Digital Technology and the Representation of Himalayan Cultural Geography

Dynamically plotting the features of a multilingual XML Gazetteer and connecting to associated data and resources

David Newman, University of Virginia

Goals

This paper outlines three general goals within the context of a digital library of Tibet and the Himalayas: to create a method for tracking data about places and in turn media pertaining to those places; to access media by location as opposed to theme or media type; and to manage the acquisition, input and retrieval of multilingual data. University of Virginia's Tibetan and Himalayan Digital Library (THDL) [1] is currently building an integrated system for connecting maps, resources, and data about places. THDL sees the creation of a digital gazetteer (the *Gazetteer of Tibet and the Himalayas* [2]) with basic naming, spatial coverage and feature type details as the key to linking maps and the full extent of the library's media holdings.

The *Gazetteer of Tibet and the Himalayas* is both a collection of meticulously documented data about places and a point of access to THDL's array of holdings. The goal of the Environmental and Cultural Geography initiative [3] in particular is to create an integrated system linking the *Gazetteer* with maps, a multitude of media objects (images, audio/video, texts), encyclopedic entries, and in depth analyses and explorations in the form of essays and web sites. Additionally under development is a set of more refined databases for handling data specific to such features as monasteries, archeological sites, and architecturally significant features.

This paper reflects a focus on the connectivity between the *Gazetteer of Tibet and the Himalayas*, maps and geo-referenced media objects. It will first address the need for the technology within the greater structure of the library. Then it will deal with the challenges in building an integrated

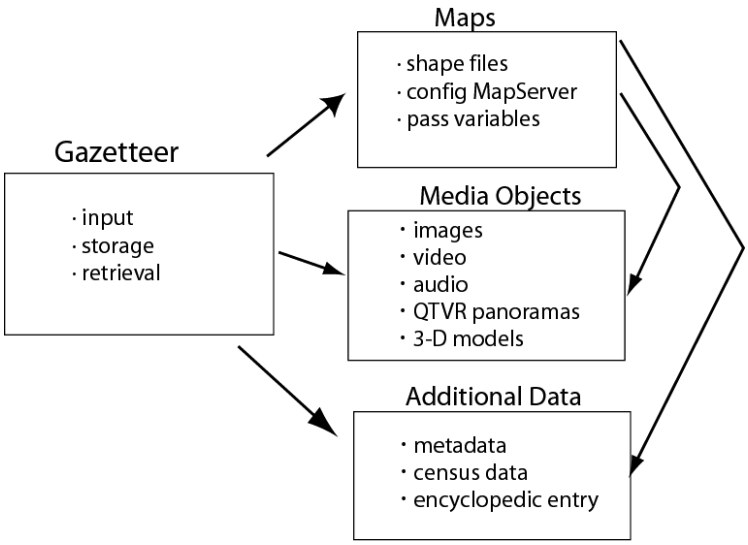
gazetteer/map/resources system, paying particular attention to issues in working with the multinational, multilingual region of Tibet and the Himalayas.

THDL Overview

THDL was founded in 2000 as an initiative to build a comprehensive and integrated digital library infrastructure concerning the geographical and cultural regions of Tibet and the Himalayas. The goal of THDL is to support interdisciplinary, multimedia academic work in these fields of study. The content and organization or THDL is cooperatively run by many institutions and individuals around the world, with the University of Virginia Library committed to long-term support. THDL functions as a centralized workspace, archive and publisher providing tools, long-term repositories, and collaborative networks.

THDL offers:

- Infrastructure: databases, web storage, community;
- Digital tools: fonts, software for language learning — time coding transcriptions and translations of movies and audio, codified system and standards for digital contributions;
- Reference tools: dictionary, encyclopedia, bibliographies, community roster, gazetteer;
- Raw data;
- Analyses;
- Movies, maps, images, 3-dimensional visualizations and audio in digital form.



There are multiple entry points to those resources, primarily through thematic and geospatial approaches. For instance, historians can begin their exploration of THDL's materials thematically via the History Collection [4]. However the alternative (i.e. geospatial) approach is to visit an area as broad as the Himalayas or as narrow as a building in a monastery either through a text listing (the *Gazetteer of Tibet and the Himalayas*) or maps, and then cull data about the place and access additional resources, analyses, essays, and so on.

The *Gazetteer* is a mixed media, locality-based cultural and environmental guide to Tibet and the Himalayas. Access to the studies, data and media objects is provided by a cartographic interface via digital maps, by navigation through hierarchically organized lists of places, and by text-based searches. And like the other tools of the library, it is designed to be compatible with the greater Digital Library initiative of University of Virginia. [5]

Methodology

In developing a three-way system of gazetteer/maps/media resources, the starting point was GIS data. From there development began on a Gazetteer and the issue of geo-referencing the various THDL media holdings. At the same time THDL was experimenting with a number of methods for visualizing the spatial data on maps and using maps to create a cartographic interface to our resources and data.

A. GIS

At first the goal was to geographically reference materials already gathered by THDL. To that end, in 2000 the library obtained a set of geospatial data from a variety of sources, compiled by Karl Ryavec of University of Wisconsin at Stevens Point. Sources for this data included the National Geospatial-Intelligence Agency (NGA) [6], the China in Time and Space (CITAS) project [7], the Australian Centre of the Asian Spatial Information and Analysis Network (ACASIAN) [8] and print documents obtained during fieldwork in Tibet.

Ryavec compiled the data in a GIS [9], plotted township seats from medium-scale maps, approximated Pinyin names and entered Chinese GB (__; *guobiao*) ID numbers from 1990 in order to allow for the creation of a hierarchical listing of features.

China in Time and Space has made publicly available extensive socio-economic data in the form of 1990 census data at the county level. This data is keyed to the GB number.

The post-processing also included the creation of full documentation compliant with FDGC standards for the shape files.

The types of data in the GIS includes the following: [10]

- Administrative units:
  - ♦ Country: polygons (i.e. traced outlines)
  - ♦ Prefecture level: polygons; name in Pinyin
  - ♦ County-level: polygons; extensive socio-economic data
  - ♦ Township-level: points of township seats; thiessen polygons approximating general area covered; minimal socio-economic data
- Land cover: 52 types as polygons with names; not classified according to any greater land cover typology
- Hydrology
- Major roads
- Satellite imagery (1000-meter resolution)
- A digital elevation map generated from the satellite imagery

Special Points about Tibetan data

The ethno-cultural region of Tibet and the Himalayas covers areas in multiple countries, including China, India, Pakistan, Nepal and Bhutan. Because of the alliances formed with the international scholarly community and the data that is freely available, China was chosen as the test bed. However there are plans to extend into other relevant countries.

With GIS, it is possible to query across geospatial, socio-economic and environmental data. However the data originally provided to THDL had serious

shortcomings with regard to minimal multilingual support. During the process of importing into the GIS the Chinese place names that had been entered using a Chinese word processing application called TwinBridge, the names were rendered illegible. Furthermore the GIS data has no Tibetan script. Although at this point the data has Unicode Chinese text, further testing needs to be done for rendering Unicode Chinese and possibly Unicode Tibetan script in a GIS application. THDL has entered the names of 1163 TAR features in THDL Extended Wylie and joined that with the GIS shape files, keyed to the Extended GB code. Although the Harvard University China Historical GIS (CHGIS) [11] has had success with Chinese names and data within a GIS, THDL still needs to apply additional resources and testing to show Tibetan script and Chinese character renderings.

Deploying the GIS

In addition to maintaining an offline GIS, the data sets that are THDL's work product will be available online. Where data sets are available from other organizations, links are provided. However THDL is committed to web deployment of as much of its data as possible. To that end, several systems have been explored, each with its own benefits and drawbacks.

ESRI ArcIMS: This requires the least in-house processing of the GIS shape files, but it is the most demanding of the server and the viewer. Although it is ostensibly supported by Windows NT servers, it is unreliable. The client is a Java application that is unwieldy for many users and unacceptably slow, particularly for overseas users or those who have slow Internet connections. Furthermore it is not supported for MS Internet Explorer for the Mac. If users manage to penetrate such access barriers, it is unlikely that they will be able to figure out how to query the data unless they are already experienced in using GIS applications. For those reasons and the unlikelihood that it could support a multilingual approach, the GIS-savvy user would most likely prefer to download the data sets. [12]

MapServer: This open source solution is currently being used in tandem with the *Gazetteer*, however it warrants further exploration. A simple HTML template and a text file is used to configure the application (a ".map" file), which renders shape files very quickly to a wide base of browsers across platforms. For this reason and because parameters for rendering the maps are passed in the URL, it was the clear choice for mapping a result of a *Gazetteer* feature query.

Macromedia Flash: This is the most time-intensive for THDL processing but seems to be the best solution for providing user-friendly maps with real control over the viewer's experience. Because the data can be pre-rendered as graphics, this is the only way to ensure that a user will be able to view names in their correct characters and scripts. Linking off to other web-based databases — whether they are XML or PHP/MySQL or HTML — results in compromised control of what the user experiences. However, as long as the user's computer is prepped with the appropriate fonts and browser, they should be able to view the associated multi-lingual data.

Though the Flash player is purported to reach 98% of web users, [13] requiring THDL visitors to have a recent version of the plug-in can be an insurmountable inconvenience, particularly to those visitors with slow internet connections or lack of facility with English. The main limitation is that "canned" demonstrations made in Flash as opposed to a true GIS offer only a minimum of querying capabilities.

IDs and Keys

Within the GIS the features and data are keyed to the extended GB code. With a shift towards a greater compliance throughout THDL as well as the University of Virginia Library of Tomorrow and its foundation in FEDORA technology, features must be keyed to a PID (persistent identification) or URI (universal resource indicator). [14]

B. Gazetteer

At its most basic, a gazetteer is an index to place names — a sort of "geographical dictionary". A more extended definition would include "time-stamped names, extents, and relationships; descriptive information about names and places; merging of information about a place from multiple sources". While according to Alexandria Digital Library, the preferred definition would be a "spatial dictionary of named and typed places". [15]

The role of the Gazetteer in THDL

THDL's intention is to document Tibet and the Himalayas through focusing on the various regions and points within that area, the history of various names given to those places (*toponyms*), and the characteristics of the environment, culture and history linked to those places. According to one description of Geography Markup Language (GML), a geographic feature is "an abstraction of a real world phenomenon; it is a geographic feature if it is associated with a location relative to the Earth". [16] This may be a tangible object such as a mountain or bridge, or it may be an intangible administrative grouping such as a country, province, region, etc. The backbone of THDL's Environmental and Cultural Geography initiative is the *Gazetteer of Tibet and the Himalayas*, which exists in two main capacities: it is an index of features, with each feature record potentially containing data on toponyms, identification codes, location (latitude, longitude, etc.), and relationships between features (what features a given feature is contained by, contains, or intersects with). It also contains a simple summary that provides a short general description of the feature for a first glance view. A record does not contain such detailed feature data as the population of a nation or the architectural composition of a building. Instead, the *Gazetteer* is used essentially as a catalog of places/features and their names, which is then used as an index throughout THDL. The Gazetteer feature code — a unique THDL ID for each feature — is used to index photos, videos, texts and other data within THDL according to location, while references to geographical

places and features in general utilize the ID for short hand reference that can be used to quickly look up the relevant documentation within the *Gazetteer*.

The THDL ID is also used to relate *Gazetteer* entries to more extended descriptions of the same feature within specialized databases. Thus THDL is creating analytical templates for a series of associated databases to document more granular and expansive information about a given type of feature. While the *Gazetteer* is a single repository with a single standard set of fields, these descriptive feature databases are plural repositories with varying sets of fields that are suited to specific groups of affiliated features. The descriptive feature databases primarily contain extended analytical data about a given feature, including both categories of information specific to that kind of feature, and extended essay length descriptions pertaining to the feature.

The following is an initial list of distinct sets of features that are under development or consideration:

- Buildings
- Monasteries
- Nations
- Administrative regions
- Ethno-linguistic regions
- Neighborhoods
- Archaeological sites
- Environmental regions and features

For each of these sets of features, the goal is to create a structured database with a user friendly interface, and most likely one which combines Gazetteer fields with descriptive feature databases fields, so that both can be edited simultaneously. Once the work is complete, scripts will then process the database into two different forms: the Gazetteer fields will be exported into the XML *Gazetteer of Tibet and the Himalayas*, while the descriptive feature database fields will be exported into a GDMS (University of Virginia's General Descriptive Modeling Scheme) model. [17]

Data model: The Gazetteer fields

The *Gazetteer of Tibet and the Himalayas* is the basis of THDL's geographically-based archive of holdings and index to catalogued features. It comprises:

- Comprehensive documentation of all the variant forms of the name of a given feature, including names in different languages, vernacular vs. official names, and names in different time periods.
- Information about the feature's location, such as latitude, longitude, and altitude.
- Specification of related features, such as what other features might be contained by the feature, or what other features might contain this feature. Thus for a county, it documents the prefecture, province and nation in which it lies, as well as townships that it in turn contains.
- Summary description from one paragraph to one page in length.
- All relevant identification numbers for the feature in government systems, or other important databases, and a unique Feature ID number within THDL. The THDL Feature ID is used in all other parts of THDL to refer in short hand to this feature, such as in citing in the image database the place where a photograph was taken, the birth place of an author, and so on.

Mapping features into the 1990 Administrative Hierarchy

In order to provide an organizing structure for the features, THDL decided to place them within a national hierarchy. Nevertheless it is also vital that Tibetan features be able to be located by a concurrent overlay of traditional Tibetan cultural regions. For reasons described above, China was chosen.

Administrative levels of PRC

Although the Chinese administrative structure is well defined, [18] multiple feature types can define units at the same administrative level. For instance provinces, autonomous regions and certain "specially administered municipalities" are at the same level. At the next level, the principal units are prefectures, autonomous prefectures, municipalities and districts. More granular administrative units include "collections of urban units", "townships with registered urban population", neighborhood committees, etc. Another potential

cause of confusion is that certain terms like "city" (*shi*) signify one level within a municipality while signifying a different level in a prefecture.

GB Codes

China uses a system of numbers for identifying three tiers of administrative units, from province to prefecture to county levels. This is called the GB or (*guobiao*) table. These numbers embedded in the feature data could be used to begin to create a hierarchical model and place the administrative units within that.

THDL, in conjunction with Ryavec and CHGIS, developed an "Extended GB code" expanded to five pairs of digits (as opposed to the standard Chinese three pairs), e.g. 5423010000 for a county-level feature. This allows for a semantic ID at the sub-township level, which allowed THDL to automate the creation of an administrative hierarchy, readily indicating the parent administrative feature. Furthermore, this system indicates in the fourth pair the type of township-level feature (e.g. Admin Level 4). Here is the semantics of the fourth pair:

- 01: a "township with urban registered population" (*zhen*) which is also a county seat (e.g., THDL ID f117 — 5401270100)
- 02-49: a "township with urban registered population" (*zhen*) which is not a county seat (e.g., THDL ID f6 — 5401010300)
- 50: a "township without urban registered population" (*xiang*) which is also a county seat (e.g., THDL ID f40 — 5401215000)
- 51-99: a "township without urban registered population" (*xiang*) which is not a county seat (e.g., THDL ID f122 — 5401275500)

The fifth pair is set aside for physical features (also known as "spot features") and sub-township level administrative features within a township level feature. Because only a minimum of physical features have been catalogued, the example of a river crossing multiple administrative units has not been extensively faced. However, since the "partof" XML attribute is repeatable in the XML structure, we will be able to attribute a feature to multiple parent features. [19]

Collating, checking, creating and documenting data

Deriving names and latitude/longitude data from the GIS

In 2001, THDL used ArcInfo to generate a data file with Chinese name, Pinyin name, GB code and latitude/longitude coordinates to five decimal places for all the province level, prefecture level, county level and township level features for which there were records in the GIS. This totaled 2038 features, but the target features were those that were in Tibet Autonomous Region — approximately 1100 features. The Chinese names which were encoded in Big-5 had been misrendered in the GIS and were not always correct, so in conjunction with Harvard's China Historical Geography project, an Excel spread sheet with traditional Chinese characters was created, from which the generation of simplified characters and exact Pinyin could be automated. Access to gazetteers, census documents and books listing variant feature names in both Chinese and Tibetan for TAR made it the clear choice for the test bed. THDL reviewed the Chinese and Pinyin that had been entered and used these resources to enter Tibetan names according to the THDL Extended Wylie scheme. [20] Variant names are a key issue because a feature may have multiple Tibetan spellings and Chinese character representations: orthographic variants can be due to different time periods or forms of speech (colloquial versus formal) or simply mistaken spellings found in various sources.

Metadata for name and spatial data such as the sources, dates, those responsible for entering the data, etc were noted in the data file.

Benefits of an XML structure for storage

After careful consideration, it was decided that an XML structure for data storage surpassed other data models for the *Gazetteer*. The following points summarize the benefits:

- Multiple organizing structures (administrative, regional) can be encoded into a feature's record with the "partof" element;
- Multiple time periods can be dealt with by the "covtime" element;

- Multilingual names and metadata are readily handled by XML with its native Unicode;
- Related names are simply documented with the "frel" element and further described by the "type" attribute;
- The use of elements is extensible: It is up to the editor whether to include optional elements, which is opposed to a relational database, which requires that all the structures be included and attached at all times. The entire system is built around the most complex example;
- Data is easily shared: The data is structured and self explanatory, so a chunk of the XML can be extracted and shared, yet still rendered meaningful as long as the DTD is available. Markers are encoded in the data, whereas with a relational database, if the data is separated from the field names, it is meaningless. In a relational database, all relations must also be explicit in every bit of exported data. With XML, requests between systems, such as the gazetteer and an XML image database are smooth because the data is defined in itself.
- Features inherit the attributes and characteristics of their parent features;
- Repeatability is conveyed in a straightforward way: For instance a name with its embedded permanent ID can appear anywhere, as an entry, part of a paragraph or something bigger. A related database requires the presence of many fields that will often have no data in order to accommodate the possibility of repeated data.
- Hierarchical structures are easily conveyed because XML is itself hierarchical.

The DTD

The document type definition (DTD) contains the rules by which the XML Gazetteer must comply. ADL's Gazetteer Content Standard was considered; it is described as

> ...designed to be a comprehensive framework for recording descriptions of named geographic places, including the core elements of toponyms (and their history), spatial location (in various representations), and classification (according to referenced typing schemes), and source attribution for pieces of description gathered from various resources for a particular place. [21]

The Content Standard was a useful starting point, but THDL's needs dictated additional elements and attributes, in particular with regard to the range of names that could be attributed to a feature in Tibet over time. Throughout the DTD development process THDL has kept with principles of GDMS and Geographic Markup Language (GML).

Displaying and querying the Gazetteer

Perl query

The mechanism for querying the *Gazetteer* is a Perl script running a SGREP search. From the *Gazetteer* search page, the search is run across the entire XML document; until a new search method is implemented later in 2004, performing searches within the results (such as the name instead of the ID of the parent feature) is limited.

XSLT to transform XML to HTML

An XSL style sheet was constructed to render the XML data as HTML for the web. This style sheet works with a servlet engine to produce the results viewed online. The HTML includes Unicode entities that are rendered as Chinese script if the client has installed the proper fonts. [22]

Tibetan-specific issues

Because no Tibetan Unicode font was in existence at the time of creating the XML data, THDL opted to store Tibetan according to its Extended Wylie scheme. As a Unicode Tibetan font is developed, a conversion system will be applied to all Tibetan names to render them in HTML.

Mapping Fields to the DTD

Due to a dearth of user-friendly XML data input tools and the need to share data with contributors at various technical levels, the source data for the *Gazetteer* was entered in an Excel spreadsheet. Because of the need to provide fields for several Tibetan names, several Chinese names, and the metadata for each, the spreadsheet had approximately 175 columns of data of Roman and Chinese scripts — the Tibetan names were stored in Wylie. A document was created mapping the fields in the spreadsheet to corresponding elements and attributes defined in the DTD. This spreadsheet was exported as UTF-8 text, which was then converted via Perl script to a single XML file of 1163 records.

C. Interaction between media

FEDORA and GDMS

By working within the standards of FEDORA and GDMS, the Gazetteer is compliant with the University of Virginia Digital Library.

Gazetteer links from a feature record

MapServer

The open source application MapServer is used to dynamically plot the queried feature on a map. In this case, the GIS shape files serve as the background for the map. The process is relatively straightforward: The shape files are stored on the server. An HTML template defines the layout of the map and a text configuration file (a ".map" file) defines the rendering of the shape files. Then variables about what feature to highlight and which feature types to display are passed in the URL. The XSL style sheet creates a hyperlink with the parameters for MapServer, primarily the Extended GB code, to which the shape file is keyed.

It is still necessary to test Chinese script, since at this point the MapServer map is just using Latin encodings. Plotting the feature based on latitude/longitude coordinates instead of adding it to the shape file must also be implemented. Additionally the THDL ID must be keyed to the shape files.

Census data

Freely available on the web is county level census data with several hundred census variables for 1990. A MySQL database was created with records for each county level feature in TAR keyed to the THDL ID.

For county level features in the Gazetteer display, the XSL renders a URL that performs a PHP query of the MySQL database of census data, passing the Extended GB number as a variable. The PHP-generated HTML page displays only a sample of census variables.

The MySQL tables are not storing Unicode Chinese, but rather entities representing Chinese characters. For instance, although feature f3 is rendered as " 拉薩 " it is stored as "拉薩" Whether this is the best method warrants further research.

Other resources and media objects

Although this is one of the more important functionalities in the THDL model, it is most under development, which is in part due to the number of media objects. When the first pass of cataloguing the over 25,000 images was conducted, a permanent THDL IDs for geo-referencing had not yet been created. At this point the Sera Monastery case study below is the only example of querying the image database from a map. However in other places throughout THDL such as web pages, searches are performed on the image and audio/video databases. As the objects in the media databases become geo-referenced by adding the THDL ID for the features they document or where they were recorded, this will facilitate the next step in the Gazetteer development: within the results of a Gazetteer feature search to be able to query across the various databases for any records that call on that feature's THDL ID. With that same querying technique the intention is to query a burgeoning Encyclopedia.

A link to a web site or other in depth study of a feature is embedded in the feature's description. Over time, the goal is to switch to a model in which any detailed web site (such as THDL's portals to Lhasa or Sera Monastery) would be just another geo-referenced object, which would then appear as a link in the relevant Gazetteer entry.

III. Case Study: Sera Monastery

Sera Monastery provides a case study in the interconnection of data, map and resources. The main points in displaying the final product are as follows:

- XML Gazetteer is rendered in HTML with XSLT;
- Results are plotted on a web deployed MapServer interface;

- Results are plotted on a more interactive map environment made possible in Flash; [23]
- PHP calls query the MySQL databases on the feature, its parent features and images;
- Ultimately an encyclopedic textual entry will be retrieved, as well as links to the various media to which it relates, i.e. audio, video, still images, QTVR panoramas and interactive 3-D models.

Here are the steps in the process:

- Conduct initial research: Professor Jos Cabez n of University of California at Santa Barbara conducted the background research on Sera Monastery.
- Obtain a base map: A 1984 1:1000 scale Chinese survey map with basic topography was used as the base map. [24] Tibet Academy of Social Science provided access to a large print version of the map which was useful for clarifying details.
- Revise the base map: Through daily visits to Sera over the month of July 2002, the base map was annotated to indicate where features had changed. There had been moderate reconstruction since the original survey.
- Collect associated data: Daily research also involved interviewing residents in order to obtain such information about features as date of construction/reconstruction, college (*grwa tshang*) affiliation, regional house (*khang tshan*) affiliation, demographics of residents, traditional and contemporary regional affiliations of residents. Additionally features were photographed and GPS readings taken.
- Create databases: At University of Virginia the raw data was entered into databases of the features, both administrative and physical (colleges, regional houses and then approximately 95 physical structures) and the images catalogued.
- Create web site: An extensive web site based on the scholarship of Dr Cabez n was created. [25] The site contains a series of XML-based, mixed-media essays, transformed to HTML on a servlet engine.
- Revise shape files: The shape files called on by the MapServer application were updated. This was necessary in order to carry out the dynamic rendering from the *Gazetteer* entry for Sera.
- Create *Gazetteer* entry: An XML Gazetteer entry was created for the monastery as a whole. A link to the Sera web site was hard-coded in the description, whereas the link to the shape file is a product of the XSL style sheet.
- Digitize satellite imagery: THDL obtained 1-meter resolution IKONOS satellite imagery, which was then digitized using ESRI's ArcMap. This resulted in approximately 100 distinct features, including large empty fields.
- Derive vector graphics: From ArcMap features were exported as vector graphics and then imported into Adobe Illustrator, wherein each was placed in

its own layer. This file was imported into Macromedia Flash, which recognized each layer and object as defined in the Illustrator document.

- Process the feature objects: This involved creating movie and button symbols, identifying the objects with the related feature's ID and adding a in include to provide button functionality.
- Process the feature records as a single XML file: In the interest of providing an XML file that is quicker for Flash to parse, only a subset of the feature's records is generated as an XML file. Though it would be desirable to access all of the data and metadata about the feature, for the purposes of configuring the objects on the interactive map, the most important attributes for the feature are its name, college affiliation and feature type. Additionally, the XML structure for the records was modified in order to minimize modification of existing ActionScripts. [26]
- Create Flash/XML interactivity: ActionScripts were written to read the XML file and then generate "tool tips" and control the appearance of the objects.
- Link to MySQL databases: A PHP script within an ActionScript allows Flash to query the MySQL tables which have extended entries on each of the physical features and the MySQL image database records corresponding to the feature. The results of the feature query is an HTML page with data about the feature and thumbnails of corresponding images. This results page allows direct queries of the feature's college and regional house administrative features and the ability to click on a thumbnail and retrieve a larger version of the image, complete with the catalog entry for it.

IV. Current challenges

Data structures

Data has been made web-accessible often on an ad-hoc basis, calling on tools developed in league with other members of the team and University of Virginia, often for other purposes.

Multiple data structures

Content has been shuttled from one system of structured data to another in nonstandard fashion. For example data was converted between XML and MySQL, which is not problematic, however it was done without standardized scripts, and therefore is not an acceptable long term method.

Version control issues

The examples wherein the systems of *Gazetteer of Tibet and the Himalayas* have been fleshed out — maps both dynamically rendered from shape files and processed in Flash, THDL web sites, and the THDL Image Database interconnect — are in large part a proof of concept and have arisen from the pressing need to present a public face for the work undertaken. The Sera map calls on a subset of the THDL Image Database with extended explanations and Unicode compliance, both of which are lacking in the greater Image Database. Furthermore Sera features exist in MySQL tables which need to be integrated into the greater XML Gazetteer. Additionally, the Flash maps read an XML file which is used to configure the objects based on XML attributes, but that file contains records that have been modified apart from their entries in the greater *Gazetteer of Tibet and the Himalayas*. There is also the need to standardize the attributes for XML files used with Flash maps and to develop an XSL for outputting the data from the greater Gazetteer.

Input tools

The multilingual nature of THDL's work continues to create extensive challenges. A pressing need is to develop an input method for creating XML Gazetteer entries in a more user-friendly way then a text editor such as jEdit. Though this may be acceptable for a more technologically savvy editor, for the majority of scholars of Tibet and the Himalayas, this is beyond the bounds to which most can be expected to go.

The development of a freely accessible Unicode Tibetan font is underway but incomplete. The paucity of Unicode Tibetan fonts drastically limits the collaboration of non-western scholars of Tibet who are not very conversant with Wylie. THDL has developed a convention for embedding Tibetan spelling, a phonetic rendering and a translation for inline entries. Here is an example:

{Dzokchen G npa (Dzokchen Monastery) [rdzogs chen dgon pa]}

The role of Unicode in PHP/MySQL has continued to be problematic, both for Gazetteer entries and media object cataloguing. Although Microsoft Access supports Unicode the lack of a Macintosh version prevents it from being viable.

Although MySQL is not fully UTF-8 compliant, it seems likely that this will be worked out soon. We have not had problems rendering the diacritics or Chinese text we have tested. The audio/video database is currently being converted from ColdFusion to a PHP/MySQL solution.

It seems that FileMaker 7 may be an excellent short-term solution because it offers true Unicode support as well as a new data model for handling multiple related tables. In the very near future, we will be testing a provisional Unicode Tibetan font with FileMaker 7.

Gazetteer search method

The search mechanism in the current version of the Gazetteer is both slow and limited. These two shortcomings will be dealt with in the next iteration as the technology used by the University of Virginia Digital Library is more clearly settled. The Perl scripts for searching are inefficient and do not allow for searches such as a feature name *within* a given feature type. In the meantime an XML search engine such as Tamino or XPath will likely be employed.

Feature typology

As more records are catalogued, a more comprehensive feature typology will hopefully develop organically. Since the majority of features entered into the *Gazetteer* have been administrative ones, the full breadth of feature types is still to be encountered.

Conclusions

Through a system of structured XML and MySQL databases, XSL style sheets, PHP, HTML, Flash and MapServer, and GIS software such as ESRI

ArcMap, THDL has made inroads in making accessible the resources it has acquired and created over the past several years. Although there continue to be challenges along several fronts, the continued development of Unicode compliant tools will improve the collection, creation and display of THDL's data and resources.

Acknowledgement

Thanks to David Germano, the Director of THDL for contributions to this paper.

Notes

[1] The Tibetan and Himalayan Digital Library <http://thdl.org>

[2] The THDL Gazetteer of Tibet and the Himalayas <http://iris.lib.virginia.edu/collections/cultgeo/gazetteer-frameset.html>

[3] The THDL Cultural Geography portal <http://iris.lib.virginia.edu/collections/cultgeo/index.html>

[4] The THDL History Collection portal <http://iris.lib.virginia.edu/collections/history/index.html>

[5] Library of Tomorrow <http://www.lib.virginia.edu/digital/info/LofT.html> and FEDORA — see reference: http://www.lib.virginia.edu/digital/resndev/fedora.html and http://www.lib.virginia.edu/digital/resndev/examples/examples.html

[6] National Geospatial-Intelligence Agency <http://www.nga.mil/portal/site/nga01/>. At the time it was known as the National Imagery Mapping Agency (NIMA)

[7] China in Time and Space <http://citas.csde.washington.edu/>

[8] The Asian Spatial Information and Analysis Network: <http://www.asian.gu.edu.au/>

[9] "Geographical information system", though GIS is increasingly referred to as "GIScience".

[10] See the "GB Codes" section of this paper for an explanation of the extended GB scheme we arrived at between THDL, Ryavec and Berman, with input from Crissman at Griffiths.

[11] China Historical GIS <http://www.fas.harvard.edu/~chgis/>

[12] A limited portion of the data sets is currently available from <http://iris.lib.virginia.edu/tibet/collections/cultgeo/gis/index.html>. More will be made available as they are ready.

[13] Macromedia Flash Player Statistics <http://www.macromedia.com/software/player_census/flashplayer/>

[14] The Fedora Project: An Open-source Digital Object Repository Management System <http://www.dlib.org/dlib/april03/staples/04staples.html>

[15] Guide to the ADL Gazetteer Content Standard. <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>

[16] Coverpages: Online Resources for Markup Language Technologies <http://xml.coverpages.org/ni2003-02-06-c.html>

[17] An advantage of GDMS is that it allows for a far richer description of the relationship of features to each other within a site — it can build a comprehensive and detailed model of the entire site that expresses all the internal relationships. See GDMS (General Descriptive Modeling Scheme) Introduction : <http://www.lib.virginia.edu/digital/metadata/gdms.html>

[18] It should be noted, however, that administrative units are reconfigured at regular intervals, marking the creation, elimination, promotion and demotion of administrative features.

[19] See the XML DTD for the Gazetteer: <http://iris.lib.virginia.edu/tibet/collections/cultgeo/documentation/gazetteer/gazetteer-new-dtd.txt>

[20] See THDL Extended Wylie scheme <http://iris.lib.virginia.edu/tibet/collections/langling/tibetan-transliteration.html>

[21] See Guide to the ADL Gazetteer Content Standard: <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>

[22] See Gazetteer XSL style sheet: <http://iris.lib.virginia.edu/tibet/collections/cultgeo/documentation/gazetteer/gazetteer.xsl>

[23] The system relied heavily on the model employed in University of Virginia's Institute of Advanced Technology in the Humanities maps on Evolutionary Infrastructure: Boston's Back Bay Fens <http://www.iath.virginia.edu/backbay/fenssite/html/maps/map01.html>

[24] The map is the frontispiece in Tshe dbang Rin chen (1995) *se ra theg chen gling*.

[25] <http://iris.lib.virginia.edu/tibet/collections/cultgeo/sera/index.html>

[26] In addition to being a mere subset of the feature's data, because of the way that the Flash map reads the XML, all of the information about the feature is attributes of each record as an element. This is at great variance with the greater Gazetteer in which multiple elements and attributes exist for each feature. Although the Flash map could be reconfigured to read down through the hierarchy of child elements, the simple 1 element per record makes it quick for Flash to parse.

Bibliography

Citations in this paper

Published Articles and Documents

Alexandria Digital Library (2001). Feature Type Thesaurus: <http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/>

Alexandria Digital Library. (2004). Guide to the ADL Gazetteer Content Standard: <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>

Berman, M. L. 2002). Multilingual Feature Classification Index for China and Japan: <http://www.fas.harvard.edu/~chgis/work/docs/papers/lex_pnc_osaka_081602.pdf>

Furlough, M., Germano, D., Newman, D., Roland, P., Staples, T. (2001). Creating the Tibetan and Himalayan Digital Library Gazetteer

Garson, Nathaniel, Germano, David, et al (2004). THDL Extended Wylie Transliteration Scheme: <http://text.lib.virginia.edu/servlet/SaxonServlet?source=http://iris.lib.virginia.edu/t

ibet/collections/langling/ewts/ewts.xml&style=http://iris.lib.virginia.edu/tibet/collections/langling/ewts/ewts.xsl&clear-stylesheet-cache=yes>

Hahn, T (2001). Structured and classified overview of independently treated, spatially relevant elements identified in 40 Chinese local mountain gazetteers: <http://www.library.cornell.edu/wason/gazetteers/>

Hill, L. L. (2000). ADL Gazetteer Content Standard: <http://www.alexandria.ucsb.edu/gazetteer/gaz_content_standard.html>

Hill, L. L. (2000). ADL Gazetteer Content Standard — Relational Database Model: <http://www.alexandria.ucsb.edu/gazetteer/gaz99.pdf>

Staples, T. and Wayland, W. (2000). Virginia Dons FEDORA: http://www.dlib.org/dlib/july00/staples/07staples.html

Tshe dbang rin chen. (1995). se ra theg chen gling.

GB 2260 — 91. "Zhonghua renmin gongheguo xingzhengquhua daima". (Codes for the administrative divisions of the PRC) - GB 2260 — 91. Beijing: Guojia jishu jianduju, 1992

TAR Statistics Bureau. (1990) TAR 4th Census Survey Manual of Collected Materials Bulletin Regarding the 1990 Main Census Data

(__1990_____)

#1: 8 November 1990; #2: 23 November 1990 (China Statistics Publishing House, 1990, Beijing, pp. 1-68)

TAR Toponymic Gazetteer, Vols I and II

Unpublished Documents and Working Papers

ECAI Electronic Cultural Atlas Initiative. (2002). Feature type thesaurus draft versions<http://www.mip.berkeley.edu/ecai/gazetteer/>

Germano, David and Newman, David (2004). Gazetteer Design.

Germano, David and Newman, David (2003). Sources for Tibetan Place Names.

Newman, David. (2004). Feature Types.

Newman, David (2004). THDL Gazetteer Entry Guide.

Newman, David and Ryavec, Karl (2003). Documentation of Ryavec GIS data.

Roland, Perry (2003). Gazetteer XML Document Type Definition

Roland, Perry (2003). XSL Transformations for Gazetteer Data

Web sites

FEDORA: http://www.lib.virginia.edu/digital/resndev/fedora.html

GDMS: http://www.lib.virginia.edu/digital/reports/metadata/gdms.html

CHGIS: http://www.fas.harvard.edu/~chgis/

ACASIAN: Australian Centre for Asian Spatial Information and Analysis Network. Griffith Univ, Brisbane, Australia

<http://www.asian.gu.edu.au/>

CITAS: China in Time and Space <http://citas.csde.washington.edu/>

CITAS: "GB CODES FOR THE ADMINISTRATIVE DIVISIONS OF THE PEOPLE'S REPUBLIC OF CHINA". China in Time and Space (downloadable dataset)

ADL project: http://alexandria.sdc.ucsb.edu/~lhill/adlgaz/

THDL: http://iris.lib.virginia.edu/tibet/index.html

THDL Gazetteer: http://iris.lib.virginia.edu/tibet/collections/cultgeo/gazetteer-frameset.html

Sera Site: http://iris.lib.virginia.edu/tibet/collections/cultgeo/sera/index.html

Image Database: http://iris.lib.virginia.edu/tibet/collections/resources/image_search.php

Lhasa maps: http://iris.lib.virginia.edu/tibet/collections/cultgeo/lhasa/lhasa-frameset.html

Sera map: http://iris.lib.virginia.edu/tibet/collections/cultgeo/sera/map.html

Tibetan Transliteration schemes: http://iris.lib.virginia.edu/tibet/collections/langling/tibetan-transliteration.html