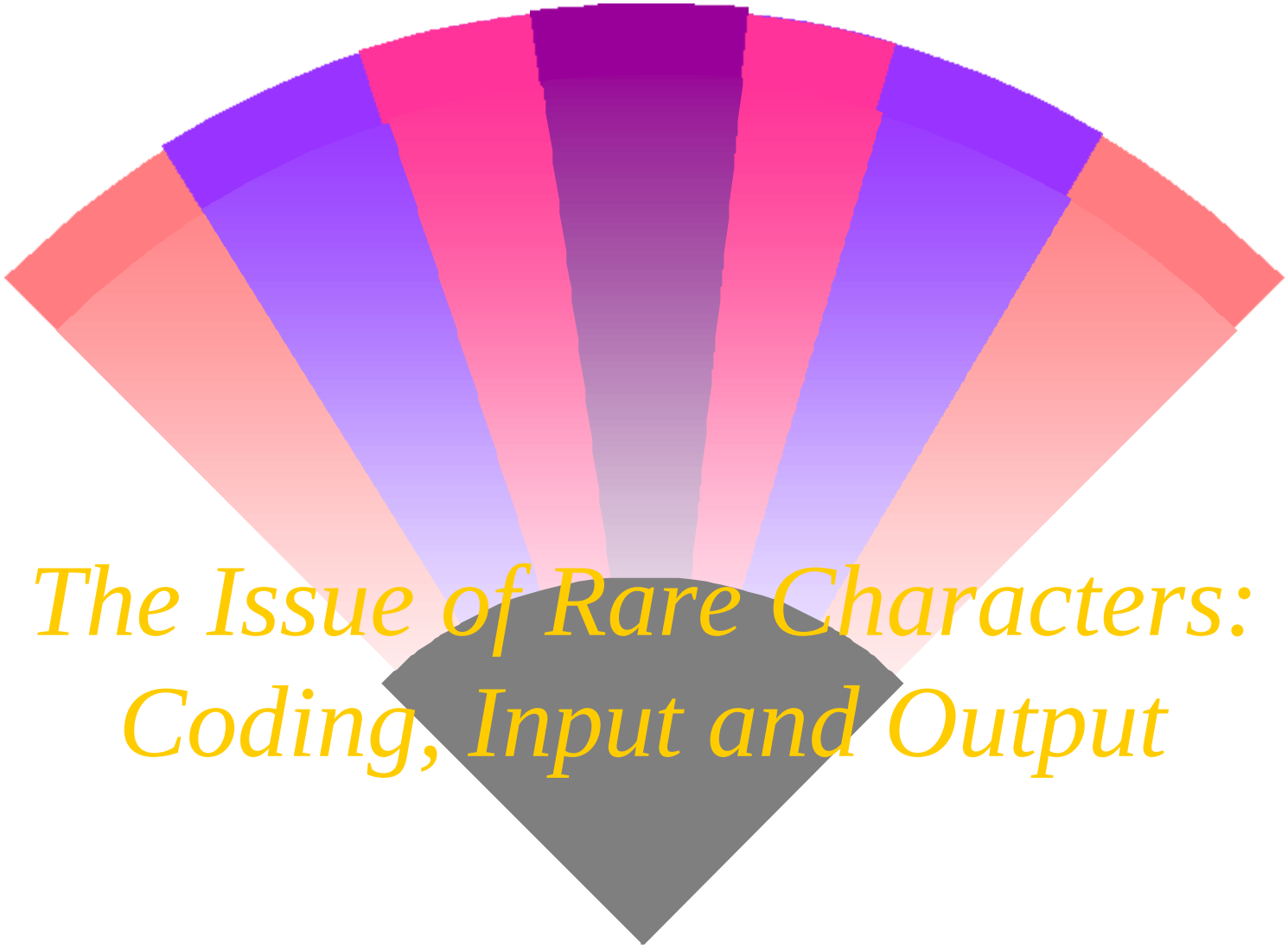


The Issue of Rare Characters: Coding, Input and Output

Approaches to the problem

A Use Black blobs or ‘similar’ characters

B Create ‘user-defined characters’

C Paste images of characters

D Use large collections of characters

E Describe the missing characters

A Use Black blobs or ‘similar’ characters

- Advantage: very easy
- Disadvantage: loss of information
- Conclusion: No solution to this problem

B Create 'user-defined characters'

- Advantage:
  - Can be done on almost every system
  - Loss of information can be avoided
- Disadvantage:
  - No portability between systems
  - Data from different sources cause confusion
  - Difficult to use on the Internet

C Paste images of characters

- Advantages:
  - Can be done in standard applications
  - Can be used on the Internet
- Disadvantages:
  - Loss of information
  - Characters not searchable

D Use large collections of characters

- Advantages:
  - Most characters can be directly used
  - Allows interchange of information
  - Allows search of characters
- Disadvantages:
  - Can not possibly include *all* characters
  - Useless without access to the reference database

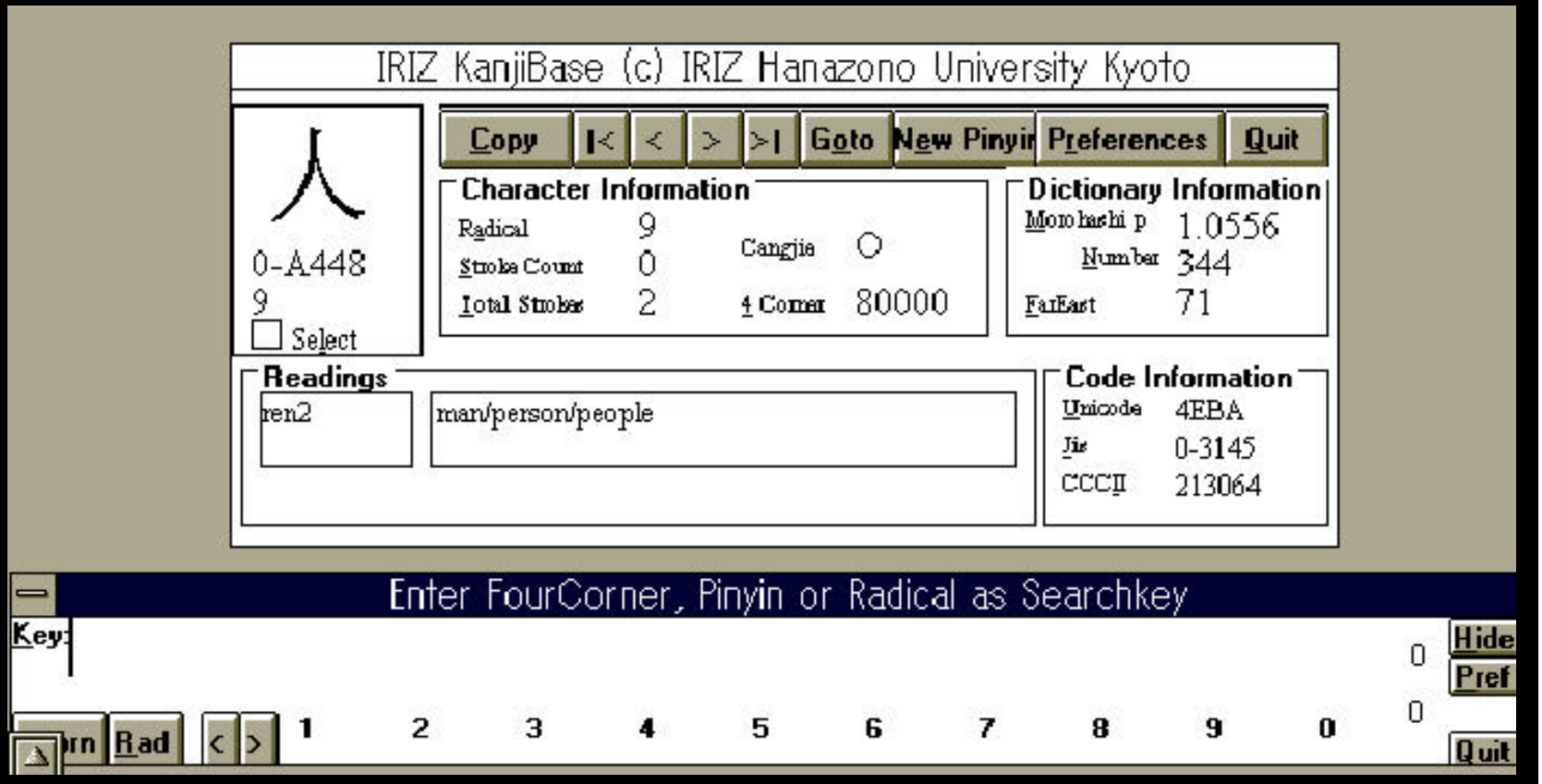
E Describe the missing characters

- Advantages:
  - Information about the character is transmitted
  - Allows interchange of information
- Disadvantages:
  - A processing system is required
  - Display and printing on standard systems not satisfying

Recent developments

- Large reference collections:
  - *KanjiBase* (1995): 48000 Characters
  - *Mojikyo* (1997): 80000 Characters
  - *eKanji* (1997): 65000 Characters
  - *Hanziku* (1998): 56000 Characters
- Character descriptions
  - *Academia Sinica* (since 1993)
  - *Mojikyo* (1997)
  - *CBETA* (1998)

KanjiBase


IRIZ KanjiBase (c) IRIZ Hanazono University Kyoto
Copy
|<
<
>
>|
Goto
New Pinyin
Preferences
Quit
0-A448
9
Select
Character Information
Radical 9
Stroke Count 0
Total Strokes 2
Cangjie
4 Corner 80000
Dictionary Information
Morohashi p 1.0556
Number 344
FarEast 71
Readings
ren2
man/person/people
Code Information
Unicode 4EBA
Jis 0-3145
CCCII 213064
Enter FourCorner, Pinyin or Radical as Searchkey
Key:
Hide
Pref
Quit
Rad
1 2 3 4 5 6 7 8 9 0

KanjiBase

- Uses CNS 11643-1992 as encoding
- Access to all 48027 characters by Fourcorner Code, Radical and Strokecount
- Many characters can also be searched by Pinyin romanization
- Display of character properties, dictionary information and codepoints in other codes
- Accessible on the Internet at http://www.gwdg.de/~cwitter

Mojikyo

漢字検索システム

今昔文字鏡

Version 1.00

Mojikyo


漢字検索システム [今昔文字鏡]
コード表(T)
ヘルプ(H)
文字部品
1
2
3
4
5
部品コピー(X)
部品消去(Z)
検索キー指定(K)
部品検索(S)
検索オプション
派生(Y)
送り省略(U)
解字検索(B)
最小一致(J)
検索実行(E)
部品転送(G)
終了(C)

Mojikyo

Mojikyo

- Encodes all characters in Morohashi's *Daikanwa jiten*大漢和辭典, all Han characters in Unicode and many more.
- Access through character fragments of inputted characters
- The serial number can be used to uniquely encode each character
- Database frontend only on Japanese Windows 95 or NT

eKanji

- Union database of Morohashi's *Daikanwa jiten*大漢和辭典, the *Kangxi Zidian* 康熙字典 and Unicode with currently 65000 characters
- Bitmapped fonts have been made available on the URL http://www.zinbun.kyoto-u.ac.jp/~ekanji
- No attempt has been made to provide an encoding that can be used in texts

Hanziku

- Encodes all characters in *Hanyu dazidian* **漢語大字典**
- Access through the printed dictionary (page and character number)
- Outline fonts of all 58000 characters for Chinese Windows 95 and NT

Recent developments

- Large reference collections:
  - *KanjiBase* (1995): 48000 Characters
  - *Mojikyo* (1997): 80000 Characters
  - *eKanji* (1997): 65000 Characters
  - *Hanziku* (1998): 56000 Characters
- Character descriptions
  - *Academia Sinica* (since 1993)
  - *Mojikyo* (1997)
  - *CBETA* (1998)

Academia Sinica

- Complete character description system on a sound theoretical base
- Operators and character parts use ‘user defined characters’
- Database frontend usable only on Windows 95 or NT

Mojikyo

- Character splitting without operators
- Purpose: Database retrieval, not complete description of characters
- Stand alone solution without reference to ‘user defined characters’ on the system

CBETA (吳寶原)

- Use only characters from the system character-set to form glyph expressions, i.e
  音＝立/日 or 因＝口@大 or even
  繞＝組-且+((土/(土*土))/兀)
- No formalized definition for each character
- Identical characters may have different representations

Evaluation

- Cultural bias
  - System environment
  - Access method(s) for characters
- Availability
  - Open and free standard for method and access
  - Commercial solutions

Conclusions

- No single method provides a complete satisifying solution at this moment
- Combination of different methods provide some basic usability
- Interchangeability between the various methods is of high importance

Recommendations

- PNC should encourage efforts to create crossreferences to the different existing databases
- PNC should develop recommendations for the implementation of interoperable encoding strategies