Kantanjur and Tibetan Information Technology

The Kanjur and Tanjur, together the Kantanjur, form two of the mainstays of the Tibetan Buddhist canon. They record the words attributed to the historical Buddha as well as subsequent commentaries.

By Tashi Tsering

The Kanjur, as a collection, is characterized by Buddhist teachings that traditionally are attributed to Shakyamuni Buddha. Comprising 108 volumes, it includes the three *pitakas* (*vinaya, sutra, and abhidharma*) as well as the four classes of *tantra* and forms a major section of the Tibetan Buddhist canon. The Tanjur consists of commentaries on the Kanjur and related Buddhist literature by subsequent scholars and translators. Comprising about 225 volumes, it includes, among others, the canons of philosophy, literature, art, language, astronomy, medicine, and architecture. The Kanjur and most of the Tanjur are translated from Sanskrit. Tibetan and Mongolian monasteries nearly always house a set of the Kanjur and Tanjur, and most Tibetan and some Mongolian families aspire to have a set in their homes. An essential part of certain ceremonies, the Kantanjur is recited for special occasions and is revered as a sacred object by Tibetan Buddhists. Obviously it is very important material for Tibetological research.

Since Buddhism was introduced in Tibet in the seventh century, many copies of the Kanjur and Tanjur have been produced in both Tibet and other parts of China. Woodblocks for about thirteen different editions of the Kanjur have appeared in the past. Not only do the versions vary from one publishing house to another, variations in a particular publisher's woodblocks have been introduced over time. Some have different characters, different syllables, different sentences and even different texts. Of the many woodblocks of the Tanjur, the Derge edition is recognized as superior, with its rich content, unique format and clear layout. Although having been consistently used for more than 200 years, the Derge woodblocks of the Tanjur are still undamaged.

In 1986, the National Centre for Tibetan Studies at Beijing began to collect and collate the different editions of the Tanjur. With the Derge edition as the base, it was compared sentence by sentence with three other Tanjur editions, with each difference added as an annotation appended to the end of each text. Thus having a collated edition of the Tanjur is tantamount to having many editions. A group of Tibetan experts and laity are carrying out the project in Chengdu. So far, the complete Derge edition of the Tanjur with comparative annotations of three other woodblock editions has been entered into computer-files, half of which the National Press for Tibetology has published in a modern layout. In the next two years, all of the Tanjur books will be published. For the Kanjur, eight editions of woodblocks have been collected and the work of collation will be finished next year. The input into computer-files has begun autumn of 2000.

Because of the low price, search and retrieval capability, ease of storage and delivery, and the rapid advance of electronic publication technology and the Internet, the National Press for Tibetology is planning to publish the Kanjur and Tanjur in electronic form. It is also hoped that in the future the Kanjur and Tanjur can be made available worldwide via the Internet.. This article marks the beginning of this undoubtedly great project and monumental task.

A Digitized Kanjur and Tanjur

The electronic publication of the Kantanjur (EPKT) aims to fulfil the needs of two types of users: one is the scholar engaged in research, the other monasteries and homes, primarily using the Kantanjur for recitation and storage. It is not necessary to go into detail regarding recitation and storage, as these are not complicated issues. For research and selected recitation, however, user inquiries to the EPKT must be all-inclusive. The programme must be able to handle diverse queries from any user. An analysis of user queries to the EPKT is one of the purposes of this article. The types of query which have been identified to date are listed below:

- Search for a text in the canon by caption;
- Search for a text by the name of the translator or author;
- Search for a text by category;
- Search for a text by category and a volume sign within the category;
- Enter the name of a paragraph, then search for the text or texts that match the name;
- Enter a string of Tibetan characters, then search for the text or texts which match the string, regardless of whether they are in captions or the body of the text. The string could be the name of a person or a master, a sentence, and so on;
- Enter any combination of the above items, then search for the text or texts that match the conditions;
- Link any caption of a text to the body of the text. Display the text by clicking the caption of the text at any point;
- Statistical function: calculate any statistics which users may desire, i.e., number of texts in a category, number of texts translated by the same translator, and so on;
- An online Tibetan dictionary.

A Design Structure for the Electronic Publication of the Kanjur and Tanjur

The electronic publication programme manages the data and responds to the users operating the database. The Kanjur and Tanjur database must be arranged by text, because the text is the main element of the books. Operations to data are based on the user-requirements listed above.

Each text in the Kanjur and Tanjur has many attributes, which must be dealt with in the database. When a page of any text is opened, there are some attributes linked to the page and to the text. They are:

Book Name (Kanjur or Tanjur); Caption: caption of the text; Category; Translator's Name or Author's Name; Text; Paragraph Name; Volume Sign; Serial Number of the text in the Kanjur or Tanjur; Serial Number of the volume in any category; and lastly Annotation of Differences between the different editions of Kanjur and Tanjur.

The primary database key is the serial number of the text. The texts in the database are stored in serial number order, but every entry in the database contains all ten of the items listed above. Each text is stored separately in its own file.

The functions of the Kantanjur electronic publication programme are based on the user queries listed above. Depending on the query, there are two types of functions: information retrieval and data presentation. The information retrieval functions are:

- Category Retrieval: the user enters the name of a category or clicks on it, the programme lists the captions of texts of that category, then the user can click on a caption to display the text;
- Caption Retrieval;
- Translator or Author Retrieval;
- Serial Number Retrieval;
- Volume Sign Retrieval: the user enters a category and a volume sign;
- Volume Number Retrieval: the user enters a category and a serial number of a volume;
- Paragraph Retrieval: the user enters a paragraph or part of a paragraph;
- Text Retrieval: the user enters a string of Tibetan characters (a word, more than one word, part of a sentence or a whole sentence, even a paragraph);
- Combination Retrieval: the user submits a query using a combination of items listed above to find a text or texts, which meet the conditions.

The programme holds standard data presentation functions, such as edit, print, and zoom functions, choice of layout, background, sound, online dictionary, view annotation and colour, and statistical functions.

Nine information retrieval functions and nine data presentation functions have been noted in this article. We hope to receive feedback and advice from relevant specialists all over the world in order to improve the design structure of the electronic publication programme.

Kantanjur and Tibetan Information Retrieval

Large publications such as the Kanjur and Tanjur call for an excellent retrieval system, which may exact a high toll on resources of time and space. In order to reduce the expense, the operation of the information retrieval system must be optimized.
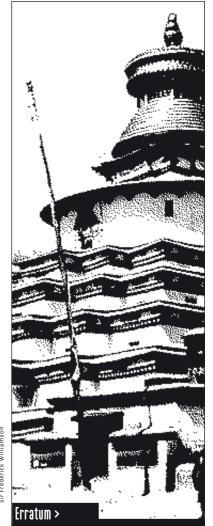
For the Kanjur and Tanjur, all forms of retrieval, except for text retrieval, must be based on a database system and hyperlink model (for instance: caption retrieval or translator retrieval), which will improve the speed of retrieval and save system space. Text retrieval must be based on Tibetan syllables, i.e., the smallest unit of retrieval must be a syllable, not a character. The submitted content or query from a user must be a string of Tibetan syllables. For optimizing the retrieval, spelling and grammar of the string must be checked before retrieval, which means that a spelling and grammar check must be included in the retrieval system. A spelling and grammar check, running under a Tibetan DOS-system, is in fact already in existence. Preparing one for Tibetan Windows system should not be too difficult.

Other applications

Apart from the above-mentioned aspects of Tibetan information technology, electronic publication of the Kanjur and Tanjur may also involve many other aspects of Tibetan information processing. Such as: sorting or ordering Tibetan in computer; a Tibetan online dictionary; a computational analysis of Tibetan; a Tibetan knowledge library; and a machine translation for Tibetan.

The Kanjur and Tanjur contain important Tibetan classics. To preserve and present them to people all over the world is our genuine desire. Computer technology can play a very important role in their (electronic) publication, thus combining Tibetan culture and modern technology. To enhance this process, we hope that experts from all over the world will come together to advance the development of Tibetan information technology. ◄

Sir Frederick Williamson

Tibetological Collections & Archives Series

Tashi Tsering, is the seventh contribution to the Tibetological Collections & Archives Series, which is devoted to important projects on cataloguing, 'computerization' (inputting and scanning), editing, and translation of important Tibetan language text-collections and archives. In this series various colleagues briefly present their initiatives to a larger public, or update the scholarly world on the progress of their already well-established projects. Some are high-profile projects, of which at least Tibetologists will generally be aware, yet some may also be less well known. Nevertheless, I trust that it will be useful to be informed or updated on all these initiatives and I also hope that the projects presented will profit from the exposure and the response that this coverage will engender. In case you would like to introduce your own (planned) work in the field, please contact the editors of the IIAS Newsletter or the author of this introduction. We should very much like to encourage our contributors to keep us informed on the progress of their projects by regular updates. ◄

Henk Blezer, Research fellow at the Universiteit Leiden E-mail: h.w.a.blezer@let.leidenuniv.nl