Case study: Typesetting old documents of Japan

NAKANO Ken and KOBAYASHI Hajime

This is a report on the typesetting of *Komonjo books* (reprint books compiling old documents of Japan) published by Shiryo Hensan-jo, with pTEX (pLATEX).

We would like to report that typesetting of *Komonjo books*, which are classified as difficult to typeset, can be made efficiently using TEX.

About Shiryo Hensan-jo

From their web site (`http://www.hi.u-tokyo.ac.jp`): Shiryo Hensan-jo (the Historiographical Institute, HI), the University of Tokyo, has as its primary objective, rather than historiography in general, analysis, compilation, and publication of historical source materials concerning Japan.

The institute has become a major center of Japanese historical research, and makes historical sources available through its library, publications, and recently, databases.

About us and TEX

Our company, Livretech, is a printing firm mainly producing textbooks and education-related books.

We've been using TEX since 1988, and have been concerned with some TEX-related topics. We developed a DVI driver program for the Shaken typesetter in 1989 (Shaken is a famous manufacturer in Japan making fonts and typesetting systems). We also typeset and output Japanese editions of *The TEXbook* (1992) and *The METAFONTbook* (1994).

1 Until TEX was chosen

Before TEX was chosen, *Komonjo books* were composed with hot metal types or phototypesetting systems developed for professional users. Several typesetting systems had been used, but commonly speaking, *Komonjo books* were difficult to typeset.

After typesetting was completed, contents of books were stored in the database named SHIPS (Shiryo hensanjo Historical Information Processing System), but another problem occurred.

In those days, Japanese typesetting systems worked with the creators' original character codes. For technical reasons and/or the creators' policies, typesetting data could not easily be re-used for another purpose. So, the final typesetting data would not automatically convert into database texts. To make the database text, typesetting contractor had to revise the final data manually, or in the worst case, input all texts again. And texts made in this way had to be proofread again.

To solve these problems, the professors of Shiryo Hensan-jo focused on TEX. This was in about 2001. They tried to typeset *Komonjo books* with TEX, and they found the possibility of typesetting and flexibility of text data.

And then, they contacted us and we started on this new challenge in 2002.

2 Typesetting problems

2.1 Problems with fonts

2.1.1 Shortage of kanji characters

Massive numbers of kanji are used in *Komonjo books*. At that time, the Shaken system that was used to typeset *Komonjo books* had about 10,000 kanji characters. However, the fonts which could be used with pTEX had only about 6,500 kanji characters.

2.1.2 Simplification of kanji shapes

In Japan, shapes of many frequently used kanji have been simplified to make them easier to read and write. Kanji of traditional shapes are called *Seiji* (正字), and those of simplified shapes are called *Ryakuji* (略字). Here are some examples of *Ryakuji* and *Seiji* character shapes:

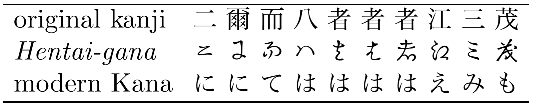
Ryakuji	*Seiji*	meaning
医	醫	medical, doctor
円	圓	circle, yen
塩	鹽	salt
害	害	harm, damage
学	學	study, learn
国	國圀	country

Like the letter "国," some *Ryakuji* characters have two or more *Seiji* characters.

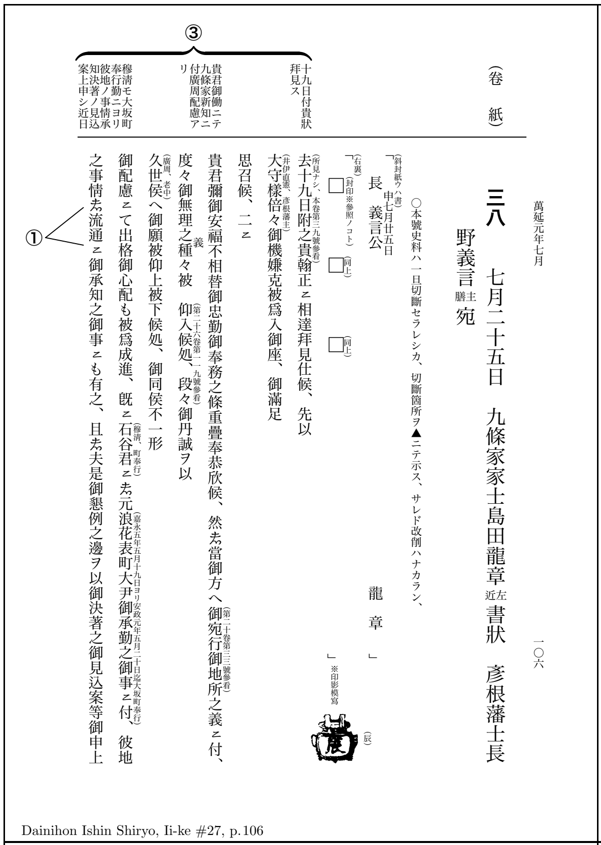
In essence, *Komonjo books* are composed with *Seiji* characters. On the other hand, we use *Ryakuji* kanji in everyday life, and we rarely use *Seiji* characters. Because of this, *Seiji* characters have been put away in the dark recesses of fonts and kanji input methods. So it is difficult for us to input *Seiji* characters directly.

2.1.3 *Hentai-gana*

変体仮名 (*Hentai-gana*, see ① in the figure on the next page) are Japanese old syllabary characters. *Hentai-gana* were made from kanji characters as well as modern *kana* characters. Here we show some *Hentai-gana* examples used in *Komonjo books*, with related kanji and kana characters.



original kanji	二	爾	而	八	者	者	者	江	三	茂
Hentai-gana										
modern Kana	に	に	て	は	は	は	は	え	み	も



萬延元年七月

三八　七月二十五日　九條家家士島田龍章左近書狀　彦根藩士長野義言主膳宛

一〇六

○本號史料ハ一旦切斷セラレシカ、切斷箇所ヲ▲ニテ示ス、サレド改刪ハナカラン、

（斜封紙ウハ書）
「申七月廿五日
長　義言公　　龍　章」

（右裏）
「□（封印※參照ノコト）　□（同上）　□（同上）」　※印影模寫　（辰）

（巻　紙）

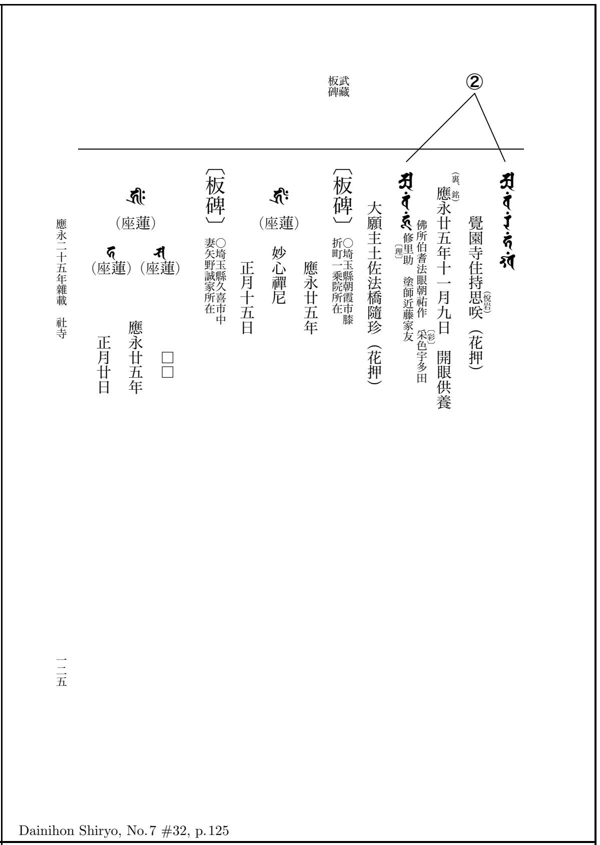
十九日付貴狀拜見ス

貴君御働ニテ九條家新知ニ付廣周配慮アリ

穆清モ大坂町奉行ニヨリ彼地ノ事情承知決著ノ見込案上申シ近日

去十九日附之貴翰（所見ナシ、本巻第三九號參看）正ニ相達拜見仕候、先以
大守様（井伊直憲、彦根藩主）倍々御機嫌克被爲入御座、御満足
思召候、二ニ
貴君彌御安福不相替御忠勤御奉務之條重疊奉恭欣候、然者當御方へ御宛行（第二十巻第三三號參看）御地所之義ニ付、
度々御無理之種々被（義）仰入候処（第二十六巻第一一九號參看）、段々御丹誠ヲ以
久世侯（廣周、老中）へ御願被仰上被下候処、御同侯不一形
御配慮ニて出格御心配も被爲成進、既ニ石谷君（穆清、町奉行）ニ者元浪花表町大尹（嘉永五年五月十九日ヨリ安政元年五月二十日迄大坂町奉行）御承勤之御事ニ付、彼地
之事情者流通ニ御承知之御事ニも有之、且者夫是御懇例之邊ヲ以御決著之御見込案等御申上

Dainihon Ishin Shiryo, Ii-ke #27, p. 106



武藏板碑

覺園寺住持思咲（悦岩）（花押）
（裏、銘）應永廿五年十一月九日　開眼供養
佛所伯耆法眼朝祐作　〔彩〕采色宇多田
修〔理〕里助　塗師近藤家友
大願主土佐法橋隨珍（花押）

〔板碑〕○埼玉縣朝霞市膝折町一乘院所在

應永廿五年
（座蓮）
妙心禪尼
正月十五日

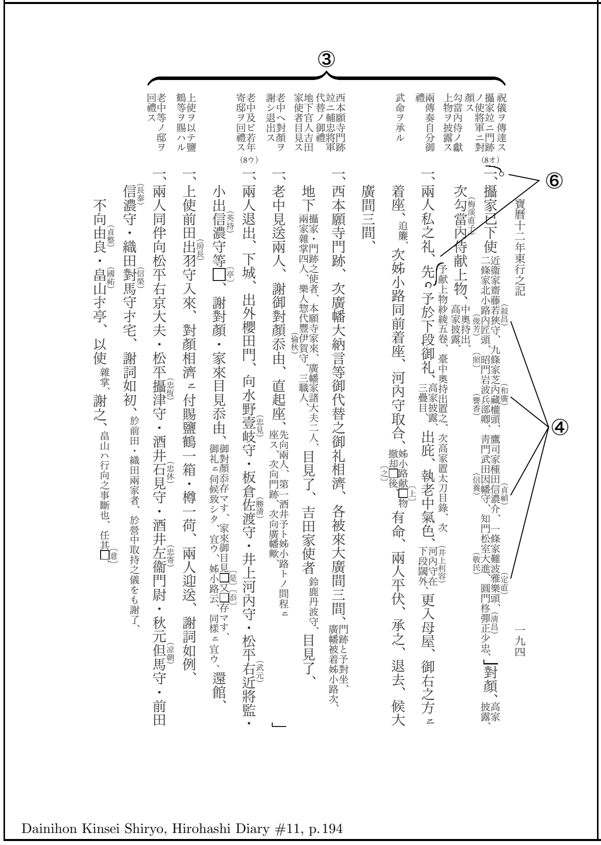
〔板碑〕○埼玉縣久喜市中妻矢野誠家所在

（座蓮）
（座蓮）（座蓮）□□
應永廿五年
正月廿日

應永二十五年雜載　社寺

一一五

Dainihon Shiryo, No. 7 #32, p. 125



寶曆十二年東行之記

一九四

祝儀ヲ傳達ス

攝家竝ニ門跡ノ使將軍ニ對顔ス

勾當内侍ノ獻上物ヲ披露ス

兩傳奏自分御禮

武命ヲ承ル

西本願寺門跡竝ニ輔忠將軍代替ノ御禮

地下官人吉田家使者目見ス

老中へ對顔ヲ謝シ退出ス

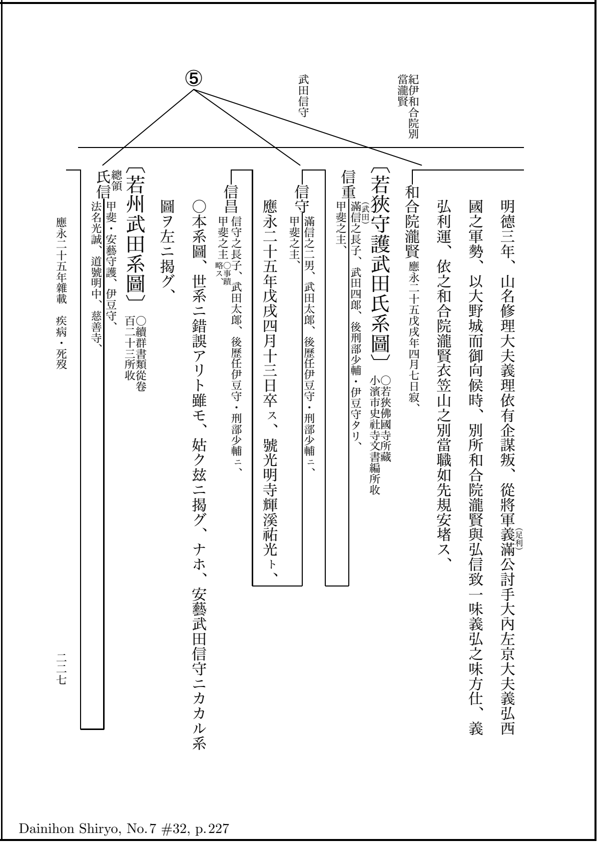
老中及ビ若年寄邸ヲ回禮ス

上使ヲ以テ鹽鶴等ヲ賜ハル

老中等ノ邸ヲ回禮ス

（8オ）
一、攝家已下使（近衞家齋藤若狹守（敍昌）、九條家芝内藏權頭（和廣）、鷹司家種田信濃介（貞禎）、一條家難波雅樂頭（定直）、二條家北小路内匠頭（俊芳）、昭（照）門岩波兵部卿（譽香）、靑門武田因幡守（信養）、知門松室大進（敬民）、圓門（清昌）柊彈正少忠、）」對顔、（高家披露、）
次勾當内侍（梅溪直子）獻上物、（中奧持出、高家披露、）
一、兩人私之禮、先〇予於下段御禮、（予獻上物紗綾五巻、臺中奧持出置之、高家披露、三疊目、）出庇、執老中氣色、（次高家置太刀目錄、次河内守（井上利容）在下段鬮外、）更入母屋、御右之方ニ
着座、（追簾、）次姉小路同前着座、河内守取合、（姉小路獻□（上）物、撤却□（之）後、）有命、兩人平伏、承之、退去、候大
廣間三間、
一、西本願寺門跡、次廣幡大納言等御代替之御禮相濟、各被來大廣間三間、（門跡と予對坐、廣幡被着姉小路次、）
地下（攝家・門跡之使者、本願寺家來、廣幡家諸大夫二人、兩家雜掌四人、樂人惣代豐伊賀守（倫秋）、三職人、）目見了、吉田家使者（鈴鹿丹波守、）目見了、
一、老中見送兩人、謝御對顔忝由、直起座、（先向兩人、第一酒井予ト姉小路トノ間程ニ座ス、次向門跡、次向廣幡歟、）
（8ウ）
一、兩人退出、下城、出外櫻田門、向水野壹岐守（忠見）・板倉佐渡守（勝清）・井上河内守・松平右近將監（武元）・」
小出信濃守（英持）等□（亭）、謝對顔・家來目見忝由、（御對顔忝存マす、家來御目見□（是）又□（忝）存マす、御礼ニ伺候致シタ、宜ウ、姉小路云、同樣ニ宜ウ、）還館、
一、上使前田出羽守（房長）入來、對顔相濟ニ付賜鹽鶴一箱・樽一荷、兩人迎送、謝詞如例、
一、兩人同伴向松平右京大夫・松平攝津守（忠恒）・酒井石見守（忠休）・酒井左衞門尉（忠寄）・秋元但馬守（凉朝）・前田
信濃守（長泰）・織田對馬守（信榮）ホ宅、謝詞如初、（於前田・織田兩家者、於營中取持之儀をも謝了、）
不向由良（貞整）・畠山（國祐）ホ亭、以使（雜掌、）謝之、（畠山ハ行向之事斷也、任其□（意）、）

Dainihon Kinsei Shiryo, Hirohashi Diary #11, p. 194



紀伊和合院別當瀧賢

武田信守

明德三年、山名修理大夫義理依有企謀叛、從將軍義滿（足利）公討手大内左京大夫義弘西國之軍勢、以大野城而御向候時、別所和合院瀧賢與弘信致一味義弘之味方仕、義弘利運、依之和合院瀧賢衣笠山之別當職如先規安堵ス、

和合院瀧賢　應永二十五戊戌年四月七日寂、

〔若狹守護武田氏系圖〕○若狹佛國寺所藏　小濱市史社寺文書編所收

信重　（武田）滿信之長子、武田四郎、後刑部少輔・伊豆守タリ、甲斐之主、

信守　滿信之二男、武田太郎、後歷任伊豆守・刑部少輔ニ、甲斐之主、
應永二十五年戊戌四月十三日卒ス、號光明寺輝溪祐光ト、

信昌　信守之長子、武田太郎、後歷任伊豆守・刑部少輔ニ、甲斐之主、○事蹟略ス、

○本系圖、世系ニ錯誤アリト雖モ、姑ク玆ニ掲グ、ナホ、安藝武田信守ニカカル系
圖ヲ左ニ掲グ、

〔若州武田系圖〕○續群書類從卷百二十三所收

總領
氏信　甲斐・安藝守護、伊豆守、法名光誠、道號明中、慈善寺、

應永二十五年雜載　疾病・死歿

二二七

Dainihon Shiryo, No. 7 #32, p. 227

Examples of *Komonjo books* (reduced 53%).

Hentai-gana have different shapes with the same reading, while modern kana have one shape with one reading.

2.1.4 *Bonji*

梵字 (*Bonji*, see ②) characters are mystical letters used in Japanese Buddhism. *Bonji* consist of consonants, vowels and derivative letters, so many characters can be created. The following are some typical *Bonji* characters and their meanings:

Bonji	sound	meaning
𑖀	a, ア	the most basic character
𑖁𑖾	āḥ, アーク	Dainichi Buddha, 大日如来
𑖮𑖿𑖨𑖱𑖾	hrīḥ, キリク	Amitabha, 阿弥陀如来

2.2 Problems with notes

Komonjo books contain many notes to aid readers' understanding. Two kinds of notes occur frequently.

One is a "head note" (③), similar to `\marginpar` but a long note appearing at the end of a page has to be split over pages.

The other is an "interline note" (④) which can occur anywhere in the text. To avoid overlapping of the interline note and text or other materials, position adjustment is needed.

2.3 Problems with picture-like materials

Picture-like materials such as family tree diagrams (⑤) and handwriting-like lines (⑥) interrupt the flow of text. Handling them is troublesome in TEX.

3 Solutions to problems

3.1 Solutions for font problems

3.1.1 New kanji fonts are available

The Adobe-Japan1-5 character collection was published in 2002. This collection has 20,316 characters total, including 12,668 kanji characters. The number of kanji characters approximately doubled from the font that we used till then. AJ1-5 fonts could mostly cover the range of *Komonjo books*.

After that, the UTF package was developed by SAITO Shuzaburo. With this package, we could now use all characters of AJ1-5 fonts with pTEX.

In addition, we made an OpenType font named `ut-gaiji` to gather the additional characters which AJ1-5 did not have ("ut" means "u-tokyo" and "gaiji (外字)" means "external character").

3.1.2 Input *Seiji* using *Ryakuji*

To avoid the difficulty of *Seiji* input, we decided to make *Seiji* texts from *Ryakuji* texts. Preparation in advance was as follows:

- Write a style file named `jitai.sty` ("jitai (字体)" denotes "character shape") to describe pairs of *Ryakuji* and *Seiji*, in order to specify *Seiji* kanji via *Ryakuji* kanji.

```
\def\seiji{\@ifnextchar[{\@seiji}{\@seiji[0]}}
\def\@seiji[#1]#2{\expandafter
 \ifx\csname @SJ#1#2 \endcsname\relax
  \typeout{!!! Seiji #2(#1) undef.}#2\relax
 \else \csname @SJ#1#2 \endcsname \fi }
\let\S\seiji
\def\DefSeiji[#1]#2#3{\expandafter
 \gdef\csname @SJ#1#2 \endcsname{#3}}
\DefSeiji[0]{亜}{亞}
\DefSeiji[0]{唖}{\CID{7633}}
\DefSeiji[0]{逢}{\CID{8266}}
\DefSeiji[0]{悪}{惡}
\DefSeiji[0]{医}{醫}
\DefSeiji[0]{学}{學}
\DefSeiji[0]{国}{國}
\DefSeiji[1]{国}{圀}
       :
```

  `jitai.sty` has about 1,000 such pairs. For example, after these definitions, the input `\S{学}` produces "學," and you can input `\S[1]{国}` to get optional *Seiji* character "圀."

  The `\CID` command in a kanji pair specifies the internal character code from AJ1-5 (Character ID) directly.

- Write a Ruby program which converts *Ryakuji* texts into *Seiji* texts, referencing the `jitai` file. When the Ruby program finds *Ryakuji* kanji in a text file, it inserts the *Seiji* command.

  For example, the kanji string "文学青年" ("Literary youth") is converted into "`\S{文}` `\S{学}\S{青}`年," and typeset as "文學靑年."

Then, work steps are as follows:

1. Input text file with *Ryakuji* kanji.
2. Convert *Ryakuji* texts into *Seiji* texts.
3. Typeset *Komonjo books* with *Seiji* texts.

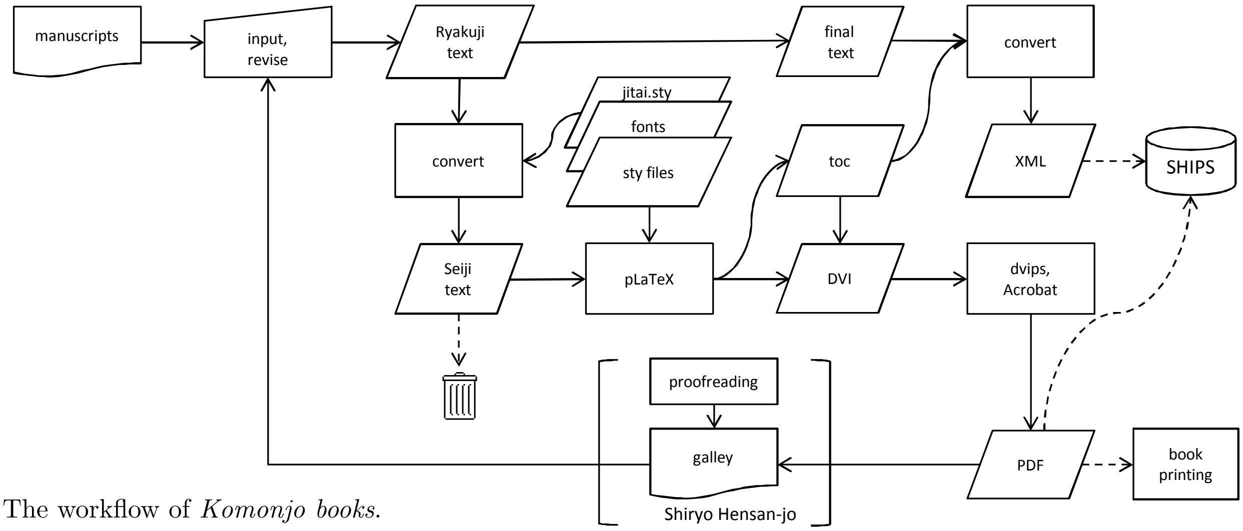
We revise only *Ryakuji* text files. We discard *Seiji* text files after typesetting has completed, since *Seiji* texts are hard to read due to the automatic insertion of *Seiji* commands.

3.1.3 *Hentai-gana*

We made *Hentai-gana* as external characters and put them into `ut-gaiji`.

3.1.4 *Bonji*

First, we used the free *Bonji* font, but it had only typical characters, so we often had to create additional characters. Therefore, now we use the 今昔文字鏡 (Konjaku-Mojikyo) font package which has about 2,000 *Bonji* characters. And we also use the `mojikyo` macro package developed by HONDA Tomoaki, to specify *Bonji* characters.



The workflow of *Komonjo books*.

3.2 Solutions for notes

3.2.1 Customize `\marginpar` for head note

A head note must be split when it "falls off" the end of a page. To implement this, we customized the definition of `\marginpar`. The original maintains a `\@marbox` to store a page of marginal notes. We just `\vsplit` `\@marbox` to `\textheight`, and put the remaining note text at the top of the `\@marbox` of the next page.

3.2.2 Interline note

An interline note is placed at the right or left side of the text. We defined `\rnote` and `\lnote` respectively, and an optional position adjustment, as follows:

`\rnote[`*v-adjust*`][`*h-adjust*`]{`*note text*`}`.

3.3 Picture-like materials

We made the complicated family tree diagram as an integral number of line units, so that it can divide at the end of page.

For handwriting-like lines, we defined a `\Curve` macro, taking three points of a curve. Our `\Curve` is essentially the `\bezier` function with `\unitlength` of 1 zw (= width of a kanji). The second curve of ⑥ (∫) was drawn with 3 lines as follows:

```
\Curve(.4,-.5)(.6,-.5)(.6,.2)
\Curve(.6,.2)(.6,.5)(.6,.5)
\Curve(.6,.5)(.6,1.1)(1.1,1.1)
```

3.4 Generate database texts

Another purpose of typesetting with TeX was to generate database texts from final TeX files. We wrote a Ruby program for this, and it works as follows: delete unnecessary information except text, unify *Seiji* in *Ryakuji*, convert character code from Shift-JIS to Unicode, and convert from TeX file to the XML format of the database.

3.5 Result

We show the whole workflow of *Komonjo books* in the figure above. Using our methods, more than half of the *Komonjo books* of Shiryo Hensan-jo have been typeset with TeX now.

As a result, we think that the two purposes to be expected in TeX (efficiency of typesetting and re-using of input for the database texts) have been accomplished.

Acknowledgements

First of all, we'd like to thank the professors of Shiryo Hensan-jo. They focused on TeX and selected us as a partner, and they were kind enough to teach everything about *Komonjo books*. We also give thanks to SAITO Shuzaburo for the development of the AJ1-5 font support packages. And also thanks to HONDA Tomoaki for the development of Mojikyo-related packages. Finally, many thanks to Karl Berry and Barbara Beeton for their kind support in reviewing this article.

About rights

Products of this activity (style files, programs, etc.) belong to Shiryo Hensan-jo. They are not publicly available at this time.

◇ NAKANO Ken
`k-nakano (at) livretech dot co dot jp`

◇ KOBAYASHI Hajime
`koba (at) livretech dot co dot jp`