic clans, and other corporate entities — can possess Person objects as members, and Place objects as estates or areas of transhumance. Objects may be combined to form new objects that may again be attributed to yet other objects — so, for example, a teacher-student relationship between two people becomes an object in its own right, and this new object can be applied as an attribute to a Lineage object. Thus we can trace any religious lineage we like from its founder right down to its present-day adherents.

TBRCDat can be viewed differently from different angles. Viewed from one angle, we are creating an On-Line Public Access Catalog (OPAC) for a body of work that is not adequately cataloged using standard library practices. Viewed from a slightly wider angle, this OPAC becomes a gateway into a vast labyrinth of biographic, geographic, and genealogical information.

1. **Works**

The core of TBRCDat is its ability to track works. What is so special about a Tibetan work that makes it difficult to catalog using "standard library practices?" There are a number of features of Tibetan books that make them difficult to classify using standard rubrics like Author, Title, and Subject Headings.

Many Tibetan books contain an inherent complex hierarchical structure. Western books are likewise hierarchical; the structure is reflected in the table of contents. The Tibetan topical outline or *sa bcad* is somewhat more complex than a table of contents. It is generally not explicitly recorded on its own page but is rather embedded in the text, to be parsed out by the alert reader. The various nodes in the structure are frequently

subject headings unto themselves. The Western approach is to enumerate all the subjects that the various parts of a book treats of, and attach them all as subject headings to the work itself. To do this in the case of a Tibetan book is to gloss over the fact that many works are in fact encyclopedic compilations of a teacher's life work. To ignore the structure and apply all subject words to the top level of the hierarchy is to lose an important data-finding tool. Learned Tibetans, who spend a lifetime memorizing these topical outlines, can quickly locate a target datum simply by referring to its position in a work's outline. We can do the same using a relational database or SGML-type data definition table; it all depends on the granularity with which we choose to analyze a work.

In TBRCDat, the topical outline is a database object existing apart from a Work; it is related to a Work in a many-to-many relationship although, practically speaking, a single outline will generally only be related to a single Work. Each node in a topical outline's hierarchy is related to a page (or range of pages) in a one-to-many relationship; the pages themselves are joined to a particular Work in a many-to-one relationship. (Each page, as it is scanned, becomes an attribute of a particular Work; however, the page itself is an object in its own right, and can receive attributes like subject headings. This gives us the ability to search down to the page level, although we lack word-level searchability that true text input would give us).

Thus there are numerous ways to drill down into the database to get at the text we want. By applying standard author, title, or subject searches, a user can find the top-level Works, which might be collections or compilations. Further, if in our data input we properly qualify the structural nodes, a user can employ the same search techniques to find groups of relevant pages. And finally, by qualifying the pages themselves, the user can retrieve individual pages corresponding to the terms of the search. (This last is actually a massive undertaking, and is a fairly low priority for us.)

It should be noted that the topical outline-tracking feature would probably best be implemented using an SGML or XML data-capture mechanism, rather than the relational database front-end that we currently use. It was a fairly late addition to our design, however, and changing the implementation would involve a not-inconsiderable retooling of our data input workflow, which is not practical right now.

Then, there is the problem of how to track "manifestations" of works. Many works of Tibetan literature are editions or translations of no-longer extant Indian originals; how do we indicate this kind of relationship? The Western library solution is to qualify the published work with the title and author of the non-manifested work as subject headings. While this may seem adequate, it seems to us to be a shortcoming: a "title" search for a work like the *Bodhicaryavatara* could conceivably turn up no matches, despite the fact that the database might contain numerous manifestations of that famous work (with slight title variations, depending on the edition). Forcing the user to do a "subject" search for what is essentially a "title" seems contrived, which is why we permit non-manifested works to exist as works in the database; they can be linked to other Works using an external table that defines the two Works being linked and the relationship type ("edition of", "translation of", etc.). Such works are simply not qualified with publication information, and it is this information that marks a work as "manifested". Placing this publication information in its own database table means that non-manifested works are not saddled with a lot of empty fields, and this helps reduce the size of the database.

2. **People**

The example just given also relates directly to a problem we had to solve on the biographical side of our database: How do we track the works written by a historical figure (like Shantideva), but which do not exist in the "real" world? We could just type in a list of works for each entry in our Persons table, but this would reek of redundancy. Far better is to link the Persons table entries with the Works table in a zero-to-many relationship. By allowing non-manifested Works to exist as Works in our database we can list all a person's writings even when such writings do not exist as standard bibliographic entities.

When we started tracking extended biographical information in our database, we discovered that there was much more involved than tracking birth and death dates and a few family relationships. For one thing, we decided that we wanted to keep track of any number of significant dates in a person's life, and this meant creating a repeatable generic "events" object, with an open-ended list of events (birth, death, marriage, ordination, etc.). The fact that this object is repeatable means that variant dates of birth in different sources can be included in the database, and the citations can be noted (cited works being, themselves, objects in the Works table).

3. **Places**

Our Places object presents the most complex difficulties. Here we track monastic seats, family estates, places of birth, death, and other life events, regions of nomadic movement, and so on. Eventually, it is our goal to link up the historical data in TBRCDat with the various GIS (Geoscience Information System) projects that are developing for the Himalayan region. But how this linkage will be implemented presents a special problem. A GIS system is typically a two-dimensional snapshot of a geographical area at a given time. To see the area at another time, another snapshot must be taken. Historical data can be difficult to extract, because the categories (such as geopolitical boundaries) that exist in one snapshot might not exist in a previous one. Historical data, then, must be tied to something in the GIS data that persists universally through all time periods. To tie a monastery just to a given county is inadequate, because the political boundaries have probably changed markedly since the monastery's founding. It might seem ideal to be able to identify a toponym with its latitude-longitude coordinates, but unfortunately this information is largely missing from the historical literature. Added to this problem is the migration of monasteries and towns from one geographical location to another over time. Even the names themselves can change, making it hard to identify monasteries that have moved, been destroyed, and then rebuilt; they would therefore appear to us to be distinct places, but in the popular imagination they are the same.

One way out of this difficulty is to treat a Place (whether a point, such as a monastery or town, or a polygon, such as a township or county) as a persistent entity independent of

time and (paradoxically) space. Time and space are regarded as attributes of that entity, and as repeatable fields they can change. We might not know the exact point coordinates for the place in question, but we probably know the larger political structure it was part of at a given time, and the polygon coordinates of these larger structures can be ascertained. By further qualifying these spatial coordinates with a timestamp, we can with some accuracy create a historical dimension for the base GIS dataset.

Dates, obviously, require special handling. In the case of a published work, there is no problem: just transcribe the date from the publishing data (although, especially in books published in the East, there is frequently more than one date given). When dealing with other objects, however, the matter is more complicated. Since we are dealing with historical information that is frequently of dubious accuracy or certainty, a date becomes not a single attribute but a cluster of at least three attributes: start date and end date, known date, and degree of accuracy. The "start date/end date" pair and the "known date" are mutually exclusive: for example, we might have it on authority that Person X was abbot of Monastery Q between 1256 and 1263. On the other hand, we might only know that he was abbot in 1261, and we have no further information. So being able to deal with both of these possibilities is important, and it is crucial to be able to cite the source of this information with a hard bibliographic reference. Finally, we must be able to add a free-text annotation, to substantiate the "degree of accuracy" field, which will of necessity be subjective.

Linkage to the Outside

Developing the data capture mechanism for TBRCDat and exploring ways to offer our texts and data over the Web has been a consuming task. However, there is yet another issue to address, that of data linkage to the outside. This is where the true promise the Internet lies.

The familiar prototype for the Internet-as-information-provider is the library. Any good academic library is a vast store of information which, depending on your field of interest, can give you access to hundreds, even thousands of years of cumulative data. Within a relatively small space a researcher can browse much of the accumulated wisdom of human civilization, and the scope of inquiry can be broadened indefinitely if one is prepared to wait a few days for the wheels of the Inter-Library Loan process to turn. The traditional library complex is essentially a great database spread across the entire planet.

Two factors limit the usefulness of this sprawling database.

1. The *index* or *catalog*, which offers a one-dimensional view of library holdings, sorted according to standard conventions like author and title. This top-down view is somewhat expanded into two dimensions by assigning standard subject headings to objects in the library. True three-dimensional access — by which I mean the ability to interrogate the content of items in the library, without actually picking them up and looking at them — is only allowed for textual items that have been digitalized in a way that makes them searchable.
2. The other limiting factor is *time*. Wandering through the stacks just takes too long.

These limiting factors are not insurmountable, needless to say. A library is like a vein of ore, and the various cataloging mechanisms are tools of the miner's trade. Learning how to mine the resources pertinent to one's field is part of the training of any researcher, whether in academia or in business.

For several decades now, computers in general have been set dangling before the researcher as a way to bring down some of the traditional barriers to research. Even a modest digital database, stored on a mainframe computer and accessed via dedicated public terminals, offers tremendous economies of scale over traditional card- or book-based cataloging mechanisms. Personal computers, again, effected an important change in personal research habits, as they streamlined tasks which used to require armies of graduate students sweating over stacks of shoe boxes filled with recipe cards. But the Internet offers us something of a silver bullet: a way to access data from a seemingly endless range of topics directly, from any angle, using a simple, cheap, and standard interface that is available everywhere.

And yet, alas, there is no silver bullet. For what we are beginning to see is an Internet that becomes a vast, entropic sea of data, some of which is not worth the cost of hard disk space it takes to store it. The indexing of this sea is left up to a collection of public search engines, many of which (by selling high search rankings to the highest bidder) are in clear conflict of interest against their stated mission of providing access to information. Again, we are being offered one-dimensional access to a sprawling database: in the old library, we could search by metadata (subject headings and broad categories) but not by content; on a typical Internet search engine, we can only search by content, not by metadata. (The HTML specification does, of course, provide a facility for specifying metadata in addition to the content of a web page, but this feature only works for static web pages, and anyway very few webmasters use it properly.)

Standing out from this gray sea are numerous islands of well-formed data: information providers who offer specialized gateways into their own private data domain. There are different levels of access, different types of output, and different prices. There are also vastly different standards. Many of these high-calibre information providers still use standards that are based on binary protocols that were developed while the Internet was still in its infancy, before the now-familiar Hypertext Transfer Protocol had been developed, when available bandwidth was small and transmission times were potentially quite large. Such protocols do not operate very well with the Internet as we know it today. Others present data in more friendly text-based protocols, of which there are competing standards. XML has been touted as one of the more promising of them, but there is no guarantee that it will survive.

So what we have, in effect, is a reconstruction of the old library, where old barriers to research have been replaced with new ones. Data providers are islands on the Internet, and each island has its own language. Again, the researcher must commit significant time to learning the tools of the trade. Now, however, these tools come in the form of expensive software that must be used to communicate with

the island databases. Often you can search the islands using a public HTML interface, but the output from such searches generally does little more than fulfill idle curiosity. It is as if each floor of the old library now speaks a different language, and you must hire an interpreter or pay a fee each time you want to look at something.

However, things are not really as grim as all that. In fact, what we have is a golden opportunity to break down the two big barriers to research mentioned above. By narrowing our focus to the field that interests many of us here — that is, Tibetan studies — we can come up with a model that will exploit the Internet to full advantage, one that will be extensible, portable, and modular; one that can easily be adapted for use in other fields.

Interoperable, but Independent

But what is wrong with having islands of data on the Internet? After all, research has always been conducted this way, has it not? By analogy, a book, an index, or an encyclopedia can be regarded as a "data island": it is the result of an individual's or group's long work, offered to the world as a structured interface to a largely unstructured mass of information. People often do their best work when they are allowed complete control over what they do. While it is human nature to want to help others, it is equally natural to want to help *on one's own terms*. Why, then, would we want to remove boundaries that actually seem to heighten progress, and force people to conform to standards that may constrain them? These Internet data providers, by retaining complete proprietary control over their systems, can no doubt serve up a product that is far superior to any that could come about through consensus and compromise. If a choice must be made between interoperability and the robustness of a research tool, then it seems logical that interoperability must yield.

Indeed, we are at a point where we can have the best of both worlds: independence, as well as interoperability. By wrapping existing database servers in a commonly agreed-upon protocol, we can quite easily place a large, multi-disciplinary database at the fingertips of a researcher. The database itself is distributed across the Internet; its various pieces are tied together with the common protocol that creates the illusion of a single, seamless database.

The field of Tibetan studies is vast, and no one data provider could hope to bring together all its disparate threads into a coherent tool. Each provider — be it a university or an independent data provider like TBRC — has its proper expertise. TBRC's expertise happens to be digital books, as well as a wealth of historical, biographical, and geographical information. It is beyond our scope to start tracking things like paintings or sculptures in our database; nor do we have the capability to set up a comprehensive GIS project. Nonetheless, we want to be able to link our database objects to such information. How can we achieve this, without duplicating work that is being done with more competence by others? This question is the driving force behind the creation of a modular, distributed meta-database.

By *distributed*, we propose a scheme that leverages the unique talents of individual data providers. By *modular*, we mean a portable model that permits a data provider to join in with a minimum of software retooling. We wish to imply here that a data provider may retain complete control over its individual product. However, each provider must be willing to compromise in the area of the common data model that it exposes to the other parts of the larger database.

There are two main design issues to address: the data model, and the protocol.

The data model is simply an extension of what we spoke of earlier: what kinds of things are we going to deal with? How shall we qualify them? In terms of the TBRC data model, we can get as complicated as we like with our Persons module, qualifying it with all sorts of tangential family and lineage information. However, another data provider might not care for this kind of granularity. Thus we might choose not to expose this part of our data model to the larger, shared model. Alternatively, we might agree to different levels of exposure: objects could be arranged into "themes", exposing more or less complex data structures depending on a run-time parameter passed by the caller. Such a data model with its themes would be a subject of discussion for a standards body, and the activities of such a body should in no way constrict the design of individual data products.

The protocol would be the programmer's implementation of this model. It could permit dynamic queries of one database by another in response to a Web-based user request; on the other hand, we might choose static linkage of the various databases, whereby each data provider would be a mirror of the entire meta-database, downloaded across the Internet at periodic intervals.

This model is not without its difficulties. One serious problem is authority. How do we know that one provider has not duplicated the work of another provider? What mechanism will we have to resolve such authority conflicts? What about data entry — do we allow real-time data input? If so, how do we resolve record locking conflicts over the Internet? These are serious issues, not to be taken lightly. We feel, however, that the most expeditious course of action is not to resolve the difficulties before the design can commence, but rather to probe for collaborative interest and create a small-scale working model to demonstrate its workability.

Summary

With its expanding library of digital works, the TBRC is well poised to make a significant contribution to the field of Tibetan studies. Our database, too, shows promise as a potential tool for serious research. A vital part of our stated mission, though, is to coordinate our efforts with other data providers, and we feel that this should involve more than just putting hypertext links on our website. The interoperability model here stated is a fairly large undertaking, and will require grassroots participation from all sides, rather than top-down coordination. It is hoped that this model will receive the interest and support it will take to bring it to life.