京都大学
KYOTO UNIVERSITY

Title	Integration of a Chinese character ontology and Historical Glyph Examples
Author(s)	Morioka, Tomohiko
Citation	Proceedings of the 9th International Conference of Digital Archives and Digital Humanities (DADH2018) (2018): 287-300
Issue Date	2018-12
URL	http://hdl.handle.net/2433/241079
Right	
Type	Conference Paper
Textversion	publisher

Integration of a Chinese character ontology and Historical Glyph Examples

Morioka, Tomohiko

Ph.D / Assistant Professor

Center for Informatics in East Asian Studies

Institute for Research in Humanities

Kyoto University

Integration of a Chinese character ontology
and Historical Glyph Examples

Morioka, Tomohiko

Ph.D / Assistant Professor

Center for Informatics in East Asian Studies

Institute for Research in Humanities

Kyoto University

Abstract

This report describes an attempt to integrate the "CHISE" ("Character Information Service Environment") character ontology and the "HNG" ("Hanzi Normative Glyphs") database / dataset.

The CHISE character ontology is a large scale character ontology which includes 357 thousand character-objects including Unicode and non-Unicode characters and their glyphs, etc. It was developed for CHISE which is a character processing system not depended on character codes. The framework of CHISE is based on a graph storage named "CONCORD". We developed a Web service to display and edit objects of CONCORD, called "EsT" (or "CHISE-wiki").

The CHISE character ontology uses the "Multiple Granularity Hanzi Structure Model" to support various glyphs and multiple unification granularity of Chinese characters. This model works fine for modern glyphs of Chinese characters. However, before we started the study to integrate CHISE and HNG, it was not clear that the model is sufficient for premodern Chinese characters. In addition, to design reasonable unification rules for each unification granularity, we need various glyph examples of Chinese characters. In these senses, the CHISE character ontology should integrate glyph database and/or glyph corpus. Therefore, we tried to integrate HNG and the CHISE character ontology.

When viewed from the HNG side, this integration has the following significance. The original HNG web service has been stopped since the spring of 2015. Therefore, we applied research on the integration of CHISE and HNG, we provided HNG search function and data browsing function on the CHISE Web service.

Table of Contents

Keywords

Chinese character, glyph, linked data, database integration, dataset preservation.

1. Introduction

This report describes an attempt to integrate the "CHISE" ("Character Information Service Environment") character ontology (1) and the "HNG" ("Hanzi Normative Glyphs") database (2) / dataset (3).

Glyph database and glyph corpus including historical glyph examples of Chinese characters, such as "HNG (Hanzi normative glyphs) database" or "Character Database of Digital Rubbings" (「拓本文字データベース」) [4], are useful tools in considering the historical transition of the Chinese character glyphs and their normative consciousness. In particular, HNG is designed to display a relatively small number of typical examples of glyph-images of *kaishu* (楷書) to demonstrate that each time period and geographical region (country, state) had its own orthographic standard which differed from that of other time periods and geographical regions.

HNG is designed and developed based on various (deep) knowledges about Chinese characters and codicology, however these background knowledges may be tacit knowledges and/or they are not machine-readable knowledge. HNG does not have machine readable knowledge about unification granularity (namely how glyph-images are classified into abstract glyphs (字體)) and structure of characters (compositions of components described by IDS or other formats). Therefore, users can search through abstract characters instead of directly searching for glyphs (although HNG has its own criteria for abstract glyphs).

On the other hand, we have a character information service named "CHISE" ("Character Information Service Environment") (5) that can complement HNG. "CHISE IDS Find" (「CHISE IDS 漢字検索」) (6) is a one of the most powerful tools to search complicated Chinese characters. In a result page of CHISE IDS Find, the first column of each line indicates an entry for a character. It has a link for a page in the "CHISE-wiki" ("EsT") to display details of the character. CHISE-wiki has links for Chinese character related Web services such as "GlyphWiki" and "UniHan database", or other service such as the "Classical Chinese Morphological Linked Open Data" (「古典中国語形態素 LOD」) and the "Bibliography of Oriental Studies" (「東洋學文獻類目」) database [7]. CHISE-wiki is a good tool to find information of each character and it also displays usage of each character.

These Web services of CHISE are based on the "CHISE character ontology". It is a large-scale character ontology which includes 357 thousand character-objects including Unicode and non-Unicode characters and their glyphs, etc. It uses the "Multiple Granularity Hanzi Structure Model" to support various glyphs and multiple unification granularity of Chinese characters. This model works fine for modern glyphs of Chinese characters. However, before we started the study to integrate CHISE and HNG, it was not clear that the model is sufficient for premodern Chinese characters. In addition, to design reasonable unification rules for each unification granularity, we need various glyph examples of Chinese characters. In these senses, the CHISE character ontology should integrate glyph database and/or glyph corpus. Therefore, we tried to integrate HNG and the CHISE character ontology. In addition, the original HNG web service has been stopped since the spring of 2015. We urgently needed to set up an alternative service for the original HNG. Therefore, we applied research on the

integration of CHISE and HNG, we provided HNG search function and data browsing function on the CHISE Web service.

2. HNG

"Ishizuka Register of Chinese Character Standards of Writing" (「石塚漢字字体資料」) is a set of paper cards [Figure 1] comprising 400,000 character instances from 69 manuscripts. It is created by Emeritus Professor Harumichi Ishizuka of Hokkaido University during his tenure in Japanese linguistic seminar or other educational and research programs.
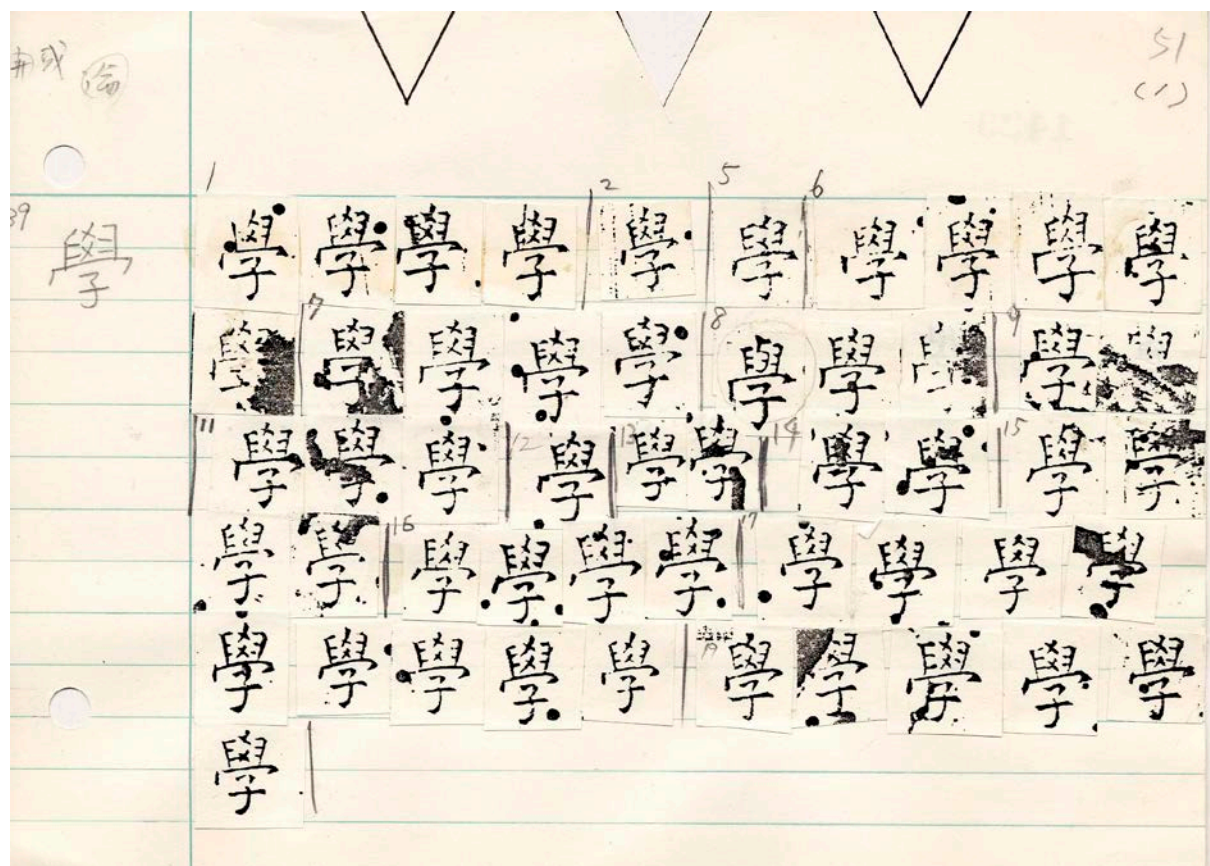

Figure 1: A sample of paper card (「學」 at 「開成石経論語」)

In order to deal with the deterioration problem of paper cards and to provide the data to the academic circle, it was digitized by researchers and students in Linguistic Sciences, Graduate School and Faculty of Letters, Hokkaido University. The result of this project was released on the Internet in 2005 under the name "Hanzi normative glyphs (HNG) database" (「漢字字体規範史データベース」) (2).

The original HNG database continued its Web service for ten years, but it stopped in the spring of 2015. In order to solve this problem, we have offered alternative service using CHISE technology, we concluded that Web-based search service alone is insufficient to maintain stable data availability over the long term. Therefore, we started new project to release HNG as a dataset, named "Hanzi normative glyphs (HNG) dataset" (「漢字字体規範史データセット」) (3). Currently, it contains 48 sources listed in Table 1.

In the original HNG database, not all glyph examples in each source of the specified character were displayed, only the representative glyph image was displayed (if a source has two or more glyphs for the specified character, two or more representative glyph images were displayed). However, we received many requests from users wanting to confirm all examples. Therefore, images of paper cards of the "Ishizuka Register of Chinese Character Standards of Writing" were also added to the HNG dataset.

No. (folder code)	ID	Type	Name of source	Abbrev	Age
1 (H03)	dng	南北朝写本	S81 大般涅槃経巻十一	S81	506
2 (H02)	keg	南北朝写本	S2067 華厳経巻十六	S2067	513
3 (H01)	jou	南北朝写本	P2179 誠実論巻八	P2179	514
4 (H05)	mam	南北朝写本	P2160 摩訶摩耶経巻上	P2160	586
5 (H06)	drt	隋写本	P2413 大楼炭経巻三	P2413	589
6 (H07)	kgk	隋写本	賢劫経巻二	賢劫経二	610
7 (H08)	myz	隋写本	P2334 妙法蓮華経巻五	P2334	617
8 (H10)	khi	初唐写本	今西本妙法蓮華経巻五	宮廷今西	671
10 (H11)	khm	初唐写本	守屋本妙法蓮華経巻三	宮廷守屋	675
11 (H13)	hok	初唐写本	S2577 妙法蓮華経巻八	S2577	7C 末
12 (H14)	kyd	初唐写本	上野本漢書揚雄伝	漢書揚雄	初唐
13 (H15)	sok	則天写本	守屋本花厳経巻八	華厳守屋	則天期
14 (H16)	yhk	盛唐写本	S2423 瑜伽法鏡経	S2423	712
18 (H09)	kda	高昌写本	大品経巻二十八	京博大品	高昌期
19 (H22)	sys	吐蕃写本	S5309 瑜伽師地論巻三十	S5309	857
20 (H49)	wan	大和寧写本	花厳経巻三十八	和寧花 38	9-10C
24 (H18)	kar	開成石経	開成石経論語	開成論語	837
25 (H19)	kae	開成石経	開成石経周易	開成周易	837
26 (H17)	kak	開成石経	開成石経孝経	開成孝経	837

29 (H25)	tzj	北宋版	東**禅**寺版阿毘達磨大毘婆沙論巻百七	東**禅**毘婆	1100
30 (H29)	jhk	北宋版	開**禅**寺版道神足無極変化経巻四	開元神足	1126
31 (H24)	tsu	北宋版	通典巻一	通典巻一	11C
32 (H26)	hos	北宋版	斉民要術	斉民要術	12C 初
34 (H28)	nak	南宋版	華厳経内章門等雑孔目巻一	華厳孔目	1146
35 (H30)	hod	南宋版	法蔵和尚伝	法蔵和尚	1149
36 (H31)	gok	南宋版	後漢書光武帝紀	光武帝紀	1198
37 (H74)	smk	西夏版	妙法蓮華経巻一	西夏法華	1149
38 (H50)	okd	日本写本	小川本金剛場陀羅尼経	金剛小川	686
39 (H54)	wad	日本写本	和銅経大般若経巻二百五十	和銅 250	712
40 (H55)	kmi	日本写本	高山寺本弥勒上生経	弥勒上生	738
41 (H56)	zkd	日本写本	守屋本五月一日経**続**高僧伝	五一**続**高	740
43 (H57)	doh	日本写本	高山寺本大教王経巻一	金剛大教	815
45 (H60)	tzs	日本写本	東**禅**寺版写大教王経巻一	仏説大教	12C
47 (H64)	kss	日本写本	明恵自筆華厳信種義	華厳信種	1221
48 (H66)	kyo	日本写本	親鸞自筆教行信証巻四	教行信証	1224
49 (H58)	jyu	日本版本	寛治二年刊本成唯識論巻十	成唯識 10	1088
52 (H33)	ink	日本書紀写本	岩崎本日本書紀巻二十四	岩崎紀 24	10C
53 (H34)	nto	日本書紀写本	図書寮本日本書紀巻二十四	図書紀 24	1142 頃
55 (H39)	nkk	日本書紀写本	兼方本日本書紀巻二	兼方紀 2	1286
56 (H36)	nkm	日本書紀写本	兼右本日本書紀巻二十四	兼右紀 24	1540
57 (H41)	kcc	日本書紀版本	慶長勅版日本書紀巻二	勅版紀 2	1599
58 (H42)	kcj	日本書紀版本	慶長十五年版日本書紀巻二	慶長紀 2	1610
59 (H43)	kbk	日本書紀版本	寛文九年版日本書紀巻二	寛文紀 2	1669

60 (H37)	k24	日本書紀版本	寛文九年版日本書紀卷二十四	寛文紀 24	1669
61 (H44)	sik	韓国写本	新羅本花厳経卷八	華厳新羅	754-755
62 (H46)	skk	韓国印刻本	晋本華厳経卷二十	古麗華 20	10C
63 (H47)	kyu	韓国印刻本	高麗初彫本瑜伽師地論卷五	初麗瑜 5	11C
64 (H48)	ksk	韓国印刻本	高麗再彫本華厳経卷六	再麗華 6	13C

Table 1: List of sources

3. Data structure of HNG

Currently, the HNG dataset is published in a Git repository: https://gitlab.hng-data.org/HNG/hng-data hosted by GitLab Community Edition. In the Git repository, each source is stored if own folder named <folder code>_<name>. For example, "10_妙法蓮華經卷五（今西本）" is the folder for 「今西本妙法蓮華経卷五」, and "10" is the folder code. Each folder has two subfolders, "cards/" and "glyphs/". The subfolder "cards/" stores images of paper cards. The another subfolder "glyphs/" stores representative glyph images selected from paper cards and cropped out.

Each card is numbered in decimal four digits. Each representative glyph image cut out from a paper card is given an ID composed of a main code and a subcode. The main code is the same as the card number. The subcode is used to distinguish between variants if they exist. If a card (corresponding with a character example of a source) does not have any character/glyph variants, subcode is empty. Otherwise, subcodes "a", "b", "c", ... are assigned to each variant. In addition, the relationship between each ID crossing the source is managed in a table of CSV format.

4. CHISE character ontology

The "CHISE character ontology" is a lightweight ontology developed by the authors for character processing. The CHISE character ontology contains information of characters included in Unicode, and other information for Chinese characters.

4.1. Multiple glyph granularity

As for Chinese characters, in addition to Unicode's unification rules (to define abstract characters of CJKV Unified Ideographs), it has information related to the plural glyph granularity of Chinese characters such as super abstract character, unified glyph, abstract glyph (字體), abstract glyph form and glyph image (字形).

4.2. Multiple Granularity Hanzi Structure Model

In the syntax of Ideographic Description Sequence (IDS) defined in ISO/IEC 10646, only coded ideographs (Chinese characters included in UCS) and radical characters can be used as

terminal components (leaf nodes of a IDS tree). However, theoretically, any component can be used as a leaf node. CHISE can represent and process characters not included in UCS. Therefore, in CHISE, Chinese characters and special components not included in UCS are also available as components of extended IDS represented by the ideographic-structure feature.

CHISE supports inheritance of character definition and CHISE character ontology uses this to represent relationships among different unification granularity, such as abstract character, abstract glyph, and glyph image. If we use abstract characters as terminal components of an IDS, the IDS represents a structure of an abstract character. If we use abstract glyphs, the IDS represents a structure of an abstract glyph. If we use glyph images, the IDS represents a structure of an abstract glyph image. Thus, the extended IDS of CHISE can represent unification coverage (granularity) of a character object (Figure 2).
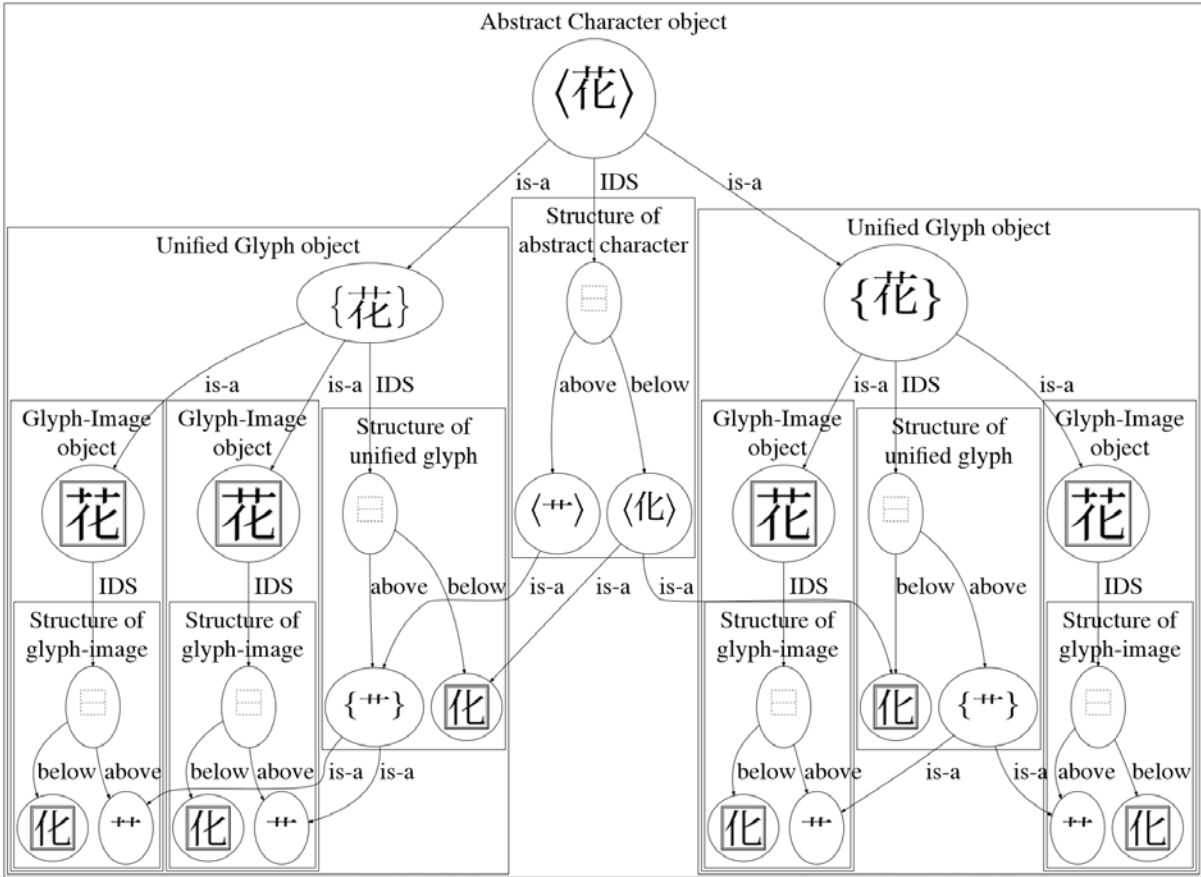

Figure 2: Conceptual graph of multiple-granularity IDS(花)

5. Representation of HNG glyphs in CHISE

5.1. Integration of HNG glyphs into the CHISE character ontology

Several methods can be considered to incorporate HNG information into the CHISE character ontology, but the following method was adopted here: Each representative glyph image of HNG is regarded and defined as a glyph image object of CHISE, and it is attached to any existing abstract glyph form object of CHISE if possible. Otherwise, it is attached to any existing abstract glyph object of CHISE if possible. Otherwise, it is attached to any existing

abstract character of CHISE if possible. Otherwise, it is regarded as new abstract character or new abstract character object will be defined to attach it. In this way, every HNG glyph object can be placed somewhere in the CHISE character ontology.

In addition, if an HNG glyph object to be newly added can be subsumed into an already existing abstract glyph object (If it can be subsumed into an abstract glyph form object, the abstract glyph form object is subsumed into an abstract glyph object. Therefore in the case, the HNG glyph object can be subsumed into the abstract glyph object which subsumes the abstract glyph form object which can subsume it.), there is no need to newly describe Hanzi structure (IDS).

5.2. Encoding of HNG glyph image object

A glyph image object of HNG is represented by a pair of glyph ID (see Section 3) at its source and ID feature of glyph image granularity indicating its source.

In HNG, an ID consisting of three alphanumeric characters is attached to each source. Using this, in CHISE, each source of HNG is represented by ID features with the prefix '===hng-' ("===" is the prefix for glyph image and "hng-" is the prefix of HNG) before the three alphanumeric ID to indicate the source of HNG.

For example, in the case of 「開成石經孝經」, the source ID is "kak", so the ID feature in CHISE is '===hng-kak'.

For the glyph ID, use the value obtained by adding 10 times the card number to the number corresponding to the suffix (0 for no suffix, 1 for suffix "a", 2 for suffix "b", and so on) as the feature value. For example, a glyph ID is "0100", the corresponding feature value is 1000; a glyph ID is "0123b", the corresponding feature value is 1232.

6. Implementation

6.1. Classification of HNG glyphs

Currently, in the CHISE project, we are formulating a guideline to detect unifiable range of glyph or abstract glyph form objects named "CHISE Guidelines for Glyph Granularity of Chinese characters" (CHISE-GGG) (8). In Section 5.1 we have discussed how HNG glyph objects are arranged in the graph of the CHISE character ontology based on Multiple Granularity relations. In the procedure, this guideline is used in detection of unifiability for abstract glyph objects or abstract glyph form objects.

In principle, "IRG Working Document Series (IWDS) 1: List of UCV (Unifiable Component Variations) of Ideograhs" (IWDS-1) (9) is used for determining the unifiable range of abstract characters.

Such relatively rigorous HNG glyph classification work is currently manual worked and requires a lot of labor. For the reason, it is difficult to do with all HNG data at present. In order to obtain the maximum effect from minimum samples, we chose the following three sources: 「今西本妙法蓮華経巻五」(khi), 「守屋本妙法蓮華経巻三」(khm) and 「開成石経論語」(kar). The first two (belong to Early Tang Court Sutras) are sources representing

an early Tang standard, and the third source representing normative *kaishu* glyphs of the Kaicheng Stone Classics (開成石經).

6.2. Integration without classification

To avoid taking a lot of effort, we decided to integrate HNG into CHISE using only the mechanically convertible part from the information of HNG for the remaining parts other than those classified by hand described in Section 6.1.

First, we defined HNG glyphs as glyph image objects of CHISE by the method described in Section 5.2.

Next, linking between the HNG glyph image object and the UCS abstract character object by using mapping information to UCS (or mapping information to "Daikanwa" (「大漢和辭典」 (10))) in HNG. Each link from HNG glyph image object to the corresponding UCS abstract character object is represented by relation feature '<-HNG'. In CHISE, there is a mechanism to automatically generate reverse links for relation features, so the relation feature '->HNG' from UCS abstract character object to HNG glyph object is automatically generated.

However, in order to bring it closer to the original HNG, we are now using the following relation features shown in Table 2 with domain specifiers attached for each source type instead of relation features '<-HNG', '->HNG'.

Source Type	HNG glyph to entry	Entry to HNG glyph
Chinese manuscripts	`<-HNG@CN/manuscript`	`->HNG@CN/manuscript`
Chinese printed books (including Stone Classics)	`<-HNG@CN/printed`	`->HNG@CN/printed`
Japanese manuscripts	`<-HNG@JP/manuscript`	`->HNG@JP/manuscript`
Japanese printed books	`<-HNG@JP/printed`	`->HNG@JP/printed`
Korean sources	`<-HNG@KR`	`->HNG@KR`
Other sources	`<-HNG@MISC`	`->HNG@MISC`

Table 2: Relation features between HNG glyph image objects and entry objects

By this method, when displaying a CHISE-wiki page for a UCS abstract character, it is now possible to display unclassified HNG glyph image objects as the same as classified objects.

However, in the method, the search target of CHISE IDS Find is limited to entries characters (may be UCS abstract characters) of HNG glyphs, and it is not possible to search HNG glyphs which have different structures from their entries characters.

Conversely, in the three sources that integrated with the classifications described in Section 6.1 (「今西本妙法蓮華経巻五」(khi), 「守屋本妙法蓮華経巻三」(khm) and 「開成石経論語」(kar)), each HNG glyph image object knows own Hanzi structure even if its structure and abstract glyph are different from its entry character. For example, by searching Chinese

characters having 「十」 and 「刀」 as components in CHISE IDS Find, you can find 「切」 (U-0002D0C4) in the search result and you can see 「切」 in its CHISE-wiki page.

6.3. **Web applications**

CHISE-HNG IDS Find (CHISE-IDS HNG 漢字検索) (11) [Figure 4] is a Web service to search Chinese characters included in the HNG dataset. It is like CHISE IDS Find (CHISE IDS 漢字検索), if a user specifies one or more components of Chinese characters into the "Character components" window and runs a search, the characters that include every specified component are displayed. However, unlike CHISE IDS-Find, only Chinese characters with examples of HNG are displayed.

**Figure 3 CHISE-HNG IDS Find**

In a results page of the CHISE-HNG IDS-Find (or CHISE IDS-Find), the first column of each line indicates an entry for a character. It has a link for a EsT page [Figure 5] to display details of the character. In the EsT page, links for HNG examples are displayed in the "→ [HNG] …" field. (cf. Table 2)

昇 *昇*

結合：

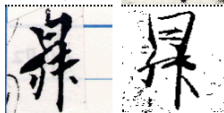- **+ U+E0100 :** 昇
- **+ U+E0101 :** 昇
- **+ U+E0102 :** 昇

部首：日部 (R072)

漢字構造：⬚ 日升
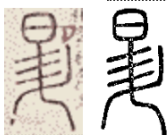
= UCS：U+6607 (26119) [-] [+]

→ [HNG] 中国写本：

→ [HNG] 中国版本：

→ [HNG] 日本写本：

→ [HNG] 日本版本：

→ 説文小篆：

→ 包摂：昇

古典中国語形態素用例：彥昇 昇 昇州

東洋学文献類目用例（題名・キーワード等）：

- 陶鈍(au)「上昇」, 1949年10月

Figure 4: A sample CHISE-wiki (character object viewing in EsT) page (昇)

7. Conclusion

This report outlined the integration of the HNG dataset and the CHISE character ontology, and also introduced the outline of the HNG dataset.

By realizing this integration, it became possible to browse information of HNG via CHISE Web services such as "CHISE IDS Find" or "CHISE-Wiki" ("EsT"), and we were able to strengthen the search function for HNG data. In addition, by realizing "CHISE-HNG IDS-Find" that limits the search range of CHISE IDS Find to the range of HNG data, it has become possible to search for glyph examples of HNG having common components.

References

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

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

3. *HNG dataset*. https://gitlab.hng-data.org/HNG/hng-data.

4. *Character Database of Digital Rubbings*（拓本文字データベース）. http://coe21.zinbun.kyoto-u.ac.jp/djvuchar.

5. *CHISE Project.* http://www.chise.org.

6. Morioka, Tomohiko, *CHISE IDS find*（CHISE IDS 漢字検索）. http://www.chise.org/ids-find.

7. *Bibliography of Oriental Studies database* （東洋学文献類目検索 [第 7.4 α 版]） http://ruimoku.zinbun.kyoto-u.ac.jp/.

8. 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.

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

10. Morohashi, Tetsuji (諸橋轍次) et al. *Dai Kanwa Jiten* (大漢和辭典). Tokyo: Taishūkan.

11. Morioka, Tomohiko. *CHISE-HNG IDS find* （CHISE-IDS HNG 漢字検索）. http://www.chise.org/hng-ids-find.