ISO/IEC JTC1/SC2/WG2/IRG N 1858

Date: 2012-06-15

ISO/IEC JTC 1/SC 2/WG 2/ Ideographic Rapporteur Group

Source: SAT project (Prof. Masahiro Shimoda)

Contribution Identifier: Expert Contribution

Meeting: IRG#38 @ Gyeongju, R.O.Korea

Title: Request to add Un-encoded Characters in the Taishō Shinshū Daizōkyō (Taishō edited series of Buddhist scriptures)

Keywords: (none)

Status: Input to IRG #38

Short Description:

Proposed Conclusion / Requested Action:

Arguments / Text of Contribution: See attachments.

Request to add Unencoded Characters in the Taishō Shinshū Daizōkyō (Taishō edited series of Buddhist scriptures)

A. Charles Muller[1] and Kiyonori Nagasaki[2]

Abstract:

The Taishō Shinshū Daizōkyō (Taishō edited series of Buddhist scriptures, hereafter, Taishō), which contains the basic texts of the Buddhist canon written mainly in classical Chinese has played a prominent role not only in the field of Buddhist studies, but also in various fields connected to Buddhism since its publication in 1934. With such a major and influential literary corpus containing over 6,000 unencoded characters, one of the leading digitization projects of the Taishō, the SAT Daizōkyō Text Database Committee (SAT) proposes encoding those characters in UCS, based on the results of its research. This document reports the situation of the unencoded characters in the Taishō and explains its significance as a resource for research.

1. Unencoded characters in Taishō Shinshū Daizōkyō, which we would like to encode in UCS.

The Taishō Shinshū Daizōkyō (大正新脩大藏經, Taishō edited series of the Buddhist canon, hereafter, Taishō is a collection of Buddhist scriptures consisting of 85 volumes, approximately 120,000,000 characters in total and 27,600 distinct characters, the largest portion of which are Han characters.

There are 6,444 unencoded characters in the texts of the Taishō digitized by the SAT Daizōkyō Text Database Committee (SAT) led by Professor Masahiro Shimoda of the University of Tokyo. This number is arrived at as the result of several comparative analyses with existing CJK Unified Ideographs in UCS up to CJK Unified Ideographs Extension D, but it may decrease slightly in the final comparative effort.

---

[1] Member of SAT / Professor, Faculty of Letters, University of Tokyo

[2] Member of SAT / Senior fellow, International Institute for Digital Humanities / Project Associate Professor, Graduate School of Interdisciplinary Information Studies, the University of Tokyo

Among the 6,444 characters, 1,024 characters occur in two or more individual texts, with the rest occurring only within single texts. Furthermore, The Taishō consists of three types of collections: the first is the collection of scriptures written in India and translated into classical Chinese. The second group is of those written in China/Korea, and the third contains texts composed in Japan. Each group is comprised of: (1) (India): 1,753 distinct characters, (2) (China/Korea): 3,995 distinct characters, (3) (Japan): 696 distinct characters.

2. Current Status of the Taishō

Taishō was critically edited from 1924 to 1934. The base corpus of this canon, the Daizōkyō (大藏經; Skt. Tripiṭaka) took form over more than a millennium, starting in the fourth century. In the earliest stages, it was reproduced by hand copying. From the tenth century, it began to be reproduced by woodblock printing in China, Korea, and Japan. After Japan began to import western technologies in the late 19th century, attempts were first made to print it by metal type and engage in critical editing modeled on Biblical studies. Under these new conditions, several forward-thinking Buddhist researchers began to publish the Taishō, comparing their editions with numerous manuscripts and several authoritative series of xylographs while availing themselves to the latest research in the field of Buddhist studies. At that time, many variant characters were unified (or "normalized") by the determination of scholarly researchers. After that, the Taishō became the primary textual resource for Buddhist studies, being distributed to libraries around the world.

Entering the digital age, Taishō has been fully developed for digital usage. The project of its digitization was initiated in 1994 by the SAT. In the early stages of the project, its contributors input the text manually. The full set of texts was published in 2007 and opened on the Web in 2008. During the course of the project, the treatment of unencoded characters was one of the major issues that faced project members. This character issue has been resolved provisionally by using glyph images, which are sufficient for human readability, but not machine readability and distribution. SAT initially planned to release a font using the private use area (PUA), but gave up the idea, since use of PUA seemed to cause a lot of troubles.

The digitized Taishō has come to be used not only as a resource unto itself, but also as a hub for

linkage and interoperation of a wide-ranging matrix of research resources. For example, SAT has released a Web service integrating a digital dictionary, a parallel corpus and several major bibliographical databases with the Taishō at the center. However, when any interoperating partner attempts to achieve deeper integration, his/her activities are often obstructed by the presence of unencoded characters. It is an urgent matter that needs to be resolved as soon as possible.

3. Significance of Taishō: Printing by Metal Type

According to one traditional narrative of Buddhism, the first Tripi□aka was transcribed in the first century BCE in Pāli language — a kind of Middle Indic. Translation of original Indic scriptures into Chinese started around the second century and scriptures were transcribed individually after that.

A remarkable upsurge in the formation of Chinese Buddhist canonical collections through cataloguing ensued from the fourth century. Dao'an 道安 edited the *Zonglizhongjing mulu* 綜理衆經目錄 (Comprehensive Catalog of Scriptures; not extant). Afterwards, in the Tang dynasty (730), Zhisheng 智昇 compiled a large catalog called the *Kaiyuan shijiao lu* 開元釋教錄 (Taishō 2154.55 "Record of Śākyamuni's Teachings Compiled During the Kaiyuan Period;" usually abbreviated as *Kaiyuan lu* 開元錄) which included 1076 texts written in 5048 fascicles, kept in 480 boxes.

Following the age of hand copying, woodcut printing began in the late Tang.[1] The technology was applied to a series of Buddhist scriptures in 972 at Chengdu 成都 by order of an emperor of the Beisong 北宋 dynasty. This is the so-called *Kaibao Zang* 開宝藏, completed in 977 in 5048 fascicles according to the catalog of the *Kaiyuan lu.* This was the first Chinese version of the canon printed by xylographs, and it included over 130,000 woodcut blocks. This version of the canon was used as the source for the *Goryeo Daejanggyeong Edition* 高麗大藏經 (Korean canon, usually abbreviated as *Goryeo Edition*) around 1011, which was in turn adopted as the base text for the Taishō. The first *Goryeo Edition* was burned by the Mongols in 1087. The existing xylographs were carved again in 1251, creating the so-called second Goryeo Edition.

Regarding the Chinese version of the canon, there were other two genealogical families differentiated by their style of printing. One was the *Qidan Zang* 契丹藏 (Khitan Canon) which seems to have root in manuscripts that were disseminated around Chang'an. Another was a group of several series of Buddhist scriptures that were disseminated in Jiangnan area. Each reflects its own culture.

Entering the Ming dynasty, woodcut printings of the Chinese canon proliferated. One of the best known was the emperor's edition called the *Nan Zang* 南藏, which was printed in the private sector and distributed among the people who wanted to buy their own copies of the canon, and the edition spread throughout all areas of China.

In Japan, editions of the Chinese Buddhist canon were imported from China and Korea since the Middle ages. Due to the vast size and variations within the canon, it was difficult to publish it in Japan, but the Tenkai Edition 天海版 was published from 1648 to 1684. This project printed using wood types whose typeface followed glyphs of an edition published in the Jiangnan area. As the project didn't print so many copies, extant versions are only found in a number of traditional temples. In 1671 Tetsugen Dōkō 鉄眼道光 started a project to make an edition of the canon by woodcut printing which is called the Ōbaku Edition 黄檗版大藏経 which was dedicated to the current Japanese emperor in 1678. Finally, the edition was completed in 1681. While the *Tenkai Edition* was supported by the Japanese government, Tetsugen gathered contributions from the public as well. As the Ōbaku Edition was relatively cheaper, the Buddhist canon spread more widely in Japan.

Due to the popularization of the Daizōkyō by woodcut printing, the processes for the normalization of the texts were enhanced and accelerated. Thus, a condition resulted wherein most people were reading the same texts. It is not possible to study the relationship between medium and text without taking into account typography which was developed by Gutenberg. This enabled the publication of mass printed materials by use of metal types of alphabetic characters. As is well known, it brought about the "Gutenberg revolution" which brought about many changes, such as the flood of books in the 16th century and necessitated discussion on cataloging of the books. Moreover, it normalized various ways regarding enabling pluralistic referencing, such as page numbers, footnotes, uniform glyphs, and alphabetically-ordered indexes, which were not used in any standard way during the age of manuscripts. And then, as Marshall McLuhan pointed out, it brought about a revolution in experience, ethos, and expression of human beings.

Regarding the typography of the Buddhist canon, the Pāli Text Society started addressing the printing of a series of Pāli Buddhist canon starting in 1881 in London. In Burma (Myanmar), the 6th Burmese edition was published based on the 6th meeting of the compilation of scriptures held from

1954 to 1965. Regarding the classical Chinese tripiṭaka, The Dai Nippon Edition 大日本校訂大蔵経 was published using metal type from 1881 to 1885 in Japan. After publishing several editions of the daizōkyō, Dr. Junjirō Takakusu 高楠順次郎 and Kaikyoku Watanabe 渡辺海旭 started publishing the Taishō, critically editing it via comparison with earlier versions of the Chinese canon. As editions of the Taishō were printed in large numbers, the Daizōkyō was rapidly disseminated, especially to libraries all over the world. Thus we have arrived to today's state of affairs where almost everyone in the field of Buddhist studies working in Sinitic sources is reading the same text.

Regarding the printing of the classical Chinese Daizōkyō, unification of characters became one of the important issues. While it was not a matter of serious concern in the case of hand written manuscripts and woodcut printing, printing by metal type encouraged the practice of the repetitive use of the same characters. In the editing of the Taishō, the table of variant characters was revised three times. The end result is that there are about 27,600 distinct characters included in the Taishō which consists of about 12,000,000 characters in total. It also reflects the editors' intention to follow the realistic necessity as far as possible. Now, we are in the midst of the transition to the next phase, going from the technology of typography to the digital. It is our important task to inherit this cultural heritage, which has come down in Asia in this digital age.

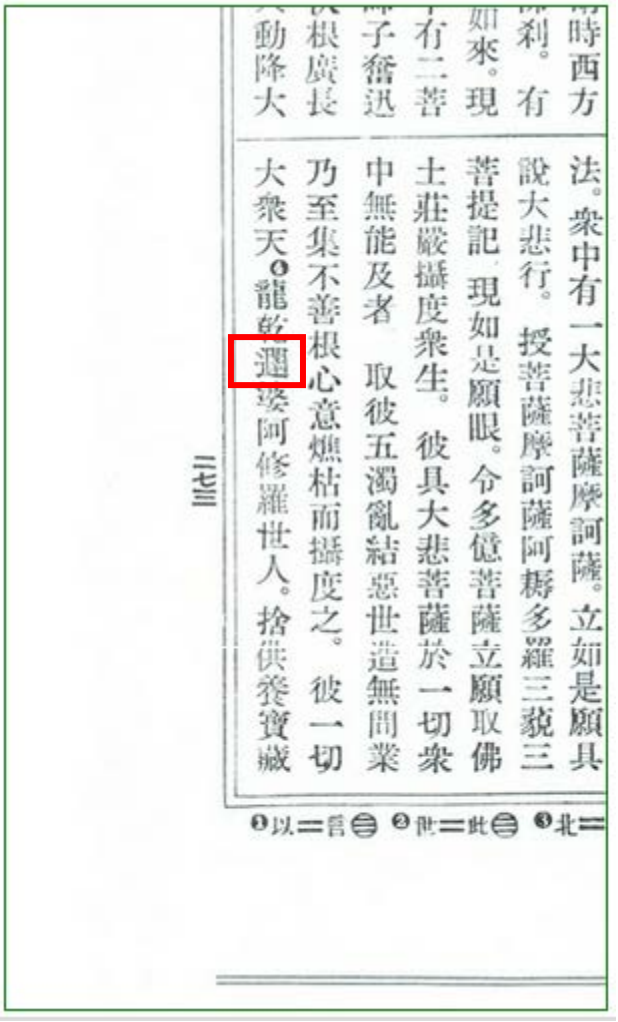
---

[1] A project of woodcut printing of Confucian classics was started in 932.

Examples of unencoded characters in the Taishō Shinshū Daizōkyō

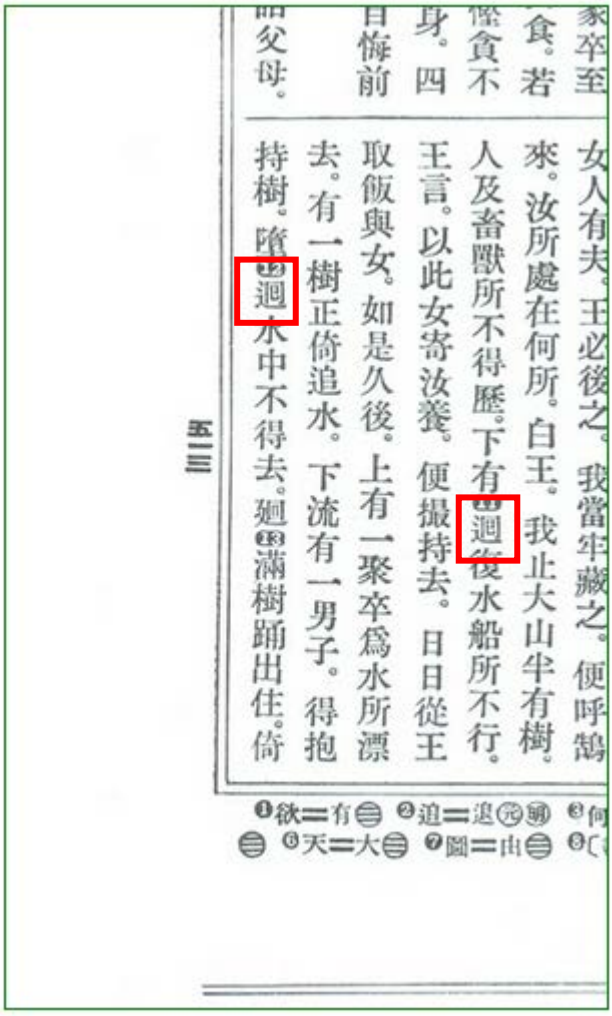
SAT00002

闡	&MT00002;	{しんにゅう2+(門<(土/羊))} -- IDS: ⿺辶⿴門羊				
		162	しんにゅう2	21	25	タツ



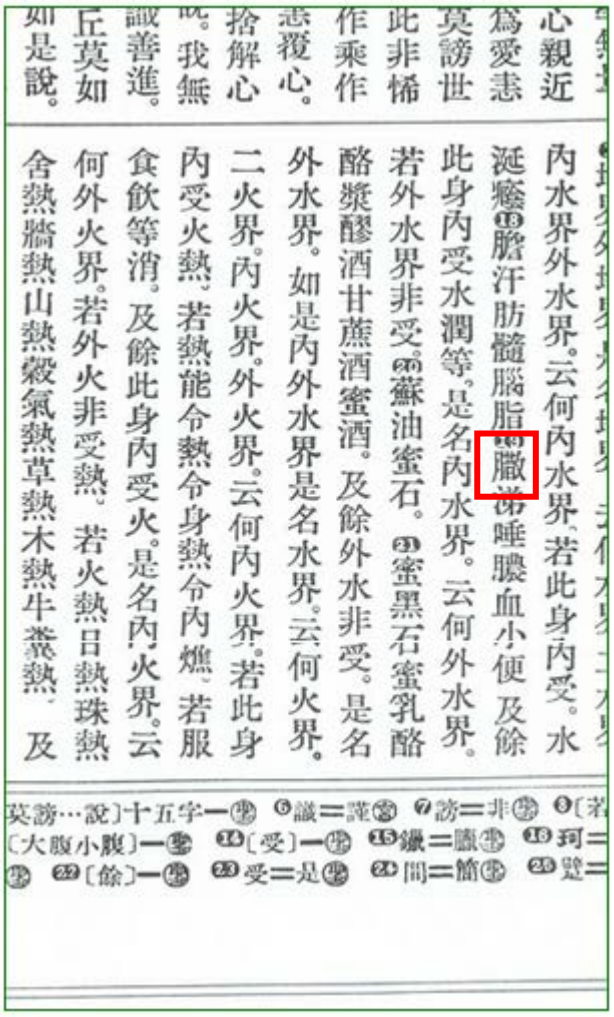
SAT00003

迴	&MT00003;	{(辿-中)+回} -- IDS: ⿺辶回				
		162	しんにゅう	5	9	カイ・エ



SAT00007

膸散	&MT00007;	{月+昔+攵} -- IDS: ⿰月散				
		130	にくづき	12	16	サン



SAT00009

膅	&MT00009;	{月+((庸-用)×巾)} -- IDS: ⿰月⿸广巿				
		130	にくづき	9	13	ヨウ


食唾子。以是之作是説。欲壞聖是説。若日食唾形。金縷彼時阿調達亦世尊以

云何。漸漸膅⑮髀故曰鹿*蹲腸。七合滿盈其義云何。⑯肉脈平正鉤⑰鎖骨。七處滿足平住色不變移。是謂七合滿盈。師子臆其義云何。身無高下不前却。是謂師子臆。味味知者其義云何。輭軟微妙皆悉能知。是⑱爲味味曉了

不養賢恐懼　慈及諸所趣
無想有想唾　曩昔云何相

以何等故。鳥畜生昔日皆能語。今不能語。或作是説。爾時從人中終生畜生中。以前所習

⑧衆+(僧)⑨仁=然 ⑩已=以
⑯肉=内 ⑰鎖=瑣 ⑱爲=

SAT00600

	&MT00600;	{さんずい+(丿+友+丶)}、{(池-也)+(ノ/(友+丶))} -- IDS: ⿰氵⿸丿犮				
		85	さんずい	6	9	


粗相問訊。在一面坐。即白佛言。汝實不欲利
益安樂諸衆生耶。豈不常讃如是法乎。佛告
之曰。我於長夜。常欲利益一切衆生。亦恒讃
嘆如是之法。閉口姓言。若如是者。何不爲諸
衆生等同說法有不說者。佛告之曰。我今問
汝隨所樂答。譬如世人。有三種田。有一上田
良美❷&MT00600;壞。極爲上好。第二田者。適處其中
好。第三田者。曠野邊遠。沙鹵鹹惡。有諸田
夫。先於何田。而下種子。閉口姓言。爲利益
者。先種良田。望獲大利。佛告之言。若良田
盡。次種何田。閉口姓言。次種中者。種中田
已。次種下田。亦復擲子。亦望愛時少有所獲。

等。於長
姓。聞佛
喩。佛告
有三種
無❹滲漏
有*滲漏
先何器。
二器。其
告言。滿
破。亦應
瓮。喩我

Statement to Support “UCS Encoding of the External Characters in the Taishō Shinshū Daizōkyō”

Concerned members of various international academic associations
related to Buddhist Studies, Asian Studies and Digital Humanities

Over the course of its 2,500 year history, Buddhism has been contributing to the spiritual culture of human beings as the major world religion based in the regions of Asia. Particularly, the “Daizōkyō,” consisting of volumes of the major Buddhist canon translated in ancient China has become the basis in the creation of spiritual culture of human beings for many years as a treasure of human wisdom that successfully merged two great civilizations in India and China. Each kanji in this canon contains an enormous background in the history of civilization.

The ”Taishō Shinshū Daizōkyō（大正新脩大藏經）,” compiled drawing upon the wisdom of Japan at the beginning of the 20th Century, introduced this unique blend of thought from the East to the Western world, and can be recognized as a historical accomplishment that realized the encounter of Eastern and Western civilizations in human cultural history. Since then, the truth of this world religion has been interpreted both in the East and West based on the ”Taishō Shinshū Daizōkyō,” and has continued to be expanded as a cultural property for the next generation.

At the end of the 20th Century, the method to store and dispatch human wisdom changed dramatically due to the appearance and dissemination of digital media, and it became an important task to share the intellectual heritage of human beings with the new media in the world. Under these circumstances, the “SAT Daizōkyō Text Database Committee” successfully created a text database of the overall ”Taishō Shinshū Daizōkyō” in accordance with today’s highest possible academic standards and at the same time completed detailed and careful research of non-standard characters. To realize UCS encoding of these external characters is an essential factor to inherit the overall history with such profound background to the future. As interested members of various academic associations relating to Buddhist Studies, Asian Studies and Digital Humanities, we desire for it to be realized at an early stage; therefore we have come together to submit this statement.

Shayne Clarke

(Professor, Religious Studies, McMaster University, Canada)

Max Deeg

(Professor, Religious Studies, Cardiff University, UK)

Qing Duan

(Professor, Beijing University, China)

Paul Groner

(Professor, Religious Studies, University of Virginia, USA)

Zhaoguang Ge

(Dean of Historical Research, Fudan University, China)

Paul Harrison

(Professor, Ancient Indian and Chinese Buddhism, Stanford University, USA)

Jens-Uve Hartmann

(Professor, Indology and Tibetology, University of Munich, Germany)

Soonil Hwang

(Professor, Buddhist Studies, Dongguk University, Korea)

A. Charles Muller

(Professor, Center for Evolving Humanities, University of Tokyo, Japan)

Bethany Nowviskie

(Director, Digital Research & Scholarship, Library of the University of Virginia, USA, President of the Association for Computers and Humanities)

Lisa Lena Opas-Hänninen

(Professor, University of Oulu, Finland, Chair of the Association for Linguistic and Literary Computing)

Juyhung Rhi

(Professor, Seoul National University, Korea)

Jan-Noel Robert

(Professor, Philology, College de France, France)

Masahiro Shimoda

(Professor, Indian and Buddhist Studies, University of Tokyo, Japan, Chair of Japanese Association for Digital Humanities)

Raymond Siemens

(Professor, University of Victoria, Canada, Chair of the Alliance of Digital Humanities Organizations)

Additional Character Proposal for the Taishō Daizōkyō Database
IRG N1858 Appendix
Charles Muller
Kiyonori Nagasaki
Masahiro Shimoda

The SAT Taishō Daizōkyō database project, based at the University of Tokyo, is proposing the addition of 6,000+ ideographs to the ISO 10646 character set. These ideographs are contained in the standard East Asian (CJK) version of the Buddhist canon (*canon* = "collection of scriptures authorized by the tradition" 藏). This canon took form over a period of 1700 years in China, Korea, and Japan, being edited into its final form in Tokyo during the Taishō era by a team of Japanese specialists. It is this version of the canon that is uniformly used as the standard for reference and citation by scholars, not only in China, Korea, and Japan, but throughout the world. The digitized version of this "Taishō" Buddhist canon presently serves as the hub of the SAT online database (http://21dzk.l.u-tokyo.ac.jp/SAT/ddb-sat2.php), where it is interlinked and is interoperative with a growing range of other related text databases, bibliographies, and lexicons.

The new ideographs that we propose to include are in no way particularly arcane, ancient, or in any other way odd. Most importantly, they are used daily in the research being conducted by scholars around the world who are relying on our database. This means that the number users who read these ideographs daily is quite large--probably around 5,000 or so--distributed around the world.

Although the final edition of this canon was developed in Japan, since, as noted above, the historical development of the canon occurred through the cultures of China and Korea (e.g. the Korean version of the canon 고려대장경 高麗大藏經) as well as Japan, it is quite conceivable that scholars working in this area in these other regions might be interested in voicing their opinions and offering input in the process. We warmly look forward to receiving such input.