Overview of CHISE:

— its Chinese Character ontology

and some applications—

2021-06-09

Morioka Tomohiko (守岡 知彦)

CHISE Project

CHaracter Information Service Environment

open source research and development project aiming to realize text processing environment that can freely use characters according to character definitions by users

Character Definition = Character Ontology

- User can define characters
- Character processing using character ontology

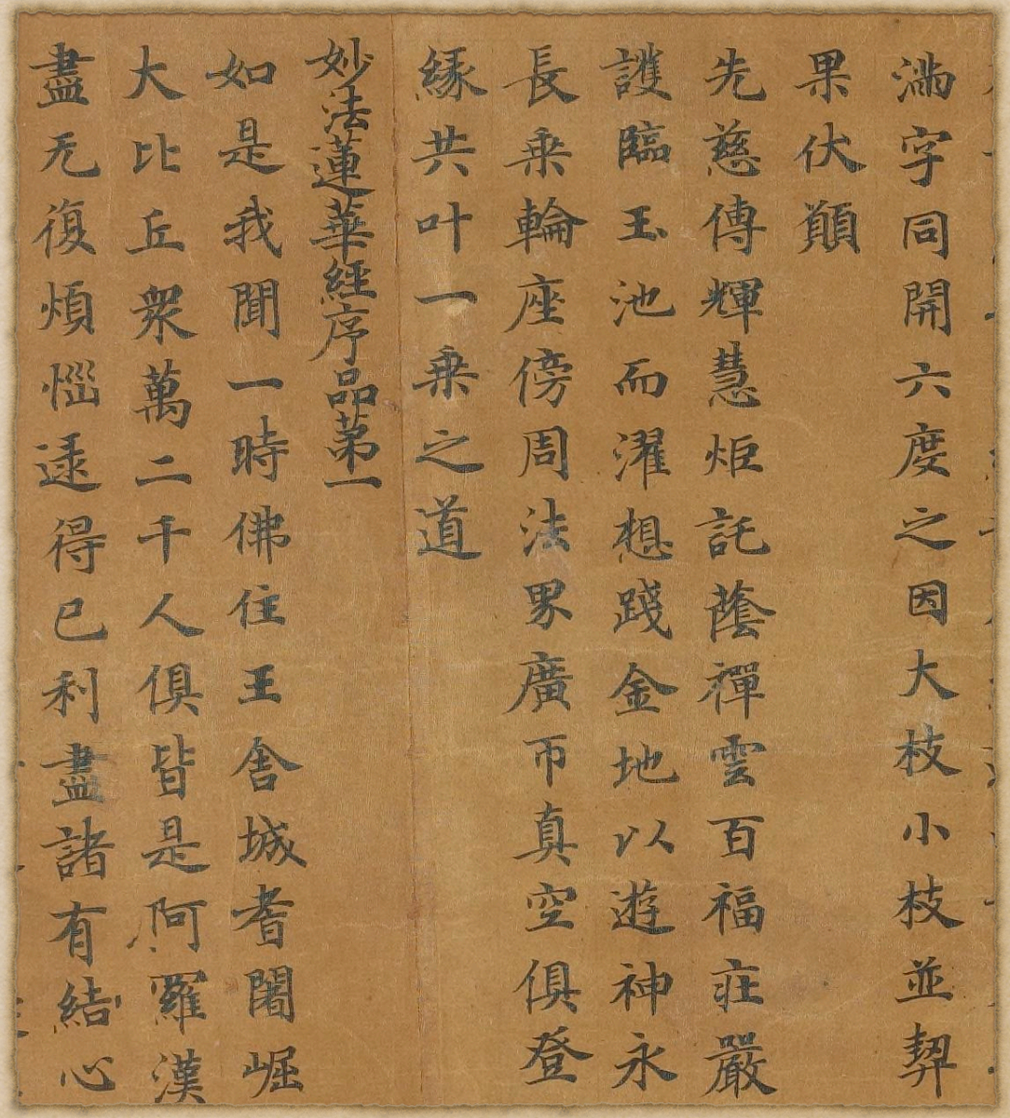
Character ontology + clients

http://www.chise.org

How to represent Chinese characters in the real world

- How to encode?
  - Which and which are the same characters?
  - What is "same character"?

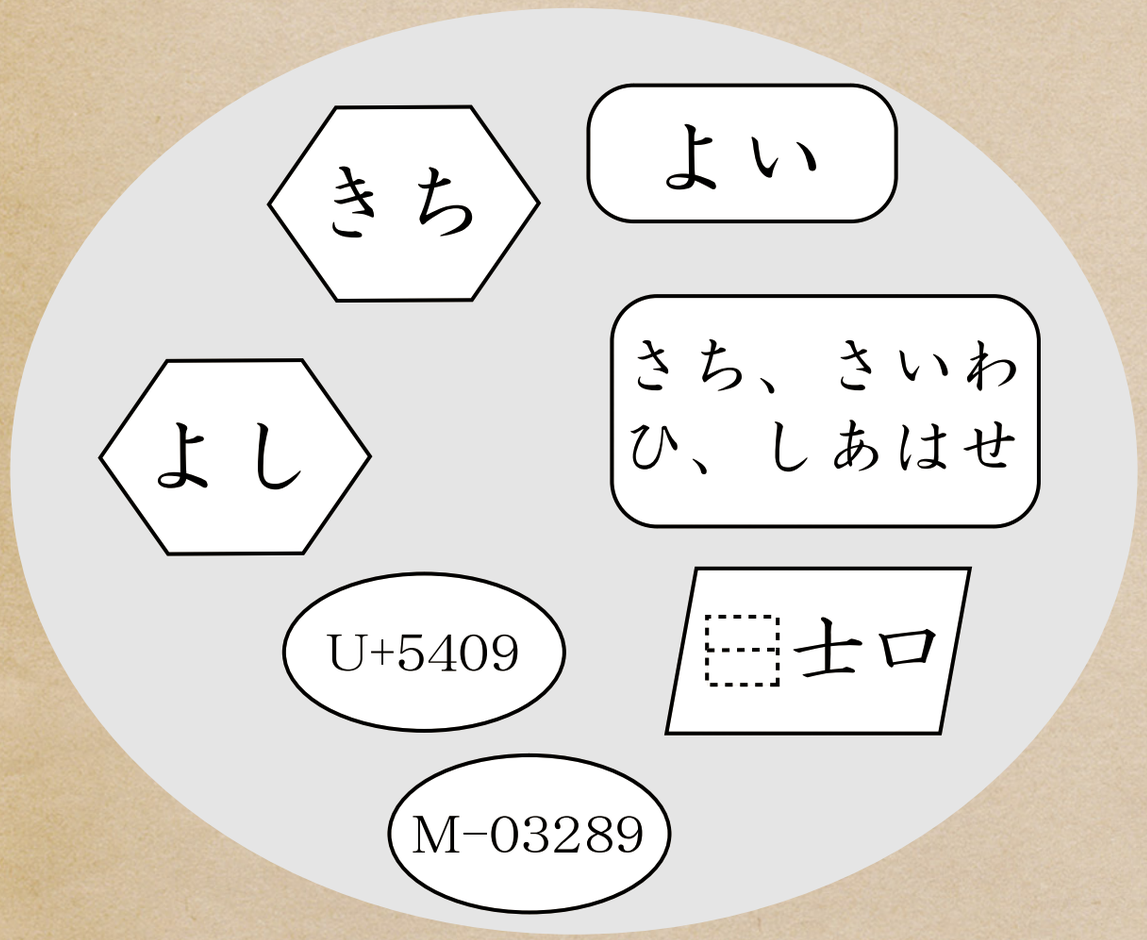
Ontology of characters


滿字同開六度之因大枝小枝並葉
果伏顛先慈傳輝慧炬託蔭禪雲百福莊嚴
長乘輪座傍周法界廣布真空俱登
緣共叶一乘之道
護臨玉池而濯想踐金地以遊神永
妙法蓮華經序品第一
如是我聞一時佛住王舍城耆闍崛
大比丘衆萬二千人俱皆是阿羅漢
盡无復煩惱逮得己利盡諸有結心

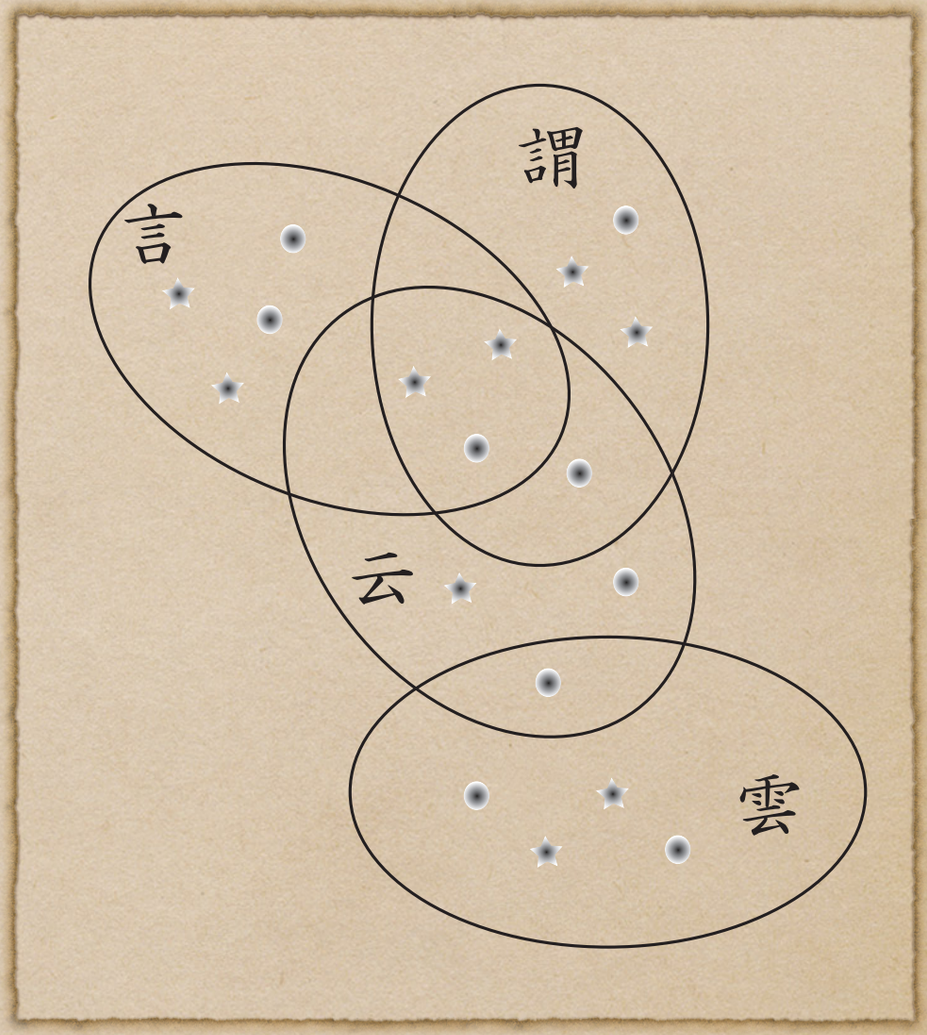
Designation of characters

- Each code point of a coded character set is a rigid designator for abstract character
- Can we use definite descriptions (set of properties) to designate characters?

Chaon model



Character representation based on set of features



Set operation of character features

Character operations

- Define character (create new object)
- Set feature (put property value) of a character
- Get feature of a character
- Map / iteration for each character with feature
- Find characters with feature(s)

What is character feature (1)

- Character feature = abstract of an operation on characters
  - Display / print
  - Save text / code conversion
  - Find character / text search
  - sort
  - etc.

What is character feature (2)

- Glyph Information : Display, print
- Various character codes : encode/decode text, code conversion, etc.
- Strokes, phonetic values : Search, sort, etc.
- Structure Information of Chinese characters : find characters, automatic glyph composition (display, print) , etc.
- Other dictionary descriptive information : find characters, etc.

→ Interface for characters

What is character feature (3)

→ Interface for characters

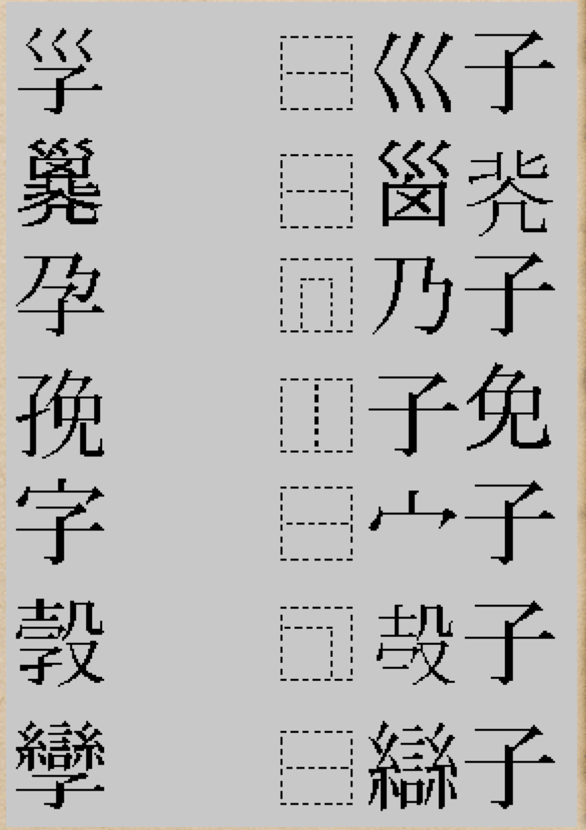
- It cannot be explained within the scope of the Chaon model

→ Necessity of fundamental research for Chinese characters and writing system

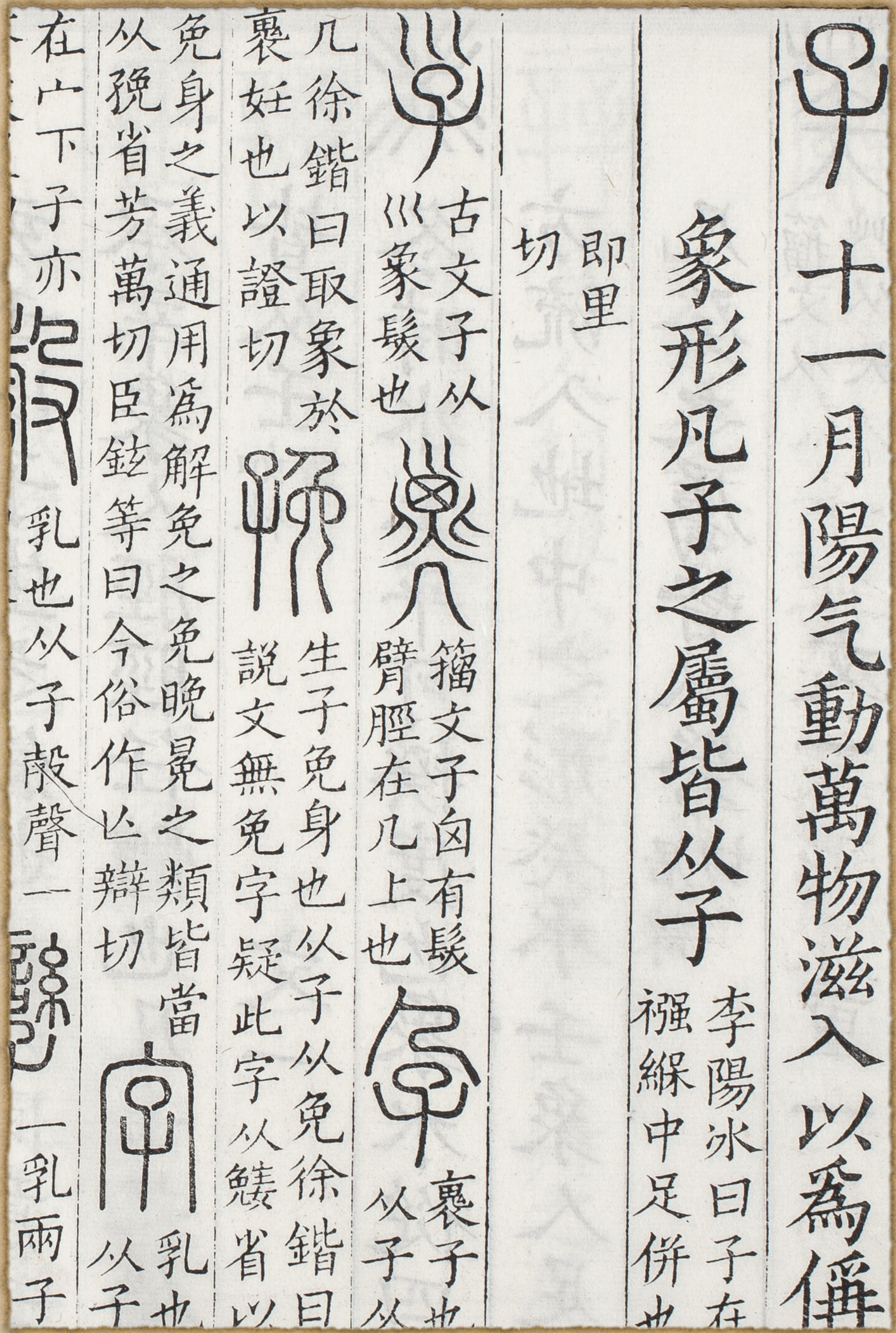
Shuowen Jiezi (説文解字) as an object model for Chinese characters

Many Chinese characters are complex characters composed of multiple components

- If we teach the knowledge to a computer

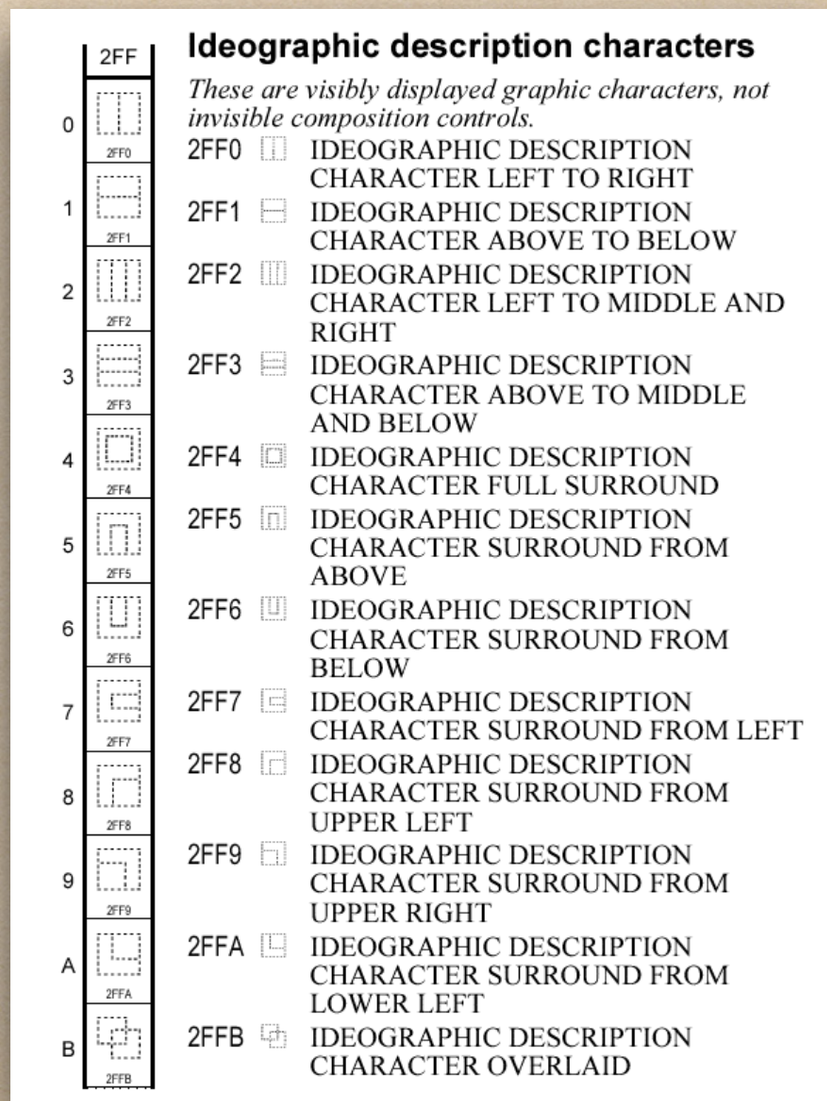


- Each character and component has features such as glyphic/phonetic/semantic values.
- Composite characters may or not may inherit features of their components



Ideographic Description Sequence

- Two or three components after IDC (it is a kind of prefix operator)
  - e.g. ⿰木寸　（村）
    ⿳口工口（嚚）
- IDS can also be used as component（IDS can be nested)
  - e.g.：⿰糸⿱不土（絰）



Ideographic description characters

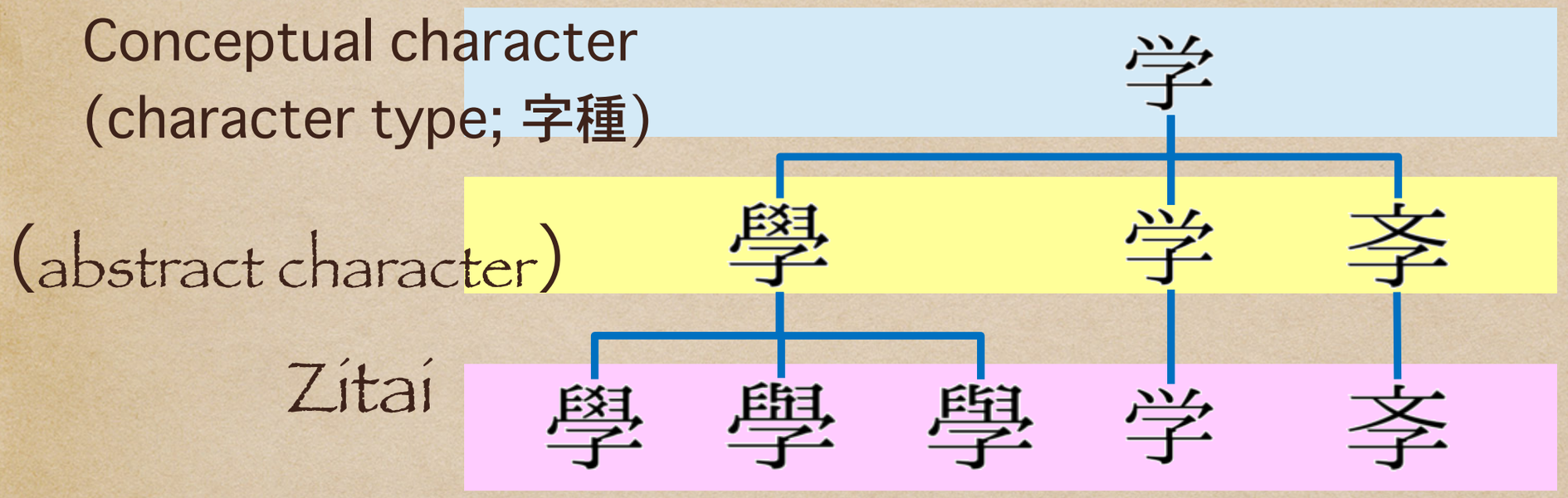
These are visibly displayed graphic characters, not invisible composition controls.

2FF0	⿰	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT
2FF1	⿱	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW
2FF2	⿲	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT
2FF3	⿳	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW
2FF4	⿴	IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND
2FF5	⿵	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE
2FF6	⿶	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW
2FF7	⿷	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT
2FF8	⿸	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT
2FF9	⿹	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT
2FFA	⿺	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT
2FFB	⿻	IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

Glyph and Glyph image

- Chinese characters can be written in various forms, so there is no time to distinguish all of them.

Unificatioin of *Zítai* (≒ glyph)



Unify similar glyphs into one abstract character

➡ Unification rules

- Abstract glyph design differences before applying the unification rules

- JIS X 0208:1997, JIS X 0213
- IWDS-1 (List of Unifiable Component Variations of Ideographs of UCS)

Glyph design differences

（３）接触の位置に関する例

岸 岸　家 家　脈 脈 脈

蚕 蚕　印 印　蓋 蓋

（４）交わるか，交わらないかに関する例

聴 聴　非 非　祭 祭

存 存　孝 孝　射 射

（５）その他

芽 芽 芽　夢 夢 夢

3　点画の性質について

（１）点か，棒（画）かに関する例

帰 帰　班 班　均 均　麗 麗　蔑 蔑

（２）傾斜，方向に関する例

考 考　値 値　望 望

（３）曲げ方，折り方に関する例

勢 勢　競 競　頑 頑 頑　災 災

（４）「筆押さえ」等の有無に関する例

芝 芝　更 更　伎 伎

2　筆写の楷書では，いろいろな書き方があるもの

（１）長短に関する例

雨 － 雨 雨　　戸 － 戸 戸 戸

無 － 無 無

（２）方向に関する例

風 － 風 風　　比 － 比 比

仰 － 仰 仰

糸 － 糸 糸　礻 － 礻 礻　礻 － 礻 礻

主 － 主 主　　言 － 言 言 言

年 － 年 年 年

（３）つけるか，はなすかに関する例

又 － 又 又　　文 － 文 文

月 － 月 月

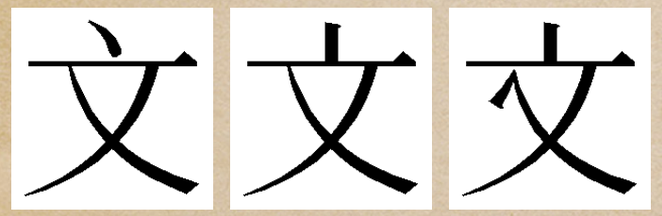
条 － 条 条　　保 － 保 保

（４）はらうか，とめるかに関する例

奥 － 奥 奥　　公 － 公 公

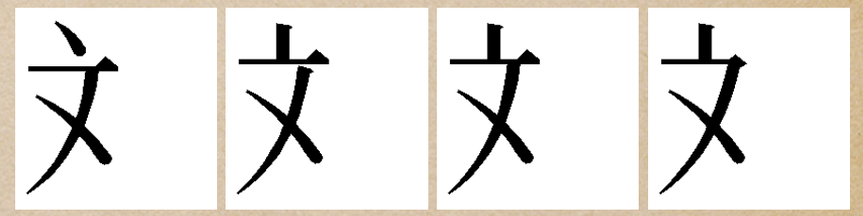
角 － 角 角　　骨 － 骨 骨

Boundary between glyph (*Zítaí*) difference and glyph design difference

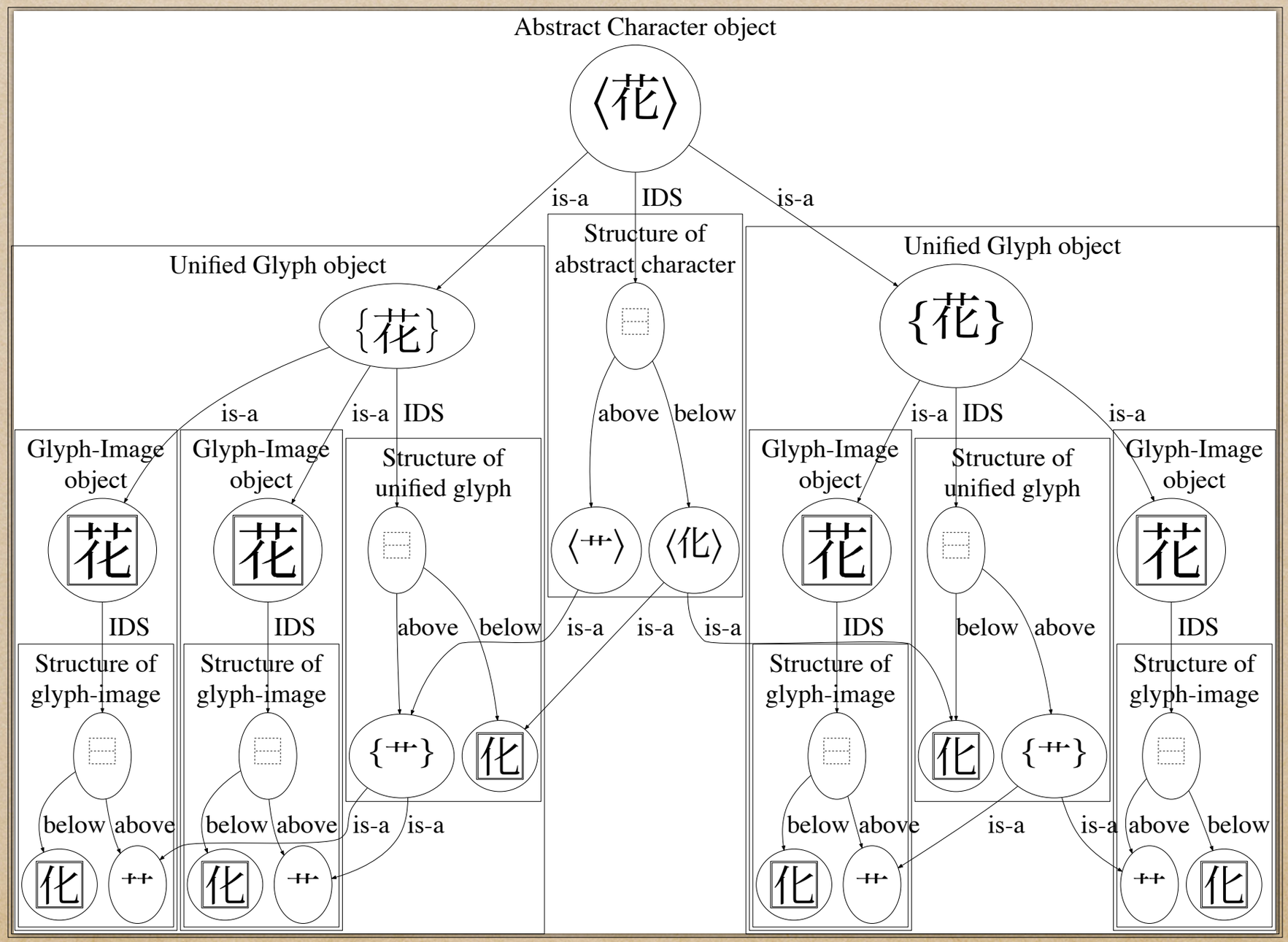

→ Sometimes difficult to distinguish clearly

- However, the unifiable range cannot be determined unless a boundary line is drawn somewhere.




different characters

Multiple Granularity Hanzi Structure Model



CHISE Web servises

- CHISE IDS Find：

  searching Chinese characters that contains specified components

  http://www.chise.org/ids-find

- CHISE-Wiki (E_ST)：

  display information of character (object)

  http://www.chise.org/est/view/character/字

- SPARQL endpoint：

  http://rdf.chise.org

Glyph Examples of Chinese characters

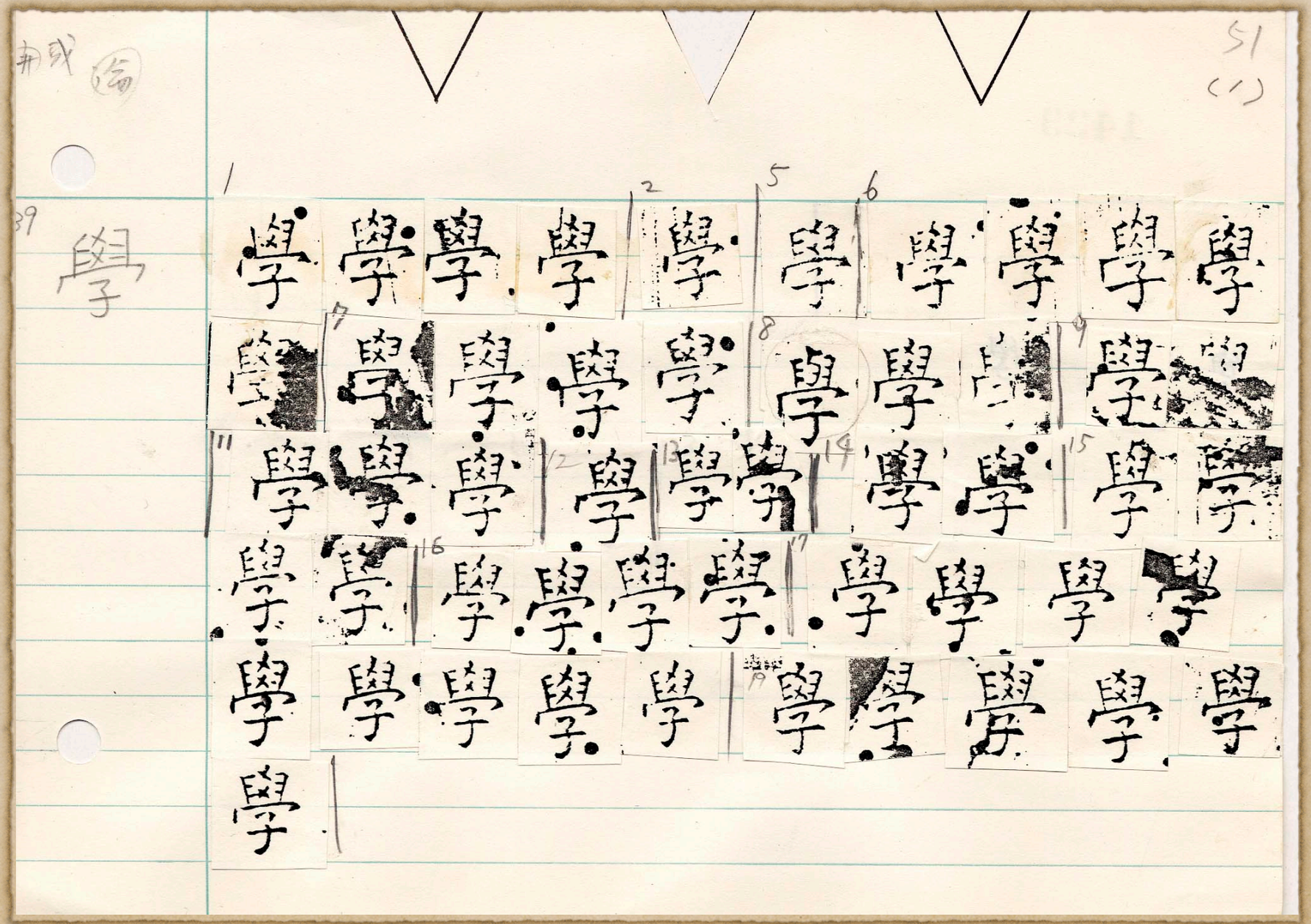
- HNG (Hanzi Normative Glyphs) dataset
  - Daijiten （大字典）dataset
- Shuowen Jiezi (説文解字) in Center for Informatics in East Asian Studies, Institute for Research in Humanities, Kyoto University
  - Clipped data was developed by Suzuki Toshiya (Hiroshima University)
- (Oracle Bones characters in Center for Informatics in East Asian Studies, Institute for Research in Humanities, Kyoto University [京都大学人文科学研究所所蔵甲骨文字] (under developed))

HNG (Hanzi Normative Glyphs) dataset

- A kaishu (楷書) glyph corpus of Chinese characters from 63 manuscripts (including stone rubbing images and woodblock printed books)
- Based on "Ishizuka Register of Chinese Character Standards of Writing" (石塚漢字字体資料)

https://gitlab.hng-data.org/HNG

Ishizuka Register of Chinese Character Standards of Writing



HNG character search

- HNG one character search「漢字字体規範史データセット単字検索（HNG 単字検索）」

https://search.hng-data.org

- CHISE-HNG IDS Find

https://www.chise.org/hng-ids-find

- (Multi-database Search System for Historical Chinese Characters: https://mojiportal.nabunken.go.jp/en/)

HNG and Daijiten

- In the development of HNG (石塚漢字字体資料), Daijiten （大字典; 1917 (ed.) 上田万年、岡田正之、飯島忠夫、栄田猛猪、飯田伝一） was used to classify glyph materials. So It is important to understand HNG dataset
- Daijiten has an aspect as a dictionary for Kanji in Japanese, so it is an important dictionary for Japanese linguistics.

Daijiten dataset

https://gitlab.hng-data.org/HNG/daijiten-data

- Daijiten character data (daijiten_DB.txt)
- Page data (daijiten_page_number.csv)
  - ✳ (Image data of two versions are available at National Diet Library of Japan, so we can link to full image)

Conclusion

CHISE (Character Information Service Environment) is:

- a character processing system based on character ontology
  - not depended on general character code (such as UCS)
  - make the computer understand various knowledge about characters
  - character processing based on machine-readable knowlege

➡ Works as a meta system of UCS or other private character set to describe characters and their knowledges

Resources

- CHISE project: https://www.chise.org/
  - CHISE IDS Find: https://www.chise.org/ids-find
  - Git repositories: https://gitlab.chise.org/CHISE
    - CHISE IDS: https://gitlab.chise.org/CHISE/ids
      ([mirror] https://github.com/chise/ids)

- HNG (Hanzi Normative Glyphs) dataset:
  - Git repositories: https://gitlab.hng-data.org/HNG
  - HNG 単字検索: https://search.hng-data.org/
  - CHISE-HNG IDS Find: https://www.chise.org/hng-ids-find

References

- Tomohiko Morioka: “Viewpoints on the Structural Description of Chinese Characters”, Grapholinguistics in the 21st Century 2020, Grapholinguistics and Its Applications (ISSN: 26818566, eISSN: 25345192), Vol. 5, pp. 683–712. https://doi.org/10.36824/2020-graf-mori

- Morioka Tomohiko: “Integration of a Chinese Character Ontology Title and Historical Glyph Examples.” In: 9th International Conference of Digital Archives and Digital Humanities (DADH 2018). Taiwanese Association for Digital Humanities / Dharma Drum Institute of Liberal Arts, pp. 287–300.

- Morioka, Tomohiko: “Multiple-policy character annotation based on CHISE,” Journal of the Japanese Association for Digital Humanities 1(1), p. 86–106.

- Ishizuka, Harumichi: “Current status and future prospects of the Hanzi Normative Glyphs (HNG) database”, http://idp.bl.uk/downloads/hng_translation.pdf.

- Morioka, Tomohiko: 2016. CHISE Guidelines for Glyph Granularity of Chinese characters （CHISE 文字オントロジーのための漢字字体・字形粒度の情報記述に関するガイドライン） [Ver.0.9.1]. http://www.chise.org/specs/ggg_v0.9.1.pdf.

- IRG Working Document Series (IWDS). http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html.