A Wen-zi-xue (文字學) data base
and
the ITRIX Chinese computing environment

May 8,1991

Ching-chun Hsieh 謝清俊 — research fellow, Inst. of Information Science, Academia Sinica

Jin-ding Sheu 許金定 — research assistant, Computing Center, Academia Sinica

Yang-juh Lai 賴洋助 — C.UNIX System Project Leader, System Software Dept.,CCL, ITRI

A Wen-Zi-Xue (文字學) data base
and
the ITRIX Chinese computing environment

1. Background

Since 1985, a long-term program , called the Computer Applications for the Humanities and the Sociology studies has been started and conducted by the Computing Center of the Academia Sinica. Under this program, many research and development projects have been carried out. These projects include developing data bases [1] ,implementing computing tools [2,3,4,5,6], and also doing some basic computational linguistic research in Chinese language [7,8,9,10]. The database and the related work presented in this paper is also a part of the mentioned program.

The ultimate goal of the program is to develop a computerized base for Sinology studies, especially in the Humanities and the Sociology fields. As a consequence, the ability of processing information in Chinese becomes a critical issue of the program. The work presented in this paper provides a necessary and basic foundation for studying ancient materials as well as improving the system's abilities to handling Chinese language. A concepture functional block diagram of the base is shown in Figure-1.

2. The Wen-zi-xue data base

The Wen-zi-xue (文字學) means the historical studies of the ideography, the phonology, the morphology, the etymology and the semantics of Chinese characters. Therefore, by nature, the Wen-zi-xue database(ZXdb in short) contains mainly a collection of dictionaries from ancient to present. The time span will cover more than 1800 years.

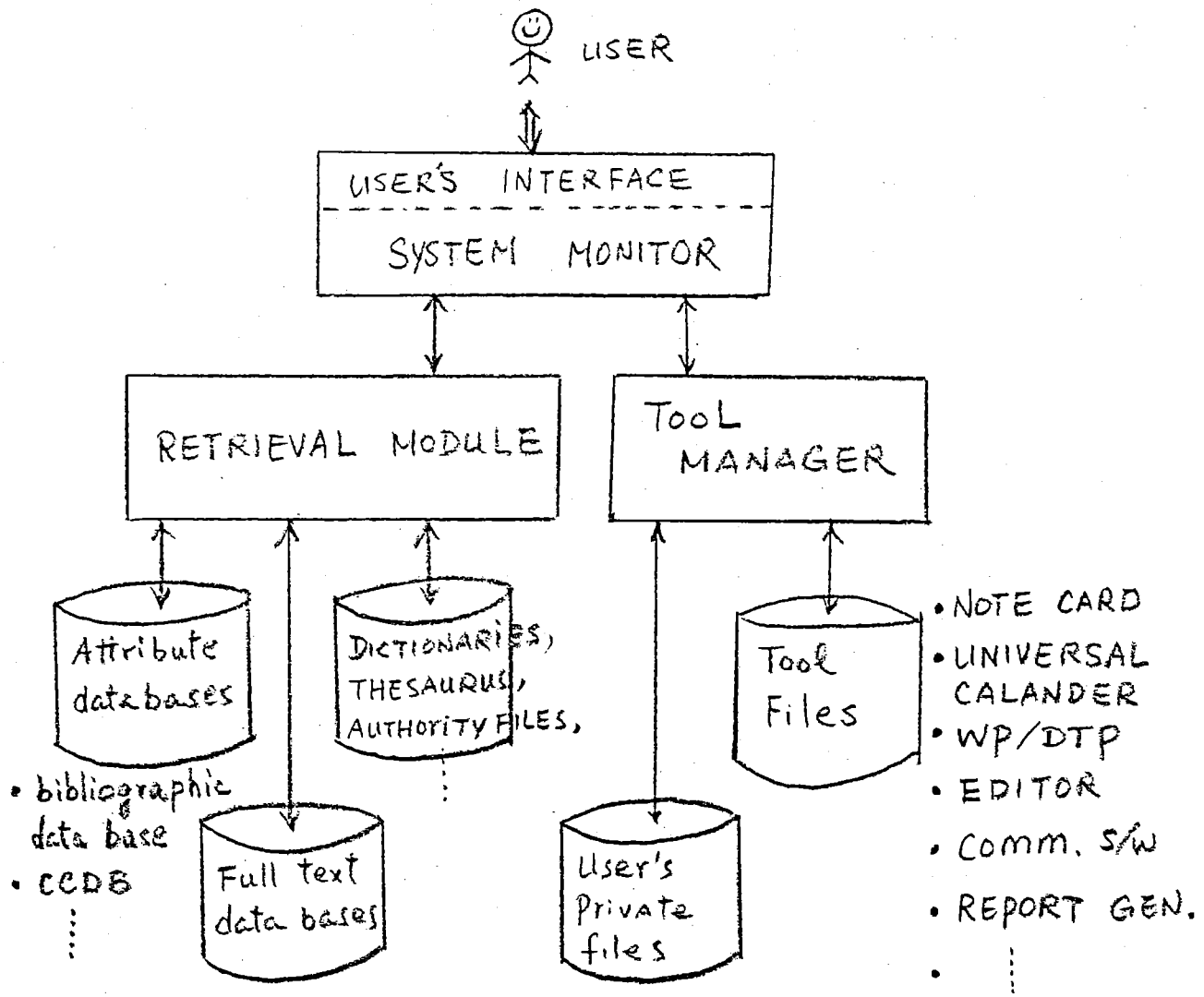


Figure-1, A conceptual functional block diagram of the computing base for Sinology studies.

2.1 Dictionaries in the ZXdb

At present, in the first phase of this project, only two ancient dictionaries were selected. The first one is the<<Shou-wen-jie-zi(說文解字)>> by Mr. Xu,Shen(許慎) [14]. It was first published in 121AD. It was the first well organized, systematic, and theoretically sound and complete dictionary in Chinese history. An example of some entries of the <<Shau-wen-jie-zi>> is shown in Figure-2. There are 9353 entry characters collected, with additional 1753 variants(異體字).

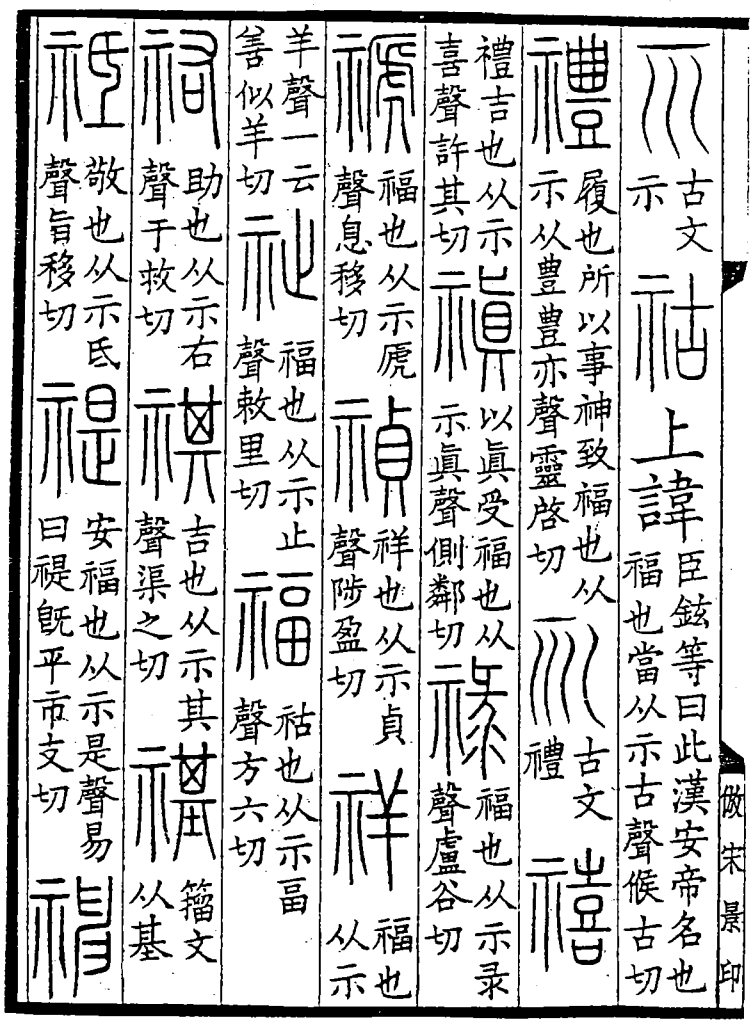
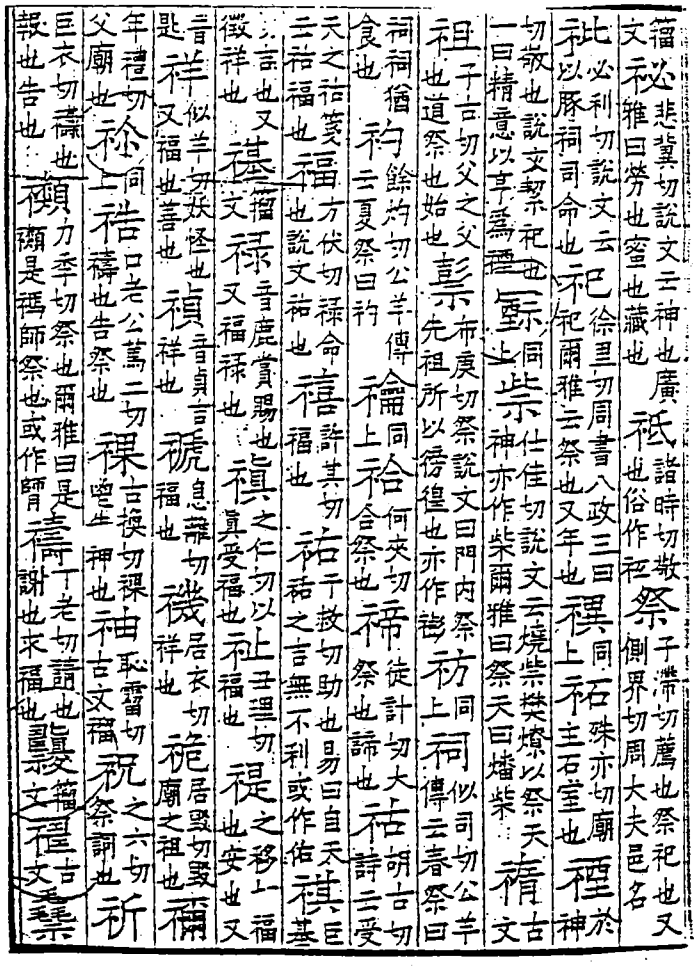


Figure-2. An example of some entries of the <<Shau-wen-jie-zi>>

The second dictionary selected is the <<Yu-Pian>> (玉篇) by Mr. Gu,yie-wang(顧野王)[15]. It was first published in 543 AD. It has 22800 entry characters. An example of the <<Yu-Pian>> is shown in Figuare-3. These two dictionaries are the two most important reference tools for understanding ancient written materials.



Figuare-3. An example of the <<Yu-Pian>>

In these two dictionaries, each entry will begin with a glyph of the selected character, and then followed by its meaning explanation, the originality and structure of that character, the pronunciation in Fan-qie (反切) form, and some examples from ancient books as illustrations. For <<Shau-wen-jie-zi>>, the leading glyph is in ancient Shyau-jwan (小篆) font which is not supplied by any computing systems today. So, we are now using

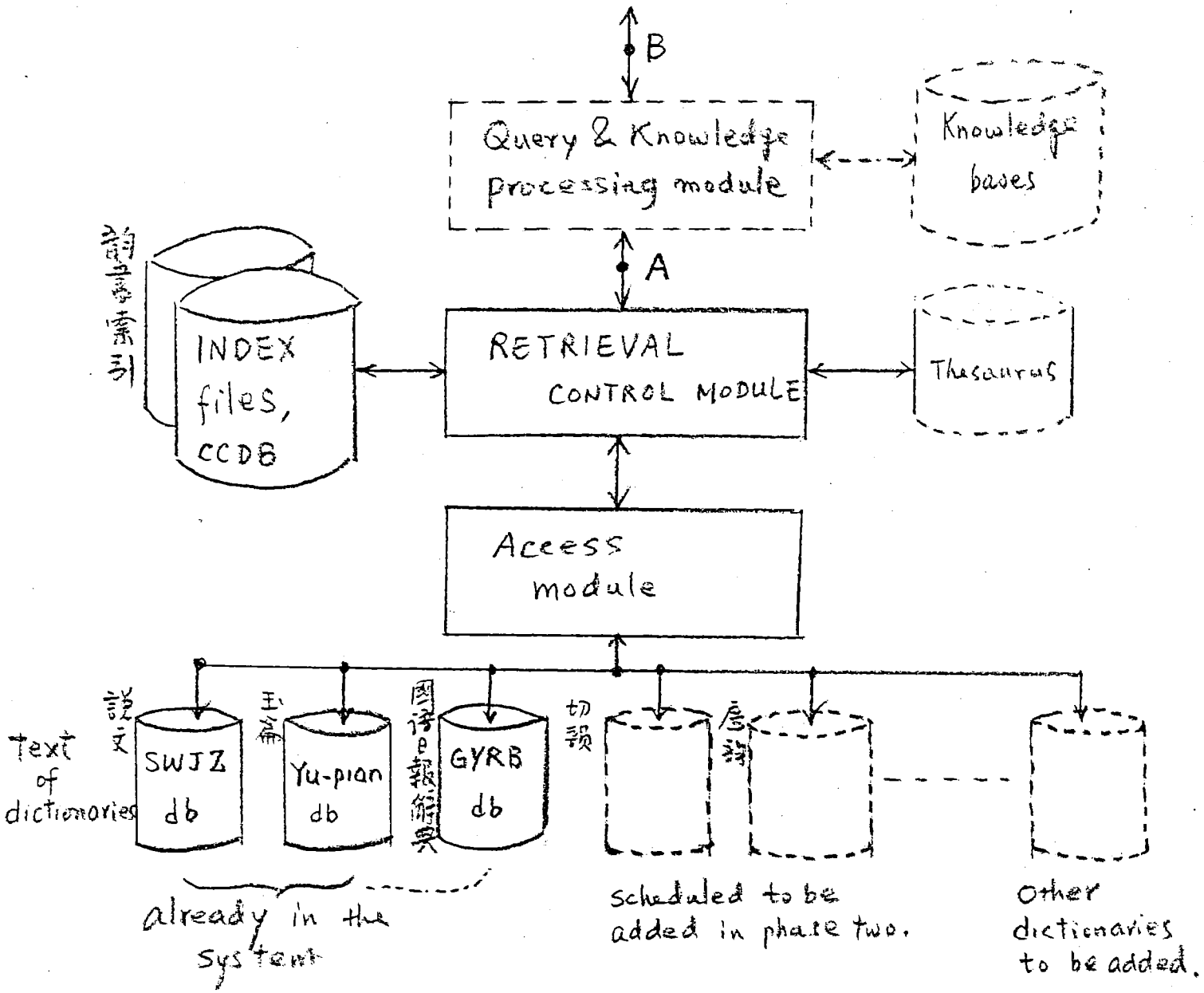


Figure 4. A concepture block diagram of the ZXdb.

Kai (楷) font instead in the ZXdb, and the Shyau-jwan font will be supplied by an image data base later in the second phase.

The content of these two dictionaries covers all the essential information for understanding Chinese character, but not enough for the phonological studies. Therefore, it is planned that the two most important phonological dictionaries will be added to the system in the second phase. They are the <<Qie-yun>> (切韻) in 601AD and <<Guang-jun>> (廣韻) in 1008AD. Also, in order to provide user a modern way to access these ancient dictionaries, we integrated these ancient dictionaries with the attribute data base of Chinese character(CCDB) [11] and a very popular present day dictionary called <<Gwoyeu Ryhbaw Tsyrdean (國語日報辭典)>> [12] by the Mandarin Daily News. By this arrangement any attribute from CCDB and any access entry from <<Gwoyeu Ryhbaw Tsyrdean>> can be used to access the corresponding content in these two ancient dictionaries,and the reverse way of the access is also available.

A block diagram of the ZXdb is shown in Figuare-4. An example of the screen form of the <<Gwoyeu Ryhbaw Tsyrdean>> is shown in Figure-5.

```
詞庫系統功能 : 部首查詢 (共  1 筆)     ※(F/B):顯示(下/上)筆資料 !!      <1>
頁/欄/序/號[884-2-13-0 ]詞彙[間                    ]詞頭部首/筆劃數[門:  4 ]
【發音】  1. [ㄐㄧㄢ
          2. [ㄐㄧㄢˋ
[共 2筆]  3. [
【詞意】  1. [音1- 1][兩個的當中. 如「兩人之間」「居間調解」.          ]
          2. [音1- 2][房屋的量詞. 如「一間房子」.                    ]
          3. [音1- 3][隔間的部分. 如「房間」「區間」.                ]
          4. [音1- 4][指時候. 如「日間」「晚間」.                    ]
          5. [音1- 5][指地方. 如「田間」「字裏行間」.                ]
          6. [音1- 6][見「間關」.                                    ]
          7. [音2- 1][空隙. 如「間隙」「間不容髮」.                  ]
[共13筆]  8. [音2- 2][隔開. 如「間隔」「間斷」.                      ]
【詞類】  1. 語位:[b]    詞意編號:[1,4,5                           ]
             詞類說明:[ng                                           ]
             屬    性:[                                            ]
[共 6筆]     附加說明:[                                            ]
詞抑制 ESC:
```

Figure-5. An example of the screen form of the <<Gwoyeu Ryhbaw Tsyrdean>>

2.2 Research objectives

Besides creating the ZXdb, this project has a few new research attempts as listed in the following objectives of the project.

(1) To find out whether the characters collected in the CCCII can meet the requirement of handling ancient dictionaries.

(2) To study the problems of handling the glyph (字形) and the font(字體) variations of ancient characters in computer.

(3) To study the feasibility and to develop techniques of switching from our old Chinese computing environment which is the AT&T 3B series mini-computers with the bi-lingual-UNIX operating system[16] onto some newly developed personnel computer(abbr. PC,afterwards) based environments.

3. The computing environment

There are two such newly developmented environments selected. One is an PC plus a CCCII card with DOS, the other one is the ITRIX with the CCCII implemented. These two environments are described as the followings.

3.1 The DOS environment

The ZXdb was firstly implemented on a personal computer with an 80386 processor and a 30 MByte hard disk, under the DOS operating system with dBASEIII software in March 1990. The operating system was extended to handle the latest version of the Chinese Character Code for Information Interchange (CCCII) [13] with 53,940 characters, by adding a plug-in CCCII card to the PC and equipped with a set of utility routines to facilitate the CCCII card. The major items of the CCCII card specification is listed in Table-1. The system is used as the data entry workstation, first, and then to build the ZXdb. An exemple of the screen forms is shown in Figure-6.

We found the 80386 processing power is OK. The CCCII card is good. But, the DOS can not support multi-tasking job becomes a major drawback of

Table-1 The major item of the CCCII card specifications

SPECIFICATIONS:

ITEM	DESCRIPTION
Chinese Code (option)	CCCII, CNS, BIG-5, IBM-5550, ANSI/NISO Z 39.64 EACC 1989, and Tele-communication code, etc. (CTSS 5226, KS)
Input Methods	Chang-Jie, Jian-Yi , Chu-Yin Fu-Hao, Phrase retrieval, Three-Corner Code Method, Internal code, Chinese Pin-Yin (Romanization of Chinese), Romaji Input System (Romanization of the Japanese alphabet–the Kana System), Romanization of the Korean alphabet–the McC-R (Hangul System), etc.
Printer Driver	EPSON, NEC, PANASONIC, FUJITSU DL Series dot matrix printer, Laser printer (OPTION).
Utility	Code convert driver, Pattern generator driver, Dictionary driver, Orthographic and simplified forms of Chinese characters convert driver, etc.
Terminal Emulation Driver (Emulating as workstation)	DEC VT100/VT220 , AT&T BX10, UNIX System, MICRODATA PRISM-4, etc.
JOIN Editor	Full screen editor for CJK Language System
DBASE III Library	For CJK Language System use.

PRODUCTS: CJK Cards

Model	Number of pattern cards	Characters set
P-22	1	22,000 chinese characters (Most Frequently Used Chinese Character Set and Next Frequently Used Chinese Character Set), and Japanese Hiragana, and Katagana.
P-33	2	33,000 chinese characters (Most Frequently Used Chinese Character Set, and Next Frequently Used Chinese Character Set and Variant and Simplified Forms of Chinese Orthographic Characters), Japanese Hiragana and Katagana, and Kanji, Korean Hangul and Hanja.
P-53	3	53,940 chinese characters (Most Frequently Used Chinese Character Set, and Next Frequently Used Chinese Character Set, Rarely Used Chinese Character Set and Variant and Simplified Forms of Chinese Orthographic Characters), Japanese Hiragana and Katagana, and Kanji, Korean Hangul and Hanja.

PATTERN CARD PHYSICAL DIMENSIONS (Net): 334mm × 106mm.
WEIGHT (Net) : 530 g

the system. Therefore, only a simplified single user system, with low cost, is considered worthwhile to be developed under the DOS environment.

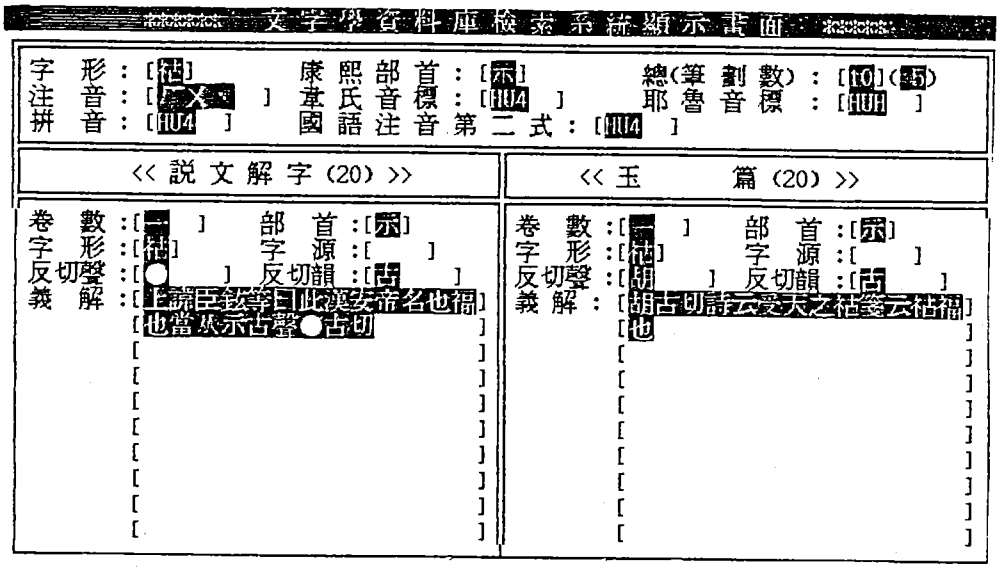


Figure-6. An exemple of the screen forms

As the UNIX system V release 4.0 (UNIX.SVR4) becomes available on PC, we think it is the O.S. we are longing for, because all our programs developed on Minis are also under the UNIX. As the CCCII is available on the UNIX .SVR4 CCCII a couple of months earlier , we have installed the second system, namely the ITRIX Chinese computing environment (蘭亭中文系統) with a 486 processor and a 300MByte hard disk , and we are now porting the system from 386 to 486 and from DOS to UNIX.

3.2 The ITRIX environment

The ITRIX, developed by the Computer and Communication Research Laboratories(CCL) of the Industrial Technology Research Institute (ITRI), is an outstanding Chinese-working environment comparing with other existing systems. The complete ITRIX includes two separate packages, the Chinese Base System(CBS) and the Chinese Window System(CWS).

The CBS includes the utilities of inputting Chinese characters under the

text mode, the console driver, the Chinese/Asia character printing system, the multi-fonts, the internal code transformation, and the application interface library (with Chinese computing capability added). The CWS includes the utilities of inputting Chinese characters for the X-Window, the Chinese-localized OPEN- LOOK and the Motif (not yet available at the time of preparing this paper). Multi-lingual input ability and multi-bytes libraries for windowing applications are also included in the CWS .

The ITRIX follows the specification of the EUC (Extended UNIX Code) [17]. The CCCII and the CNS11643 [18] are both adopted in the EUC. Furthermore, many development tools allow users to develop their own input methods, output methods and to generate new font alterations. Though the ITRIX is developed under the UNIX SVR4, yet users may choose other UNIX-based operating system, such as the INTEL UNIX ,to install the ITRIX.

3.2.1 An EUC implementation of Chinese character codes

According to the definition of the MNLS (Multi-National-Language-Supplement) [17], the EUC is comprised of a primary code set (set O) which is always assigned to the ASCII , and three supplementary code sets (set 1 through 3) for multiplying byte codes, such as the CNS 11643 and the CCCII.

The CNS 11643, used by the SEED project[19] and the SUN workstation series, uses 2 bytes and 4 bytes structures and are designated for the EUC set 1 and set 2, respectively. The CCCII uses a four-byte structure and is designated for the set 3. At the present time, 53,940 Chinese characters are included in the CCCII. This collection can soon be extended to an amount of over 74,OOO characters if needed. Both the CNS 11643 and the CCCII can be co-existed and used by user under the ITRIX.

The big collection of the CCCII minimized the chance for creating new characters and enhanced information interchange among parties. Also, the ASCII, the simplified Chinese characters, Japanese Kanji and Korean Hanja characters are all included in the CCCII. Therefore, the ITRIX with the CCCII is by nature a multi-lingual computing environment.

Code Set	EUC Code
0	0xxxxxxx
1	1xxxxxxx [1xxxxxxx [...]]
2	SS2 1xxxxxxx [1xxxxxxx [...]]
3	SS3 1xxxxxxx [1xxxxxxx [...]]

(a) the EUC Set

Code Set	EUC Representation
Code Set 0 (ASCII)	0XXXXXXX
Code Set 1 (CNS LEVEL1 & Symbols)	1XXXXXXX 1XXXXXXX
Code Set 2 (CNS LEVEL2 & Ext.)	SS2 1XXXXXXX 1XXXXXXX 1XXXXXXX
Code Set 3 (CCCII)	SS3 1XXXXXXX 1XXXXXXX 1XXXXXXX

(b) the ITRIX implementation

Figure-7. The EUC Code set and the ITRIX implementation of CNS 11643 and CCCII

3.2.2 The CWS

The CWS is a Chinese graphic user interface, which is built on the top of the X-Window system. The CWS system consists of two major components, the Chinese-localized X- Window system (C.X-Window) and the Chinese-localized OPEN- LOOK(C.OPEN-LOOK).

The C. X -Window preserves all the properties of the original X-Window and added some ingredient to make it more powerful and friendly. In the C.X-window, the X- server is remained unchanged, hence the CWS is capable to operate the X- Window system and other X- Window compatible systems, such as the PC X- Sight and X- terminal, simultaneously.

Similar to the CBS, the CWS also provides utilities for inputting Chinese character, creating and modifying input methods, creating new characters, etc. The utility library also provides the graphic capacity with a variety of fonts for C.X-Window.

The C. OPEN- LOOK is a graphic interface and windowing environment that capable of displaying and receiving Chinese text. It has five major tools, the Work Space Manager, the Window Manager, the Chinese Terminal Emulator, the File Manager, and the Administration Manager. It operates almost identically to the original English edition of the OPEN- LOOK. It has the capability of handling all information in Chinese language. In addition, C.X-Window provides the OPEN- LOOK Tool Kits with Chinese capability, including the Chinese-localized widgets and relative functions,for developing the Chinese OPEN- LOOK applications. In Figure-8,there is an example of the screen image under the C. OPEN- LOOK .

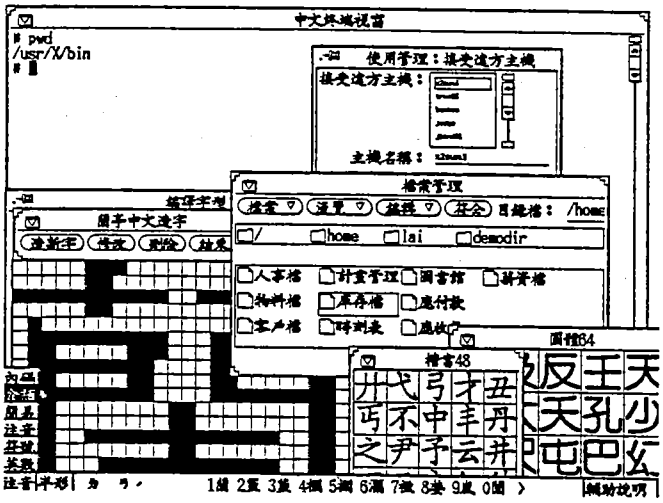


Figure-8. A screen copy of the Chinese-Localized OPEN-LOOK Environment

4. Some findings and concluding remarks

After a few months study, we have the following important findings. And, some remarks are also made there as the conclusions of this paper.

(A) The CCCII of 53,900 characters is not good enough for processing ancient dictionaries. A table of some relative numbers is shown below. Now, we are checking the missing characters against the 74,000 characters of CCCII. We don't know exactly how many characters are missing yet. But, one thing for sure, some entry characters in the

ancient font, such as  and are not included in the CCCII. Therefore, it is recommended that, for processing ancient documents, the CCCII must be extended to cover the ancient characters found by this project. It is expected that after the extension, the CCCII can be used to process most of the ancient documents.

TABLE 2. Some statistics of ancient dictionaries

Items	Shou-wen-jie-zi (說文解字)	Yu-pian (玉篇)
Total no. of Character in data base	207,244	310,173
data base size under DOS	5.7 MBytes	11.7 MBytes
No. of entry characters	11106 (9353+1753)	22800
missing entry character	1049	2231
No. of total missing character	?	?

(B) A very large proportion of the missing characters are some variants in ancient font. Therefore, we think a formal structure for organizing characters, their glyphs and associated font changes in computer system is needed for processing ancient documents. A good proposal for this problem was published by professor C.C. Hsieh[20].

(C) Some bugs were found in the ITRIX/C.X-window. This situation is not unusual for a new extension of an operating system. Although it is getting better day by day, but it does cause inconvenience.

(D) The data entry job is a painful experience. The Zang-je (倉頡) input method we used is very bad for our working material with a very large ancient character set . We know that the three-cornor-coding input method is much better than Zang-je (倉頡)[21], but since our typists are not familiar with the three-cornor-coding method, we have no other choice. Some characters are missing just because the typists can not figure out the right key strokes. After the key-in process, a lot of labor has been spending at verifying missing characters. The key-in rate is also low. A typical learning record of a typist is shown in the following Figure-9

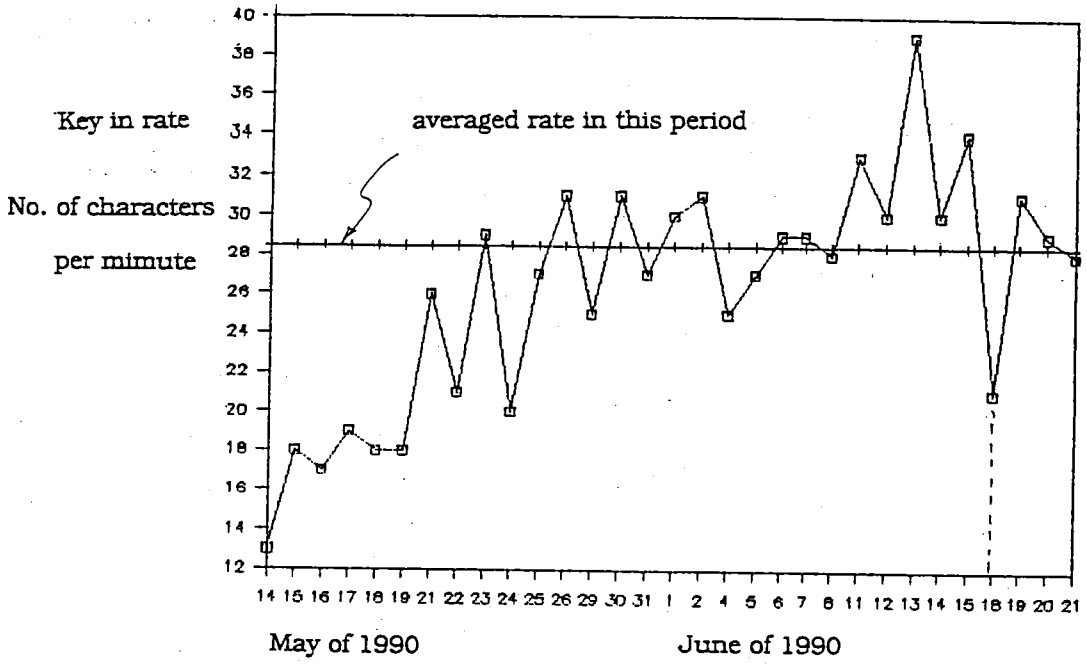


Figure- 9. A typical learning record of a typist during key-in the ancient dictionaries.

(E) Most of the program in C language used in our old Chinese computing environment can be used in the two new PC-based Chinese computing environments by recompiling the source C programs in the new system. But, in order to fully utilize the power of the ITRIX, especially for those user's interface programs, we still have to rewrite them.

(F) So far, the overall evaluation of our project shows that the ITRIX with CCCII on an 80486 based PC is a very promising candidate for replacing our old Chinese computing environment. And, a low cost version, an 80386 based PC with the CCCII card will be a good choice as a personnel equipment for Sinology studies.

Acknowledgement

The authors would like to express their deeply appreciation and thanks to Miss Fu,Wu-Chang(傅武嫦) for her excellent typing and verifying works of preparing this document. Also, thanks must present to Mr.Chris, C.L. Chang(張建良) for his valuable discussions with the authors while writing this paper.

REFERENCES

1.<<Computing Center A brief Sketch>>,Computing Center of Academia Sinica, 1990,Taiwan,R.O.C.

2.諱國蔭,<<卡系系統使用手册>>,Computing Cneter of Academia Sinica,1988,Taiwan,ROC.

3.林　晰,<<文獻層級結構運用於全文處理的研究>>,Computing Cneter of Academia Sinica, 1990,Taiwan,ROC.

4.陸念慈,<<文獻資料庫之文字統計系統套裝軟體使用手册>>,Computing Cneter of Academia Sinica,1989,Taiwan,ROC.

5.宋志隆, 何惠安, 褚台雄 ,賴彦丞,<<研究計畫用文獻資料庫使用手册>>, Computing Cneter of Academia Sinica,1990,Taiwan,R.O.C.

6.謝清俊,<Full Text Processing of Chinese Language(中文全文處理) > Annual Conference of The Association For Asian Studies,Sig Panel,1986, Chicago,U.S.A.

7.<<Proceedings of The ROCLING I.II.&III >>中華民國計算語言學會,1988-1990,Taiwan,ROC.

8.陳克健　黃居仁,<Information-Based Case Grammar> Proceeding of CoLing,1990 。

9.曾士熊, 楊鍵樵, 謝清俊,<中文全文資料庫之實驗模型>,中國工程學刊,1990, Taiwan,ROC.

10.中文詞知識庫小組,<<國語的詞類分析>>（修定版）Computing Cneter of Academia Sinica 1989,Taiwan,ROC.

11.黃克東, 張仲陶, 謝清俊, 楊鍵樵, 曾士熊,<Chinese Character Data Base(CCDB) And Coding Chinese,Japanese And Korean Machine Translation>,International Conference On Computer Processing of Chinese And Oriental Language,Toronto , Canada,1988 。

12.謝清俊, 魏文眞, 查全淑, <電子辭典在國語文教學方面的應用>,1989,<<華文世界>>,第51期 Taiwan,ROC.

13.The Chinese Character Analysis Group,<<The Chinese Character Code for Information Interchange>>,vol.I,II,III,and <<The variants forms of CCCII>>,1985,Taiwan,ROC.

14.許愼,<<說文解字>>

(1) <<四部備要>> 經部說文解字眞本(共兩册),1982,中華書局 據大興朱氏依宋重刻本景印

(2) 說文解字篆字譜,摘<<說文手册>>楊家駱主編,鼎文書局印行

15.顧野王, <<玉篇>>採用之版本: <<澤存堂五種>>張氏重刊宋本玉篇

16.An Chinese extension of the AT&T UNIX V by the ACER.

17.<<UNIX System V Multy-national Language Supplement Release 3.2 Product Overview>> ,1990,AT&T.

18.<<中國國家標準 通用漢字標準交換碼>>,經濟部中央標準局,Sept.1986,Taiwan,ROC.

19.<<SEED(中文透通計畫簡介)>>,財團法人資訊工業策進會,March,1991,Taiwan,ROC.

20.謝清俊, 張仲陶, 黃克東 ,<On the Formalization of Glyph In Chinese Language>, AFII meeting at Kyoto,Feb. 1990

21.By personnel contact with professor 勝村哲也(Prof. Katzmura) the school of Humanities Studies,Kyoyo University