The Rise of "Thematic Research Collections" in the Study, Teaching and Transmission of Buddhist Scriptures

David Germano and Nathaniel Garson
University of Virginia, USA

Part I: Introduction

Digest

This paper will deal with the rise of "thematic research collections" in the study and teaching of Buddhist scriptures, a type of collection facilitated by emergent Web technologies. The specifics will be drawn from our creation at the Institute of Advanced Technology in the Humanities of a prototyped SGML/XML-based standard for cataloging and reproducing Tibetan Buddhist literature on the Web, which uses one of the most important Tibetan canons, *The Collected Tantras of the Ancients* (*rNying ma rgyud 'bum*) as its test bed and first exemplar in *The Samantabhadra Collection* (see http://jefferson.village.virginia.edu/tibet/).[1] We will discuss the literary canon in question as background, outline the nature of a thematic research collection and its special connection to collaborative work, and conclude with a detailed outline of the DTD we have created for the project

An integrated Digital Library approach to research and teaching of Buddhist Scriptures with digital technology

Before proceeding, we would like to briefly allude to some of the larger intellectual, political and technological issues pertaining to the creation of "of a digital library system and associated "information community" for a geo-culturally defined area, and how these impact on digital initiatives pertaining to the reproduction, translation, and analysis of Buddhist literature stemming from that defined area. *The Samantabhadra Collection* belongs to *The Tibetan and Himalayan Digital Library* (www.lib.virginia.edu/infocom/tibet/), which is project sponsored by the University of Virginia to create a multi-institutional digital library devoted to all aspects of research on the geo-cultural area of Tibet and associated information community. As such, it embraces such initiatives as a GIS model of the Tibetan plateau, use of immersive multimedia resources to visualize Tibetan places and the cultural events transpiring with them, language instructional resources, encyclopedic reference resources on terms, peoples, and places, a folk

---

[1]. Our two main collaborators at the Institute are Daniel Pitti and Worthy Martin, to whom we are deeply indebted both in the creation of the project, as well as many of the ideas expressed in this paper.

music collection, and bibliographic resources. The study of Buddhist literature and Buddhist culture, should be done increasingly in an fashion which takes full advantage of these integrated digital library resources. We believe that these technologies hold the promise of eventually fundamentally transforming how we go about research, not only in terms of process, but also in terms of content.

Thus, in the present context, we would like to briefly reference a few of the broader implications for a generalized model based on specific issues we are addressing in the context of Tibetan Buddhism. Since our main focus in this paper is a more "traditional" (for the digital world, at least) approach focused on the texts themselves, we must limit ourselves to a brief sketch of five distinct areas of a digital library relevant to research and study of Buddhist literature, broadly defined.

(i) The new conception of "thematic research collections" based on XML-SGML involves a transformation of the generation, publication and use of Buddhist literature and contemporary scholarship on that literature. This is the main topic of our paper, namely an explanation of how this transformation enables enhanced direct and intellectual access to Buddhist literature.

(ii) In addition to written literature, digital technology also allows new levels of access to oral literature, both in its own right, and as spoken commentaries on written texts. In *The Tibetan and Himalayan Digital Library*, we are creating a XML-Java system for transcribing, time coding, translating, and analyzing digital video of Tibetan Buddhist oral commentaries on texts, as well as broader Buddhist cultural events. Users can thus go from the written texts in the thematic research collections, to see how that same text is commented upon in traditional oral exegesis. Not only is the audio available, but users can also see the cultural context visually in which the commentary is/was delivered.

(iii) A digital library approach also can facilitate a broader notion of what a text is beyond simply a source of intellectual discourse. In other words, it allows us to explore all aspects of a Buddhist culture's understanding of a text – the physical appearance of texts, specific contexts and processes for how texts are studied and taught, ritual uses of texts, transmission practices, dissemination of literacy, and so forth. These can all be documented with images, video, text and so forth so that users can go from texts, to associated materials documenting on how those texts were generated, used and transmitted in their own cultural contexts. This includes immersive environments in which three dimensional interactive models of monasteries allow users to witness ritual exchanges involving texts right within the architectural and communal setting in which they took place. We have thus begin prototyping such an environment within one monastery within the old city portion of Lhasa.

(iv) These contexts of texts can be dealt with to an even greater extent through development of multimedia reference works built into the digital library's infrastructure, so that users can go seamlessly from a text, to broader reference materials on people, places and terms referred to, and/or relevant to the text in question. These allow users to follow strands of significance that exceed the text proper, yet which ultimately form the contextual horizons of meaning that constitute the text's significance in its generation, transmission and use These reference works can take advantage of Web-technologies to collaboratively build resources which are unprecedented in their ability to simultaneously publish the most granular of materials – a stray paragraph on a term – and the most extensive of materials – a five page essay, say on the history of a single term. In addition, reference works are not limited to text, but can incorporate images, videos, and so forth. We have prototyped, for example, a dictionary system in Cold Fusion which allows for OED-style rich dictionary documentation of Tibetan terms, along with multimedia illustrations and encyclopedic-style essays (which appropriate). We plan over the next two years to migrate a finalized design over to a Java and XML-based system. These reference materials with their integrated media demand to be collaborative in nature, which involves complex political and informational flow issues in their generation and maintenance. One exciting development is the ability to incorporate GIS into these reference works, namely spatial and temporal databases that enable the result of queries to be output in digital cartographic representations. GIS approaches allow for a more regional-based approach to the study of textual traditions in terms of analyzing distribution of textual genres, and relating contents and styles of texts to associated cultural and environmental factors on locality-based criteria. We have thus created a base GIS model of the Tibetan plateau drawing upon contemporary data, and are now working towards expanding this back in time.

(v) Finally, a unifying thread in all of the above is the necessity to pay close attention to the politics of community formation, and how digital technology can encourage unprecedented "information communities" to take shape. A single digital library can use technology to address and integrate users with different linguistic needs, different cultural conceptions, different disciplinary backgrounds, and different levels of expertise (scholarly/popular) that all share a common interest in the Buddhist texts of a given culture.

Within *The Tibetan and Himalayan Digital Library* these different components are at varying stages of development, and hence the connections of *The Samantabhadra Collection* – historically our first major project – to the broader digital library are still nascent. However, while *The Samantabhadra Collection* primarily involves classical, written texts, it is in the process of being integrated into the larger scope of the project in a number of ways which we hope to discuss in more detail in a later publication.

Part II: The Samantabhadra Collection - A thematic research collection of Tibetan Buddhist scriptures

Overview

As the issues concerning a Tibetan digital library will be discussed in relation to a particular religious canon, it is necessary to first understand the nature of this group of scriptures, how it arose and how it is structured. Then, we will present in more details the actual SGML/XML[2] system for creating a thematic research collection of Tibetan Buddhist scriptures using *The Collected Tantras of the Ancients* (*rNying ma rgyud 'bum*) as our first test bed. Finally, we will discuss what we view as the transformations such a collection engenders in how we research, publish, and teach Buddhist scriptures. Part II of the paper will thus consist of six subsections: (i) An overview of the literary basis of the collection, namely a canon of 8th to 15th century Tibetan canonical literature known as *The Collected Tantras of the Ancients*; (ii) the nature of a SGML/XML-based thematic research collection and its collaborative dimensions; (iii) the deep cataloging dimensions of our system; (iv) the manner in which the system provides direct access to texts; (v) the manner in which the system provides expanded intellectual access to texts; and (vi) the broader issues concerning the networked management of a humanities projects consisting of multiple institutions and scholars from diverse cultures and locations.

I. *The Collected Tantras of the Ancients*: the literary basis of the collection

The Collected Tantras of the Ancients (*rNying ma rgyud 'bum*)[3] is the most important canonical Tibetan literature after the well-known Tibetan compilations of translated Indic Buddhist literature known as Kangyur (*bKa' 'gyur*) and Tengyur (*bsTan 'gyur*). It currently exists in at least six different editions, each with overlapping but divergent textual contents, with the largest (the *mtshams brag* edition) consisting of forty-six volumes and over forty thousand folio sides.

Though Buddhism had its inception in India, it was translated *into* the Tibetan language in a very self-conscious process from quite distinct languages. Thus, translation was a high profile phenomena in Tibetan Buddhism from its beginning. Buddhism had two major inceptions in Tibet, and both were driven by a massive influx of foreign texts translated into Tibetan: the first in the eighth and ninth centuries came to be called the "earlier transmission" and the second in

---

[2]. We are currently using SGML, but are in the process of migrating it over to XML. From hereon out, we will simply refer to SGML, but it should be understood that the broader project is based on XML.
[3]. Please see the web site for bibliographical details.

the tenth through fourteenth centuries came to be called the "later transmission". All three canons derive from translations supposedly done during these two time periods.

To outline the distinctive nature of *The Collected Tantras of the Ancients*, we will discuss three topics: doxographical categories, the nature of a canon of visionary literature, and finally issues of authorship and transmission.

i. Doxographical categories of tantric traditions

In general, both Indian and Tibetan Buddhism tended to classify Buddha-authored and human-authored texts as belonging to particular doctrinal-praxis systems. These systems were then ranked in a vertical nested hierarchy of "lower" and "higher" systems of thought and practice. These constitute "doxographies", i.e. "writing of the view". *The Collected Tantras of the Ancient*'s texts also are organized into distinct doxographical groups within the canon. The overarching organization is tripartite, and is based upon three distinct tantric systems: Atiyoga, Anuyoga and Mahaayoga. However, the internal organizations of each of these three then is based on further doxographical subdivisions. These classifications are crucial for understanding the creation and existence of traditions, since often these *canonical schemes themselves were important factors in the retroactive creation of traditions for these texts.*

ii. A visionary canon

The Collected Tantras of the Ancients in particular constitutes a hybrid canon of genuine translations, and visionary translations, otherwise known in some modern circles as apocrypha. Such a canon raises interesting questions as to the nature of translation, authorship, readership, and editing. It may be useful to speak of *visionary* translation, *visionary* authorship, *visionary* readership, and *visionary* editing within the broader context of a *visionary* canon marked by visions, reincarnations of past saints, emanations/incarnations of past, present and future Buddhas, and constant processes of concealment and revelation. In this way we can bibliographically and interpretatively acknowledge the tradition's own self-understanding and self-representation, but also make useful distinctions with more conventional notions of translation, authorship, readership and editing that were as familiar to Tibetans as they were to Westerners.

We have chosen this collection for electronic cataloging and digitization not only because of its great importance for understanding the history of Tibetan literature and Buddhism, but also because it offers a complex test case for addressing issues pertaining to canonical and apocryphal literature, as well as text-critical scholarship overall.

iii. Problems of authorship and transmission

In terms of authorship, we can discriminate between three principles of difference that were important to later editors of these canons of translations: the being, temporality and ethnicity of any given author.

1. The being of authors: Buddhas and humans, reincarnations and emanations

The "being" of an author signifies the strong difference between a text authored by a divine Buddha, and a text authored by a "historical" human. A Buddha's texts are always presented as transcripts of orally delivered lectures framed by a narrative structure. However, human authors are usually actual people located in our geographical and temporal contexts who originally *write* texts. Thus, Buddhists texts are typically divided up on the basis of the *being* of the authors: divine oral texts, and human graphic texts. Divine oral texts often are also partially, or fully, spoken by similar divine figures such as Bodhisattvas, Dakinis, Gods, Goddesses, and so forth. These texts are often produced in public circulation by human individuals but ascribed to a Buddha's divine voice.

In addition reincarnational and emanational beliefs can also defer compositional agency of a given text produced by a historical figure to a past or divine figure with whom they are identified as a reincarnation or emanation. In other words, I may produce this text, but claim it to be the composition by a much earlier figure of whom I am the reincarnation, or by a divine figure of whom I am a historical incarnation. Modern notions of "authorship" are thus not adequate with their assumptions of a single author producing a stably defined text with his/her name stamped on it. Instead we must identify an entire set of figures including "divine author", "divine redactor", "divine audience", "human redactor", reincarnational pedigree, emanational pedigree, and so forth, each of whom are said to, or may actually have had, a hand in the actual shape of the received text.

2. Temporality: issues of revelation

The "temporality" of an author relates to the history of how Buddhism was transmitted into Tibet. Authors and translations are divided into those belonging to the "earlier transmission" - the ancients - and the "later transmission" - the new ones. The later transmission produced new translations in conjunction with Indian scholars, while the later proponents of the early transmission instead after the Empire's disintegration in the ninth century, initiated a revelatory cult which produced hitherto unknown texts that claimed to be early translations concealed and then later re-revealed by visionary means. *The Collected Tantras of the Ancients* is thus literally

the "collected tantras" of the latter group, who styled themselves as the "ancients" (*rnying ma*). They are clearly a mixture of genuine early transmission translations, and much later "revealed" works.

Regardless of whether the text in question was Buddha-authored or human-authored, ancient authors classify texts as "continuously transmitted precepts" (*bka' ma*, "*Kama*") or "treasures" (*gter ma*, "*Terma*") The basic principle of distinction concerns the way in which the texts in question claim to have been transmitted in Tibet following their supposed translation into Tibetan from their original source language. "Treasures" claim to have been concealed for a period of time following translation, such that their circulation was at some point interrupted via concealment and then later re-revealed; "continuously transmitted precepts" claim to be texts transmitted without such intervening concealment and re-revelation. This cult of concealment and revelation entails a broad range of figures involved in the final form of the text - author, concealer, revealer and so on - and complex tensions between traditional and modern analysis of the actual identity of the author.

3. Ethnicity: Indians and Tibetans

The "ethnicity" of authors concerns whether the author is Tibetan, or whether the text in question is a Tibetan translation of a text originally authored by an Indian. With the rise of the later dissemination, the consequent emphasis on translation of Indian texts and lineages, and the creation of the Kangyur/Tengyur to canonize them as authoritative - a canon not simply of translations, but which included *translation* as a central aspect of its sanctity - the non-Tibetan character of texts' authorship came to be viewed as a hallmark of authenticity. This appears to have been an important factor in the rise of the revealed "treasure" cult and the Ancients' tendency to produce its new literature in the authorial voices of eighth and ninth century Indians, rather than in their own Tibetan voices.

II. A collaborative, networked Thematic Research Collection

The Samantabhadra Collection is what is known as a "thematic research collection" in that it is an intentional collection of research material centered on a particular, and in this case rather broad, theme. Its design from the beginning was intrinsically collaborative in character because of our conviction of the centrality of collaborative work in Buddhist Studies. The present section will outline the nature of a thematic research collection, and its special relationship to collaboration in this context.

i. Collaboration in the humanities: interdisciplinary realties and possibilities

Traditionally, collaboration in humanities scholarship of any type is rather rare. Humanities scholars are accustomed to working alone and the entire system - at least in the US - is set up to encourage this: the system of academic credit for publications, institutional and disciplinary boundaries, and so forth. Collaborative activity tends to be limited to a few broadly based ventures, such as dictionaries and encyclopedias, or biographical projects, or the rather tenuous collaboration of conference proceedings. Until now, cataloging and translation projects of Buddhist scripture have aimed at print publication, and hence had a finite period of composition terminating in the end product. This has tended to entail a series of quite bounded contributions with limited opportunities for synergistic dialogue evolving over an extended time frame.

In particular, one could point to several types of interdisciplinary collaboration which are conspicuous for their obvious benefits and equally obvious rarity:

1. collaboration between philologists and interpretative scholars
2. collaboration between textual specialists and art historians
3. collaboration between area-studies experts and scholars without such expertise

These types of collaboration – which are just a few examples of collaboration across discipline - would further the advancement of Tibetan Studies by providing a broader and integrated picture of the Tibetan society and culture from multiple perspectives.

ii. Why collaborative in this case?

Why does our project in particular demand a collaborative approach, aside from the general benefits of collaboration? There are three major reasons for this: the amount of material involved, the state of Tibetan Studies as a field, and the inadequacy of present reference material in Tibetan Studies.

1. The Size of *The Collected Tantras of the Ancients*

As with the Kangyur and Tengyur, the massive size of *The Collected Tantras of the Ancients* exceeds the capacity of a single scholar to deal with it comprehensively in any given aspect – such as, critical editions, translations, interpretations - much less as a whole. Therefore, the required work from critical editions to interpretative analysis can only be done gradually over time through the contributions of many scholars. Critical editions and translations are particularly obvious cases, because they require such a great investment of time to be done accurately. By providing a dynamic digital publishing environment for such works, the Collection can support the work of isolation scholars, provide integrated access and profile for

scattered research, and thus begin to compile the database necessary for understanding a canon of this breadth.

2. The Infancy of Tibetan Studies

Tibetan Studies began to receive common recognition as an important field only in the last fifty years. Though a great deal has been accomplished in that time, it is still in a relative state of infancy. One example of this is that very little work has been done on *The Collected Tantras of the Ancients*, despite its historical, philological, and doctrinal importance. This is gradually being addressed, but work in this vein is hindered by the paucity of resources as well as the dearth of interested scholars. It is therefore desirable to bring together the limited number of scholars that study this topic to exchange information and ideas. The problem is that these few scholars are often separated by considerable distances and may not even know of each other's work. This problem is further exacerbated by the fact that such scholars are often working in different fields that do not have the forum to adequately communicate with each other. The *Samantabhadra Collection* will provide the means for scholars from any field to share their work on this canon as well as a venue for them to engage in dialogue on any aspect of the topic and its relation to Tibetan culture and society as a whole. By proactively eliciting, enabling and stimulating scholarship in this particular area, we also hope the initiative will produce new content, not simply integrate content that would have been produced anyways.

3. Inadequacy of reference materials

Another outcome of the inchoate nature of Tibetan Studies is that there are few adequate reference materials available. In a number of areas, such as biographical information in translation, resources are often altogether lacking. While it is of course possible to compose static printed volumes containing this kind of information, the interactive, digital model provides an avenue for quickly developing such resources through the collaborative efforts of a number of scholars. The authority files - which will be discussed below - are designed to support these efforts and will serve as reference materials for the *Samantabhadra Collection*. Originally, these were considered to be resources for the textual collections alone, but ultimately we decided that they should be designed in a way to serve as overarching and collaborative reference resources that can be shared by multiple projects throughout the digital library.

iii. A new model of collaboration: technology and the humanities and the rise of Thematic Research Collections

A Thematic Research Collection: soliciting, enabling and publishing research

Digital technology has enabled the creation of what we term "thematic research collection", but what do we mean by this term, and how does it relate to collaborative ventures? An archive is traditionally understood as a secondary by-product accumulating around the life of a person or organization.[4] It thus coheres by provenance and not by intention, in contrast to an ordinary "artificial collection" which is the result of intentional acquisition of materials for the purpose of building a collection. In contrast to such a traditional archive, a thematic research collection signifies a relatively new model that has been proliferating under the stimulation of the incorporation of technology into humanities scholarship.

A thematic research collection begins with the selection of a scholarly theme, which could be an author, a genre, a movement, a city, a historical period, or a canonical body of literature, such as in our own case. It is thus intentional in that one formulates a criteria and then builds a collection based on that criteria. Following the selection of the theme, all primary, secondary and reference resources relevant to the study of that theme must be identified; in other words, everything a scholar might want at an arm's reach in an ideal universe in working on these bodies of literature. The collection then accumulates the relevant resources in primary, secondary and reference literature based upon those criteria, and actively integrates them in a single, interlinked medium, unlike a traditional print collection.

Function as publisher

An additional point of difference is that in addition to simultaneously providing direct access *and* intellectual access to these resources, a thematic research collection also functions as a publisher. In this aspect, it facilitates and organizes focused research on a particular theme or area, such that it integrates the activities and interests of archivists, librarians, publishers, and scholars. The nature of electronic on-line publishing entails that the products of its publishing activities are intimately interlinked, exegetical literature to the primary literature it comments on, reference works to the exegetical literature, and so on. In addition, the publishing can be open-ended, with products amenable to constant update, thus allowing for dialogues and conversations to emerge precisely due to the extended temporality of published texts, or texts in

---

[4]. We are using this terminology to distinguish it from what archivists and manuscript librarians usually mean when they use the terms "collection" and "archives". In the traditional sense, the difference between an "archive" and a "collection" lies in intentionality, which in the United States is often referred to as the difference between an archival collection and an artificial collection respectively. The former is a byproduct of materials accumulating around the life of an individual or corporation in the process of their ordinary activities, while the latter is an intentionally created collection, such as a library collection, which was deliberately generated to belong in that collection.

the process of publication. Another element of publishing in a thematic research collection is the possibility of proactively soliciting and organizing interlinked scholarship in ways enabled uniquely by this extended temporality. What we have in mind is projects that constitute collaborative reference works without traditional limitations on length, which are directly linked into the primary resources and scholarship rather than standing alone, and which can be constantly updated with reference to the original and revised entries. This might take the form of an encyclopedia of terminology, a biographical database, and so forth.

This unitary and interlinked character of the publishing enables a thematic research collection to integrate the traditionally distinct work of philologists and interpretative scholars, text scholars and art historians, and in general scholars from different disciplines. It also attempts to bridge the traditional distances between modern Euro-American scholarship and traditional Asian scholarship. Most interestingly, the integration of these different scholars into a single interlinked site, as well as the ability to respond to other works through revision of one's original work, stimulates unprecedented discussions across disciplinary and other boundaries. Such conversations are often largely implicitly present in the academy, but the revisable and interlinked nature of a thematic research collection renders such discussions explicit. In addition, the ability to publish short contributions, as well as eliminate temporal lag between composition and distribution, further encourages the development of such dialogues across boundaries. The ultimate result is that a thematic research collection can become the foundation for the emergence of a new community of actively interacting scholars bound together by a common subject area across disciplinary and regional boundaries. The collection becomes a site where primary resources are consulted, reference materials are accessed, scholars publish articles, scholars read new cutting edge scholarship on their subject, and explicit real time exchange take place.

What are the general advantages to a humanities project posed by collaboration, and how does technology enable a new model of collaboration in the humanities? Technology has offered new possibilities of collaboration that did not exist previously - rather than taking extant models and digitizing them, the new medium is enabling us to come up with entirely new models. Before discussing the nature of the "thematic research collection" enabled by new technology, we would like to briefly point out some of the obvious elements of technology which enable new collaborative possibilities.

イ　1. Open-ended publishing

The digital nature of the medium enables an open-ended publishing process capable of continual revision and expansion. Projects no longer have to have finite time lines that terminate in an irrevocably fixed form. This open-ended quality of digital publication enables the gradual evolution of dialogue and synergy across all types of scholarly boundaries, thereby enabling collaboration as well as stimulating it.

ロ　2. Granularity of publishing

Electronic publication enables small, granular contributions such as chapter summaries, or even mere footnotes, to be formally published as part of a larger venture. This is a function of the integration of publication, collection maintenance, and scholarship. This not only widens the participatory nature of a project, but it also helps dissolve the strict boundaries of publications and authorship.

ハ　3. Immediacy of publishing

The Internet enables the incorporation of on-line communication channels into scholarly projects, as well as the collapsing of time lags between composition and circulation. It also expands access in terms of financial and geographical barriers.

ニ　4. Publishing links between diverse types of texts, images and sounds

Hypertext enables direct connections between diverse textual work as well as images: (i) text critical work, (ii) translations, (iii) reference materials, (iv) interpretative scholarship, and (v) visual images and sounds.

ホ　5. Searching links across domains in publishing

Sophisticated searching operations are available which operate simultaneously across diverse textual domains, thereby establishing direct links between them within the medium of publishing itself.

^ 6. Integration of scholarship, publication and collection maintenance

In short, a thematic research collection both demands and enables collaboration, as outlined above.

III. *The Samantabhadra Collection*: a case study of deep-level cataloging in thematic research collection

In order to better understand the nature of a thematic research collection, we will briefly outline the scope and architecture of *The Samantabhadra Collection*. The collection is composed of, or participates in, several subcomponents. While at present only some of these are operable, all should be completely functional by the spring of 2002:

1. In-depth Catalogs of each edition of *The Collected Tantras of the Ancients*
2. Images of the primary texts
3. Typed in editions of the texts that can be view either in transliteration or Tibetan script
4. Translations of the texts into European languages
5. Scholarship on the texts from chapter summaries to full-blown textual analyses
6. Authority records which function as reference databases

As will become clear, the collection thus integrates the reproduction of texts in their original form, reformatted transcriptions, translations, deep cataloging for intellectual access, proactively organized scholarship for expanded intellectual access and contextualization, and the solicitation and archiving of a wide range of ever expanding scholarship. As a whole, it thus has the capacity to create and maintain an integrated community of librarians, archivists, publishers and scholars that is a distinctive product of the technology which articulates its infrastructure. Building the foundation for this infrastructure is our present task, and thus we are focused on composing in-depth catalogs of the canon encoded in SGML. As these catalogs are the entry-point for accessing other parts of the collection, they will be discussed in more detail. We will thus begin by examining deep level cataloging, then discuss direct and intellectual access, and conclude by briefly examining how such a complex project can be run through networked management.

Cataloging at deep level and customized forms

The most basic strata of the collection is a series of catalogs with unusually deep descriptive levels that have been custom designed to deal with the descriptive challenges posed by this class of non-Western and non-modern literature. The descriptive challenge posed by these medieval

Tibetan canons have entailed the need to modify traditional Western cataloging practices, as well as textual encoding initiatives such as TEI (which we why we use a *Tibbibl* instead of *Bibl*). In doing so, we have tried to integrate traditional Tibetan authorial and textual categories with modern literary and bibliographical conventions.

As an example, we will now briefly discuss four such descriptive challenges and how the SGML data-structure deals with them: Identification, Titles, Origination, and Text-structures.

i. Identification

The basic challenge is that print catalogs often assign single identification number that fail to link to other identification numbers for the text, and in particular with identification numbers applying to other editions of the same canon. In addition, they usually neglect to assign a text a number which indicates its location within a broader doxographical scheme, i.e. its intellectual classification. Our own scheme locates a given text within a particular association within a particular volume that varies in each edition, but each text is also classified within an intellectual classificatory scheme. We have thus addressed this issue by assigning dual identification numbers, one locating it within its unique publishing context, and the other within its intellectual context. In addition, the hyperlinked environment allows all identification numbers - those of the various editions and other external identification numbers - to be directly linked with each other. Finally, we have also identified volume numbers and text sequence numbers within the volumes, so that this identification number precisely locates a given text within its physical volume in a given edition.

A related issue is that of identifying the boundaries of an individual text. This is a rather complex issue because of the presence of pervasive intertextuality, i.e. the direct incorporation of texts within other texts. In addition, there is the phenomenon of texts which grow over time by the accretion of new sections. We have used a traditional text terminating particle as our principle of identifying text boundaries – along with other traditional considerations outlined in the we site - and decided to err on the side of granularity (see below).

The types of identification numbers used in our system are as follows:

Edition ID Number

Editions are identified by a unique sigla. The sigla has two letters, derived from the editions name, with only the initial letter capitalized. Multiple versions of the same edition are distinguished by numbers appended to the sigla. For the master electronic catalog the sigla is Ng.

Volume ID Number/Letter

Volumes can be identified by either their sequential number or alternatively the Tibetan syllable traditionally assigned to the volume. The former is prefixed with a 'v' to indicate it is a volume number, but both are preceded by an edition sigla. Thus, Tb.v23 refers to the 23rd volume of the mTshams Brag edition.

Text ID Number—Sequential

There is a unique reference number for each text contained in a particular edition of a collection. It is comprised of a sigla and the text's sequential number within the edition as a whole, separated by a period. Examples are: Tb.391 and Tk1.15 (i.e., the 391st text in the mTshams Brag edition and entry number 15 in Kaneko's catalog of the gTing sKyes edition—Tk1.)

Text ID Number—Doxographical Class

The doxographical class number represents the intellectual location of a work, either according to a particular edition's classification or as classified in the Master Catalog. These are represented by a sigla, followed by a series of numbers identifying the doxographical class and sub-classes to which the text belongs, separated by periods. Thus, Ng1.3.3.2 is in the master catalog (Ng), Atiyoga section (1), Experiental Precepts Series or *man ngag sde* (3), Seminal Heart or *snying thig* (3), and is the 2nd text in that sub-sub-section, namely *The Tantra of Unimpeded Sound* (*sgra thal 'gyur*).

Text ID Number—Volume/Text

There is also included an identification number describing the precise location of a text in the physical artifact, namely the volume/text identification number. This describes a text's location in an edition by indicating which volume it is in and which text in the volume it is. Thus, Dg.v12.3 is the third text in the twelfth volume of the sDe dGe edition.

Passage ID Number

To identify a passage, texts are divided into front, body, and back sections, represented by a, b, and c, respectively. These identifiers are followed by a number indicating the chapter-level element, and concluded with the line number. Hence, Tk.28.b3.154-160 refers to the passage that spans from the 154th to the 160th line in chapter 3 of the 28th text in the gTing sKyes edition to the 160th line. Context may often obviate the use of the text ID number so that in the above example, "b3.154-160" would suffice. Another format is to use the volume number followed by a folio-number.line-number prefixed by an 'f'. E.g., Tk.v01.f324.6-325.1. In cataloging texts

in a particular, physical edition, however, pagination is recorded by folio.line-number alone.

Other ID Numbers

An unlimited number of other identification numbers can also be added to a catalog record for cross-referencing purposes and to account for prior schemes. Those that have been included to date are:

The Library of Congress number (both the new and old versions)
OCLC number
ISBN number
Library call-number

ii. Multiple titles

The basic challenge is that Tibetan Buddhist literature is characterized by multiple titles from internal and external sources. Traditional print catalogs have been very limited in this context, both in their failure to document the full range of titles, and in their failure to allow for searches of anything but the full version of a single title. Clearly an electronic catalog allows for searches on keywords or phrases, but beyond that we have developed our DTD to allow the full recording of all possible titles.

The types of titles recorded are as follows:

Normalized title
Titles from the text itself: front, body, back
Title(s) from Margins
Titles from Oral Traditions:
Titles from Secondary Literature:
Non-Tibetan Titles:

iii. Provenance

The basic challenge lies in ascertaining the origination of visionary and apocryphal works, and identifying the agents involved in that process. Translation and transmission are as important as authorship. We are thus documenting not just the "author", but every single agent involved in this compositional process utilizing traditional and modern nomenclature.

The categories under “provenance” are as follows:

1. Authorship
    Author(s)/Speaker(s)
    Authorship type and affiliation
    Place/date of composition/delivery:
    Audience
    Requester(s) of composition
    Redactor(s) of original composition

2. Translation
    Language(s) from which translated
    Translator(s)
    Place(s)/date(s) of translation
    Redactor(s) of translation
    Place(s)/date(s) of redaction of translation

3. Transmission
    1. Transmission status
    2. Concealment: concealer(s), place, date
    3. Revelation: revealer(s), place, date
    4. Lineage history
    5. Editors

iv. Text Structures

The basic challenge is that Tibetan texts are not as clearly bounded as the western-style book. They are often contained within a multi-volume edition, in which the boundaries between texts can be obscure. Furthermore, as with any culture's literature, the texts can take on a variety of internal structures and formats. We are thus dealing not only with the problems of identifying a text and its boundaries, but of accurately describing its internal components. The general principle we have adopted for delineating the end of a text is the presence of the terminating phrase, "*rdzogs so//*". Of course, there are exceptions to this rule and other factors, such as punctuation, spacing, and the traditional view of the work, also come into play. In terms of describing the text's internal structure, we have adopted the standard bibliographic concept of the text as having a front, body, and back. Within these broad categories, however, we have had to create a typology for "chapter-level elements", which are the composite parts of the front, body, and back. This typology is modeled on the division of the body of a text into chapters and describes corresponding divisions for the front and back sections. The front consists of title pages, title lines, homages, statements of intents, introductory scenes, and so forth; the back

consists of closing sections, author's colophon, translator's colophon, lineage transmissions, concluding prayers and so forth. We have built these into the catalog and text representation to allow for coordination and navigation. These types were developed with regard to *The Collected Tantras of the Ancients*, and hence may require modification for other genres of Tibetan literature.

The chapter-level elements are as follows:

A. Front

1. Title page
2. Title Line
3. Homage/Invocation/Praise
4. Statement of Intent
5. Untitled Introduction
6. Ordinary Introductory Scene (*thun mong gleng gzhi*)
7. Extraordinary Introductory Scene (*thun mong ma yin pa'i gleng gzhi*)
8. Outline (*sa bcad*)

B. Body

1. Section Division
2. Chapters
    a. Chapter title
    b. Chapter homage
    c. Chapter colophon
3. Interstitial Chapters

C. Back

1. Closing section
2. Author's section
3. Redactor's colophon
4. Translator's colophon
5. Lineage transmission
6. Reviser's colophon
7. Editorial colophon
8. Scribal colophon
9. Printing colophon
10. Concluding prayer (*mjug byang smon lam*)

11. Closing invocation
12. Instructional colophon
13. Undetermined colophon

In addition, Tibetan texts are characterized by relatively limited punctuation and formatting in a rectangular page containing five to eight lines, such that the visual flow of the text yields little information about the semantic contents or structural divisions of the text. We have thus chosen to rely on Western formatting conventions, such as indenting citation, introducing line breaks for lines of verse, bold face for headers, and so forth, for the visual reproduction of electronic versions of the texts.

In addition, Tibetan texts are often characterized by a quite complex nested hierarchy of internal outlines (*sa bcad*). In cataloging large collections of scriptures, such as *The Collected Tantras of the Ancients*, the problem of *sa bcad* did not arise as texts are consistently divided into chapters. However, other situations, such as the Bönpo Collection, require that the outline be encoded within the bibliographic records, because their *sa bcad* contains a vast amount of information. These Tibetan outlines, however, are quite complex, regularly descending to the tenth level and beyond. This presents a problem to the standard text encoding format in TEI whose DIV elements traditionally only go down to the seventh level. See directly below for how we have dealt with this issue.

v. How SGML data-structures deal with these issues

SGML and its extensible correlate, XML, are well-suited to deal with these textual anomalies precisely because they are not just languages, but meta-languages. Unlike HTML, which is a fixed-format markup language for display and linking purposes only, SGML and XML are markup languages used to *define* other markup languages. Thus, they allow the user to construct a hierarchical and descriptive syntax or language for describing any situation. Such a language can then be used to search, display, or interconnect documents according to the designer's wishes. SGML/XML allow users to define three distinct types of elements or tags for marking up text

1. Structural elements: these elements describe the logical structure of a document—front, back, heads, chapters, paragraphs, etc. The ability to nest such elements allows one to create a hierarchical structure that directly correlates to the structure of the document.

2. Nominal/Thematic elements: these allow one to tag phrases and words according to their thematic content so that, for example, all personal names can be marked as such for searching and collating purposes.

3. Referential elements: these elements create referential links either within a document or to files and URLs outside the document.

An instance of an SGML/XML language is created by writing a Document Type Definition (DTD) for it. The DTD defines the names of the tags and the rules for using them, including where they can be placed and what attributes they have. (HTML itself is an instance of SGML and has its own relatively static DTD.) With the help of the Institute for Advanced Technologies in the Humanities (IATH) at the University of Virginia, we have developed our own DTD to address the specific problem of cataloging Tibetan texts. This DTD, known as the TIBBIBL, has as its basis a DTD developed by the Text Encoding Initiative (TEI) for the markup of electronic texts. The basic structural components of TEI were retained, ranging from the broader FRONT, BODY, BACK divisions of a text to the more specific paragraph element P. TEI also has a number of dramatic elements, which have been retained for future use. While extensive, TEI is nonetheless inadequate to cover all the oddities of Tibetan texts mentioned above. To address these specific issues, we enhanced a new TEI-based DTD known as TIBBIBL with a variety of specially-designed elements to deal with these problems. These elements allowed us to address the problem areas outlined earlier in a satisfactory way.

We will now briefly examine how the TIBBIBL handles those five problem situations (dividing the problem of text structure into two components): multiple identification numbers, multiple titles for a single work, the variety of persons involved in the provenance of a text, the variety of chapter-level elements, and the problem of Tibetan outlines.

1) To handle multiple identification numbers, a TIBID element is defined in such a way that it can be nested and appropriately labeled. The TIBID element has two important attributes that define it, TYPE and SYSTEM. Type determines whether it is an edition, volume, or text ID. System delimits how the ID is recorded as a letter, number, or sigla (for editions). Since they can be nested, the text-number TIBID can go within the edition-sigla TIBID to form a full ID reference for that text, i.e., sigla.text-number. This has many diverse applications.

2) A TITLEDIV element was added to handle the extended variety of Tibetan text-titles. These are labeled with TYPE and SUBTYPE attributes to distinguish the titles. Type is a fixed classification: front, body, back, non-Tibet, secondary, margin, oral. Subtype is open and further specifies the origin of the title by naming a chapter-level element, as listed above. A number of

other elements can be nested within the TITLEDIV, primarily of course TITLEs. However, to deal with text-titles at the end of chapters a special TITLELIST element was devised, which allows one to list a version of the text-title and all the chapters that use that particular title and the pagination of its instances. Thus, for a text with 100 chapter where 50 use the same text-title, that title does not need to be repeated 50 times, but is entered once followed by a list of the chapters and paginations.

3) The problem of documenting every person involved in the provenance of a text was handled with a RESPDECL or responsibility declaration. This was adopted from TEI but augmented. It is also delineated by a fixed list of types: audience, author, concealer, editor, patron, redactor, requestor, revealer, scribe, translator. PERSNAME and SOURCE elements are included within the RESPDECL to record the person's name and the source of the information. One can also include dates, place names, multiple people with different roles, original language, and discussions within a RESPDECL, greatly enhancing its functionality.

4) To deal with the particular aspects of Tibetan textual structure, we enabled the TIBBIBL (Tibetan Bibliography element) to be nested within itself and endowed it with a LEVEL attribute. This allows for the creation of a nested hierarchy to model a Tibetan collection. With a collection such as *The Collected Tantras of the Ancients*, there are four basic levels to the hierarchy: the edition, the volume, the text, and the chapter. These are the fixed list for the level attribute. While the description of each level has its own particular bibliographic needs, they are common enough to be described through a single DTD model. With the ability to nest TIBBIBLs within TIBBIBLs and distinguish them by their level, we can create an electronic model of an edition using a single type of bibliographic record. An edition record contains different field from a text record but they are both Tibetan Bibliographic records and can both be described through the TIBBIBL DTD. Using the nesting functionality of the TIBBIBL together with the TEI DIV structure, we can accurately represent an edition containing volumes, containing texts, and containing chapters.

5) The problem of a Tibetan outline breaks the TEI division model due to the complexity of the *sa bcad*. TEI could be expanded by adding further DIV elements (DIV8, DIV9, and so forth). However, these would always have a limit, and the Tibetan text whose table of contents receded to the twentieth layer, for example, could break that model. Instead, we devised an open-ended recursive method for dealing with such tables of contents or outlines. Within the TIBBIBL file, the Tibetan Analysis (TIBANAL) element is used to group the TIBBIBLs of the chapter-level elements into front, body, and back sections. Using the same method, we can also describe a text's outline. Each level of an outline that contains sub-levels consists of a TIBBIBL

describing that level (title, pagination and so forth) and then a TIBANAL containing TIBBIBLs for each of its subsections. These subsections can then repeat the procedure for sub-subsections *ad infinitum* or *ad nauseum*.

Six types of SGML files

The Samantabhadra Collection is composed of or makes use of six different type of files. These are:

i. Catalogue files
ii. Doxographical files
iii. Bibliographical entries
iv. Text representation files
v. Scholarship files
vi. Reference databases

Catalogue files compose the structural outline of a collection. These files use the TEI DIV elements to encode the hierarchical structure of the collection in which the individual bibliographic records are inserted to create a complete catalog. The catalog files are the final products published on the Internet. There are two types: an edition catalog file and the volume catalog files. An edition catalog file consists of the bibliographic record for the edition as a whole followed by a series of three first level division elements or DIV1s. These represent the Atiyoga, Anuyoga, and Mahaayoga sections respectively. These three DIV1s in turn contain second-level division elements, DIV2s, for each of the volumes it contains. The DIV2s include the volume's bibliographic record plus DIV3s for each of the texts within the volume.

Doxographical files are similar to catalog files in that they record an organizational structure of an edition. However, instead of being organized along physical lines, as the catalog files are broken into volumes, the doxographical files are organized into intellectual categories: the doxographical genres of the collection. The first level of an edition catalog is in fact a doxographical analysis, because most editions group the Atiyoga, Anuyoga, and Mahaayoga texts into discrete volumes. The doxographical files go further into these primary doxographical categories to map out the intellectual structure within which the texts can be grouped. Thus, using TEI's DIV elements a doxographical file can recreate, for instance Atiyoga's subdivisions into *sems sde*, *klong sde*, *man ngag sde*, *phyi ti*, and *yang ti*; and the internal divisions within those, such as the *man ngag sde* subcategories of *phyi nang skor*, *gsang skor*, and *snying thig*. The individual text bibliographic records are then inserted within the DIV element that represents their level of classification.

The bibliographic files are the files that contain the bibliographic information and are contained completely within the TIBBIBL element. They can be included at any level, being distinguished according to their level attribute. There are edition, volume, doxographical, and text bibliographic records, which are designed according to the specific needs of that level. Information is always included at the highest possible level. Thus, for an edition the publication information is included within the edition bibliographic file and not within the volume or text files, because this information is the same for all volumes and texts within that edition. Text bibliographic files include a nested TIBBIBL for each of the chapter-level elements it contains.

Text representation files are those files which contain a digital representation of a text. These can come in a number of different formats and therefore file-types. At the primary level, there are scanned images of the individual folio sides. There are also files containing the extended Wylie transliteration of the text, files containing a Tibetan script version of the typed in text, chapter summaries, and lastly files for translations of the text. With the exception of the scanned images, all these files can be marked up structurally as well as thematically, i.e., not only are the internal sections delimited, but titles, names, symbols and so forth can also be individually tagged.

Scholarship files contain all the scholarship on a text submitted to or collected by the project. The scholarship on each text will be included within an individual file, linked to that text by the file name. These files will be structured according to theme with DIV1 elements used to group scholarship along certain basic themes, such as but not limited to, general, linguistic, historical, and so forth. Subdivisions can also be used as the need arises.

Finally, the reference database are resources used by the Samantabhadra Collection, but are now being developed within the broader Tibetan and Himalayan Digital Library project as a whole. Through thematic tagging within a catalog, identifying a person, place, and so forth, links can be created to search the corresponding databases for information on that topic. Six such databases are envisioned:

1. Biographical database: containing information on both mythic and historical figures,
2. Institution database: containing information of various institutions from the monastic to the secular,
3. Movements database: containing information on historical movements within Indo-Tibetan culture,
4. Geo-referencing database: containing geographical information of the Indo-Tibetan cultural region using GIS technology and information.

5. Terminological dictionary: containing definitions and linguistic information on Tibetan terms, and
6. Image database: containing images of places, people, artwork, etc to be shared with the other databases.

By situating these databases in the larger rubric of the Tibetan and Himalayan Digital Library, they can be referenced from any of the catalogs and text files within the Textual Collections.

IV. Direct Access to texts: Cataloging as front end for reproduction of texts

As unusually detailed as the cataloging is, this intellectual access to the texts in fact operates as the front end of an collection whose full scope is far more extensive. The catalog record of any text will provide access to actual representations of that text in any of four different forms: digital images, transliteration, translation, or summary. The access will be linked to the title of the text or chapter so that by clicking on the title a new window pops up displaying the text in any one of these four modes and allowing the user to toggle between them. The Institute is also working on developing technology whereby multiple versions of a text can be viewed simultaneously in a linked manner so that all the versions will scroll in unison.

i. Direct access via digital images of texts

The ultimate goal of the collection is to scan each folio side of each addition for the sake of comparison and development of critical editions. Each folios image will be saved as an separate file. The optimal specifications of these image files are still being developed in dialogue with other similar projects. Most likely, the TIFF or JPEG format will be used with a reasonable resolution to provide both clarity and relative speed for downloads.

ii. Direct access via typed in editions of the texts

Because of the vast amount of time required to enter transliterations of the texts, we are not systematically entering transliterations for each text. Rather, we are prototyping the technology by entering a few key examples so as to provide the means for scholars to submit transliterations in the future. These represent the second and third forms of the texts that can be accessed through the catalog, the transliteration and the Tibetan script representation. The transliteration will be done through our Extended Wylie transliteration scheme and will also serve as the basis for creating a digital Tibetan font representation through converters written in JAVA or PERL. Both the transliteration and the Tibetan script versions will be encoded in SGML, using structural and thematic tagging. The method for marking up these texts will be derived from TEI’s rich set of elements for marking up prose and poetry. Original page and line

numbers will be inserted and will act as links connecting to the scanned images of the folio. Eventually, electronic critical editions of texts will also be included in the master archives, as they become available.

iii. Direct access via translations into European languages

The fourth form is the translation of the texts in European languages, especially English. These will be encoded in much the same way as the Tibetan versions of the text, structural elements delimiting the text's inner structure and thematic elements marking people, places, keywords, and the like. The presentation of both forms will be in the standard web-page format with headers, spacing, and indentation as in a western-style book. Texts will not be presented in the traditional Tibetan *pecha* format except for the digital images of the folios.

V. Intellectual access to texts: stimulating and publishing research scholarship in SGML-encoded collections

Beyond cataloging and direct access, the collection also includes a third component, namely the solicitation and SGML-encoded archiving of scholarship of all types on the texts and traditions covered by *The Collected Tantras of the Ancients*. The aim is to help integrate, and even create an international team of scholars. This begins with chapter by chapter summarization of the texts, and continues to include scholarship from any discipline dealing with some aspect of *The Collected Tantras of the Ancients*. To facilitate academic credit for major contributions, we are launching an associated electronic journal. The scope allows the project to record smaller contributions - such as summaries of chapters or even simple footnotes - that would otherwise languish for lack of a publishing forum.

i. Intellectual access via related scholarly research

Intellectual access to the work in the collection will be provided from a number of different avenues. There are three broad categories of scholarship that will be included:

1. Textual studies of all types: these studies can vary in length from the size of a footnote to that of a book. The ability to incorporate discussion fields directly into a bibliographic record of a text is available throughout the SGML record as is the ability to add footnotes. Both appear in separate pop-up windows, with the only difference being that a footnote is accessed through a footnote icon, while discussions are accessed through hot-linked text. A submission process is being developed whereby scholars can submit discussions (or footnotes) on a certain text through the Web. These submissions will be reviewed by the editorial staff and appropriately

placed. Longer article or book-length submissions will be stored in separate SGML files accessible through links in the bibliographic record.

2. The electronic journal: we are launching a broader electronic journal in the Digital Library project, which will provide a forum for peer-reviewed essays on topics in any way related to T*he Collected Tantras of the Ancients*.

3. Chapter summaries: Each text will have a separate, associated file containing a summary of its contents chapter by chapter. This will be accessible from the bibliographic record of the text or any of the digital representations of the text. As an invaluable aid to further research as well as an essential tool for users who do not know Tibetan, this facet of scholarly contributions will receive the most attention initially, and in most cases will long precede any translation of any given text.

ii. Intellectual access via integrated multimedia reference materials

In addition to these diverse scholarly contributions, we are developing authority files—as described above—operating as front-ends for deliberately integrated scholarship on particular themes. These will provide intellectual access and context for the texts, authors and traditions included in the collection. These will deliberately solicit and organize scholarship on specific areas within their scope which will then be integrated into a single on-line resource. This has the great advantage of avoiding scholarly repetition, concentrating diverse studies in a given area into one location, and encouraging scholars to contribute in small ways that would otherwise never reach the light of day in the publishing arena.

iii. Intellectual access: the benefits

As clear from the above, such a *thematic research collection* constitutes a major undertaking that in its very architecture and scope demands a collaborative approach integrating the work of multiple scholars from multiple disciplines. At the same time, it enables a synergy between scholars by putting their individual work within a greater context. In this way it begins to have a generative effect on eliciting and enabling new scholarship, while that new scholarship in turn contributes to the eliciting and enabling of further scholarship. We thus hope that the overall effect will be to generate a concrete sense of belonging to a scholarly community that currently exists only implicitly. In summary, we can point to four specific benefits:

1. Scholarly credit for text-critical work, translation and small level work

The collection enables, rewards, and gives high profile to the production of critical editions and translations of texts, scholarly activities that have increasingly been neglected in academics. It also brings the benefits of this text-critical scholarship to a generalist audience.

2. Interaction between different disciplines

The collection encourages the integration of philological and interpretative scholarship, two domains which often are at odds with each other. One can thus move seamlessly between critical editions, translations, and interpretative studies of a particular canonical passage. The collection will also stimulate conversation between disparate disciplines—history and philosophy, textual studies and art history, history and anthropology, etc. It will do this first at the implicit level through the on-line integration of studies from different disciplines, but ultimately at the explicit level as scholars from different disciplines are drawn to the same on-line site. In particular, the thematic research collection will build bridges between social and historical scholars, historical and doctrinal scholars, textual and art image scholars, and so forth.

3. Integration of materials across domains

Finally, the collection weaves together disparate items within a single domain, as well as what are usually quite distinct domains. For instance, usually a given canon with multiple editions is cataloged in a series of separate print editions, none of which are interfaced with each other. As for the latter, the collection integrates cataloging, direct access to the texts, translation, primary scholarship, and associated scholarship into a single collection.

VI. Collaboration in action: the networked management of a humanities project

What are the specific tactics we have employed for collaboration of a project done at multiple centers? The advantages are clear, but how does one set up the concrete technological and procedural infrastructure to enable such collaboration to take place smoothly? The Institute is using the project to work out complex but workable procedures for the networked management of a humanities project based in different locations across the world. We are articulating the infrastructure for communication and exchange between different groups and different locations, so as to bring together publishers, archivists, scholars, librarians, editors in a collaborative environment. We have begun by focusing on authors and editors in the context of building this collection. There are three primary phases in implementing and coordinating collaborative activity at multiple centers.

i. Humanities and technology: IATH and UVA's Tibetan Studies program

To begin with, the initial collaboration is one of simply that of building bridges between technology experts and humanist scholars. In this regards, we have been very fortunate to be based at the University of Virginia, which is home to one of the world's largest Tibetan book and manuscript collections, one of largest graduate programs in Tibetan Studies scholarship, and the Institute for Advanced Technology in the Humanities (IATH), itself dedicating to the building of these bridges. We built these bridges initially through weekly meetings between ourselves, two other advanced doctoral students in Tibetan Studies, and three IATH staff members, as well as an email list and ftp site. Since IATH itself already integrates technology and humanities scholarship, these bridges were far easier than otherwise possible.

ii. On-line handbooks

We have developed an extensive and comprehensive on-line handbook for reference *and* instruction governing all aspects of the project, both thematic and technical. This includes guides to the use of SGML editors, data entry procedures, transliteration systems, work flow, editing processes, and so forth. These digital handbooks are being used for training new staff members, and for reference by continuing staff. They thus function to better facilitate the integration of the project's various operations in North America, Europe and Asia. As new technological elements are developed at separate centers, the handbook will be expanded to keep each center abreast of new developments. We are also using mailing lists and ftp sites in order to pursue further developments in a collaborative fashion.

iii. Multiple centers: Astoria and Adept Edit for file-management in SGML

The second phase of collaboration involves developing multiple centers that also combine technology and humanities scholarship. In 2000-1, we have expanded to a new site focused on Bönpo canonical literature at Rice University, which is directed by Dr. Anne Klein, a Tibetan Studies professor, and Dr. Gregory Hillis, who is managing the site half time. This new textual collection is being used to explore procedures for collaboration between multiple, disparate centers. This makes a superb sister collection to the Samantabhadra Collection because of the close connection between the school of the Ancients (*rnying ma*) and the Bön religion, which claims to be the indigenous religion of Tibet. The Bönpo Textual Collection has presented new challenges to cataloging Tibetan scriptural works, because of the elaborate outlines of the texts that are being entered, a collection known as the *Zhang zhung rnyan rgyud*. Through the combined experience of editors at both sites and the technical experts at IATH, the method for encoding textual outlines discussed above has been developed. Another encouraging feature of expanding to multiple sites is that we now have the capability for remote sites to log-on and store files in the powerful file-management program utilized by IATH at the University of

Virginia. As the price of this software, Astoria, matches its quality, it is a distinct advantage of great economic benefit for other established locations to be able to use it remotely.

As the textual collections are encoded completely in SGML and contain a large number of documents, it is imperative that coherent methods for file-development and management be implemented to maintain the consistency and integrity of the collections. This is even more essential given that there are several editors working on different facets of the project. Fortunately, there is software available to handle such situations. The Textual Collections project makes use of two powerful software suites that are integrated to facilitate the management and version control of SGML records. Adept Editor and its most recent incarnation Epic Editor serve as our SGML editors for adding to and correcting the bibliographic records, while a program known as Astoria takes care of the file-management, version control, and publishing.

Epic Editor and its previous release Adept Editor are powerful SGML editors developed by ArborText that greatly assist the cataloger in his or her task. Each is installed in our Local Area Network (LAN) and is accessible from a number of stations. The basic advantage to such editors is that they display the SGML documents in an easily readable format. Like their HTML counterparts, SGML documents are strictly text documents. The elements or tags that mark-up the data are only interpretable in conjunction with the language's DTD. (An HTML parser containing its DTD is built into most browsers.) In a simple text editor with no tag recognition, editing such documents is difficult because it is hard to distinguish between tags and data. If the tags are changed, the document could very well become invalid. Epic Editor circumvents such problems by displaying the tags graphically as labeled arrows on either side of the data. A special window allows the user to edit the tags attributes as needed, but the tag cannot be directly changed. One of the advanced features of Epic and Adept is that they also parse the SGML against its DTD. Thus, they will not allow elements to be inserted where to do so would be invalid; they do not allow illegal values for attributes; and they provide the editor with a list of valid tags for the insertion point at the cursor. Another advanced feature is that the two editors can be bridged to the file-management program Astoria, providing an easily accessible avenue for retrieving and saving the SGML documents.

The file-management program, Astoria, is key to our project. An edition can contain upwards of one-thousand distinct SGML files, and for the *Collected Tantras of the Ancients* there are six major editions. Anyone familiar with Tibetan or Buddhist studies knows of the extent that textual collections in such fields can take on. Therefore, it is essential to have a means to manage and document such a voluminous number of files. A product of Chrystal Software,

Astoria functions as both an SGML/XML library and a publishing house. As mentioned above, it is accessed through a bridge between it and the SGML editor. However, access is password restricted for obvious safety reasons. On the analogy of filing cabinets, the documents for each project are stored within Astoria inside “cabinets”. Each cabinet can contain multiple files or folders, and folders can contain either files or other folders, allowing the user to build an organizational hierarchy. Thus, an edition would be a folder containing its catalog record, its bibliographic record, and then a folder for each of its volumes; the volume folders contain the volume’s catalog and bibliographic record and then folders for each of its texts. The text folders could contain a bibliographic record, text representations, summaries, and any scholarship specifically devoted to that text.

Document access within Astoria is based on the model of a library. Once a document is imported to Astoria, it can be checked out by only one authorized user at a time. That user is then the only person who can edit the document; others may view it but not change it until the document is checked back in. The bridge between Astoria and the SGML editor allow users to easily access and edit the documents directly from the editor, which also parses the document against its DTD to ensure validity before it is checked into Astoria.

Another extremely powerful feature of Astoria is the ability to share discrete SGML nodes between documents. Any node on an SGML document’s hierarchy tree (or in layman’s terms, any coherent chunk of SGML) can be shared or imported into another document, provided that doing so does not invalidate that receiving document. The benefit of such a procedure is that there is only a single version of the shared SGML. If the source document’s data is changed or the SGML altered, then those changes are reflected in every place it is shared. In our project this is utilized to create the catalog records. A volume’s catalog record does not contain a manually copied version of each text’s bibliographic record; instead, it contains a shared version, which is actually just a pointer to the one and only existing record. When the volume’s catalog file is published, Astoria retrieves *the latest version* of each text’s bibliographic record and inserts it in the appropriate place. Thus, when updating one of our catalogs, changes only need to be made to the components bibliographic record in order to be reflected in the published product. In Astoria, publishing a version means to create a static, unalterable version of the SGML. This involves collating and inserting shared SGML in the texts hierarchy, checking it for completeness, parsing it against the DTD, and finally writing a new complete file. As with any file, this file can be exported out of Astoria and thus displayed in other software, but unlike the other files, it cannot be edited. It is a fixed final form, like a print edition. However, the original source SGML files are still accessible. These can be changed and a new edition published, whenever the need arises. One can therefore see that along with a consistent file-naming

convention, Astoria, provides a powerful to for the management and publication of a large body of SGML files, such as are found in the Textual Collections Archive of the Tibetan and Himalayan Digital Library.

These two programs, Epic Editor and Astoria, along with some in-house conversion software, provide our catalogers with the means to create, edit, manage, and publish our SGML documents. A four-step process is involved in the creation of the catalogs:

1. Data entry: the initial bibliographic record for a text is entered into a simple word-processing table, which contains fields for all the variety of information concerning the text—titles, chapters, provenance personae, paginations, and so forth. This form is then processed through a Visual Basic macro that converts it into an SGML document. Having the editor's enter the information through such a process by-passes the tedious procedure of creating each SGML document from scratch. It also ensures the conformity of the resultant SGML documents.

2. Uploading the SGML: Notes and other peculiarities unique to that document are then added by hand through the SGML editor (Epic), which also parses the document to make sure it is a valid SGML document that conforms to the rules of the DTD. Because Epic has a bridge to the file-management program, the document can then be directly uploaded into Astoria to create the official bibliographic record.

3. Proofreading: A second editor, other than the initial cataloger, then prints out the SGML document and checks its data against the original for errors as well as checking the SGML for consistency. When the volume has been completely proofed, the proofreader checks each document out of Astoria, enters the correction, and checks the document back in. Each editor involved with the creation of an SGML document is recorded along with their role in production in the document's metadata.

4. Publishing: A volume's bibliographic and catalog record are then created. The catalog record, which stores each of its text's bibliographic record in a structured hierarchy, makes use of Astoria's file-sharing capability so that it imports the texts' bibliographic records upon publishing and there is no need for duplicating records. Once the volumes catalog file is published and exported out of Astoria, it is pasted into the editions catalog file and processed through the web-publishing software with associated style-sheets to produce the final catalog.

For major centers of the NGB project, we would like the center to be accessing Astoria directly in its location on the Jefferson village at IATH. To do so, the remote location needs to have a sophisticated SGML editor - such as Epic or Adept Editor (running between one and two thousand dollars) or Framemaker Plus (much cheaper)- which can be integrated into an Astoria Client software (which is relatively inexpensive). "Client software" provides a web-interface for the checking in and checking out from Astoria. The client software allows one to log on to Astoria over the internet and check out and check in documents just as those within the LAN. Adept Editor or comparable SGML editor is necessary to be actually able to do something with these documents once they are checked out. Obviously such access to our files is something that will be tightly controlled, and only possible for central figures in the Collection. Finally, it should be noted that the client software for Astoria is only available on Windows95/8 and NT at the moment without any plans to port to Macs.

v. Individual scholars: remote submission forms

The third phase involves expanding the project not only to multiple technologically advanced centers, but also to individual scholars, librarians and archivists who may have very limited technological competence, and limited access to technological assistance. We are exploring the use of SGML-based on-line forms and simple guidelines that allow remote submission of materials that tag new material in the background. In this way, individuals can cut and paste work done in basic word processors and thus submit SGML documents into the project which require minimal post-processing.

vi. Multiple user modalities

We have stressed collaboration at the level of facilitating the generation and revision of texts and images. However, collaboration also applies at the level of users. We are designing four modalities of user access to the thematic research collection in order to solicit use *and* contributions from four distinct sets of individuals:

1. Tibetologists: who can read Tibetan language
2. Western language scholars: scholars in general other than Tibetologists
3. Tibetans: particularly those unfamiliar with any Western language
4. Popular audience: the general, public audience with any type of interest in Tibetan culture, Tibetan religion or Buddhism.

By presenting the collection in various modalities that facilitate use by each of these four groups, we aim to build bridges of collaboration and communication between groups that often lack access to the activities and scholarship of each other.

i. Tibetologists and other scholars

To begin with, the simplest bridge is most likely that between Tibetologists and other scholars. The increasing specialization of scholarship has created separate fields of knowledge and data, the boundaries of which are reinforced by the need for philological research into a given language, and/or the use of arcane methodological procedures and terminology. While the underpinnings of our collection are highly specialized philological work, we aim to thereby make accessible these materials to a broad scholarly community. We hope to solicit scholarship on these traditions from non-Tibetologists in various disciplines who would otherwise feel unable to deal with Tibetan materials.

ii. Western scholars and Tibetan scholars

The second bridge is between Western scholars and Tibetan scholars, who are usually separated from each other on linguistic, methodological, financial, and other grounds. By enabling direct communication between these two groups and their scholarly activities, we aim to generate Tibetan scholarly contributions to the collections, and encourage new collaborative scholarship between the two groups. Not only will this enable advances otherwise hampered by the limitation of each group, we also hope it will generate new methodological approaches and understanding through the proliferation of hybrid scholarship. In support of this, we are exploring the opening of offices for text input in Eastern Tibet, as well as developing technological outreach initiatives for Tibetan educational institutions.

iii. Scholars and popular audiences

The third and final bridge is between scholars and a more popular audience. Generally scholarship targeted at other scholars and scholarship targeted at a popular audience are of necessity completely divorced from each other and require separate publications. The electronic and Internet-based nature of our project allows us to integrate these two target audiences, at times simultaneously. This enables highly specialized scholarship to generate products of interest and access to wide ranging audience.

Conclusion

In summation, we feel that such thematic research collections promise not only hold promise in greatly expanding our knowledge about Buddhist literature, but also in eventually transforming the very ways we go about our research on, and consequently our understanding of, Buddhist literature. It is thus an incipient revolution in the extent *and* character of knowledge in Buddhist Studies.