Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по Стандартизации

Doc Type:	Working Group Document
Title:	Review of Studies on WS2021–00001
Source:	Kushim JIANG (姜兆勤)
Status:	Individual Contribution
Action:	For consideration by SAT and IRG
Date:	2023–09–15

The document studies the decisions related to the character WS2021–00001. The materials used include:
本文概述 WS2021–00001 的相关议题，所选取的材料包括：

- Online Review Tool, WS2021–00001 <hc.jsecs.org/irg/ws2021/app/?id=00001>.
- Eiso CHAN, Arthur NG Hou Man (2023). *Comments on how to handle the strokes in the digitized running text*. IRG N2581 Eiso & Arthur Feedback.
- Selena WEI, Conifer TSENG, Chanhuan LIU (2023). *Comments on "𠂉" is a character or component*. IRG N2581 TCA Feedback.
- SAT (2023). *SAT Feedback to IRG WS2021 v4.0 (IRG N2581): Frequently asked questions about WS2021–00001*. IRG N2581 SAT Feedback.

1 Ideographs, radicals, components, symbols and strokes

The first part analyzes the treatment of CJK ideograph, radical, component, symbol, and stroke in ISO/IEC 10646, and discusses the treatment in various coded character sets.
第一部分辨析 ISO/IEC 10646 对 CJK 的 ideograph、radical、component、symbol 与 stroke 的处理，兼论各地编码字符集对五者的处理。

The encoding of CJK radicals began in 1997 with TCA's proposal on Kangxi Radicals to provide support for TCA–CNS 11643. Taking the 1992 version of TCA–CNS 11643, X 5012, section 5.1 "Symbols, Letters, Radicals", it states that, "There are 213 radicals in Chinese, with two radicals '夂' and '夊' unified to '夂', and this part is separated from the "ideographs in the first plane", indicating that this edition of TCA–CNS 11643 treats ideographs and radicals separately. Related documents are also listed:
对 radical 的编码始于 1997 年 TCA 的 Kangxi Radicals 提案以提供对 TCA–CNS 11643 的支持。取 1992 年版 TCA–CNS 11643, X 5012 第 5.1 节"符号、字母、部首"指出"中国文字部首 213 个，'夂夊'两部首同归于'夂'部中"，并将该节与"第一字面之字集"独立开来，说明该版 TCA–CNS 11643 将 ideograph 与 radical 分别处理。具体编码涉及文件：

- TCA (1997). *Proposal to Add 210 Kangxi Radicals and 3 Hangzhou Numbers in BMP for*

Compatibility. IRG N202.

- V. S. Umamaheswaran, Mike Ksar (1997). *Unconfirmed Minutes of WG2 #32, Singapore (1997, Jan. 20 to 24)*. WG2 N1503, section 8.1 (from page 24).
- John Jenkins, Wang Xiaoming (1997). *Naming of Kangxi Radicals*. IRG N449.

And the Radical Supplement block was further established, involving documents:
并进一步建立了 Radical Supplement。涉及文件：

- IRG Ad-hoc Group (1997). *IRG Radicals Definition*. IRG N407, IRG N407A, IRG N407B: "According to the definition in IRG N310, CJK Radicals are those ideographic components listed in index pages of *Kangxi Dictionary*, *Dai Kanwa Jiten*, *Dae Jawon* and *Hanyu Da Zidian*."
- V. S. Umamaheswaran (1999). *Minutes of WG2 #36, Fukuoka, Japan (1999, Mar. 09 to 15)*. WG2 N2003, section 6.1.1 (from page 16).

According to the documents above, the radical is characterized in that, one, the radical is used specifically in the dictionary index rather than body text, and two, the encoding of the radical is not undergoing a unification process and may be regarded as encoding the actual shapes.
依据上述文献，radical 的特征在于，其一，radical 特别地用于字典的索引部分而非正文部分；其二，radical 的编码不经过认同过程，可视作对 actual shape 编码。

The encoding of CJK strokes began in late 2001 with Resolutions of IRG #18 (IRG N880), which required that submissions of Extension C be divided into two parts, with clearly disunifiable characters in Extension C1, and otherwise in Extension C2. The repertoire by China (IRG N891) contains 〇 (XC100176), and the repertoire by HKSAR (IRG N893) contains 𠂉 (HK–8840), 𠃌 (HK–C879), 𠄌 (HK–8844), 𠃍 (HK–8846), 𠃎 (HK–8849), 𠃏 (HK–884A), 𠃑 (HK–884D), 𠃒 (HK–884F), 𠃔 (HK–8850), 𠃕 (HK–8851), 𠃖 (HK–8852), 𠃗 (HK–8854), 𠃘 (HK–8855), 𠃛 (HK–8841), 𠃜 (HK–8842) and 𠃝 (HK–8843), and the repertoire by ROK (IRG N896) contains 𠃋 (K5H00763). MSAR in IRG N927 considers that two strokes 𠃊 and 𠃍 were still missing from Extension C1. Note that at this time (late 2001 to early 2002) IRG generally considered that the remaining CJK strokes should be encoded in Extension C1.
对 stroke 的编码始于 2001 年末，其中 IRG N880 要求将 Ext. C 分为两部分，与 SuperCJK 显著不可认同的字符置入 Ext. C1，其余字符置入 Ext. C2。G 源、H 源和 K 源均提交了一些 stroke。M 源的 IRG N927 指出尚有两个 stroke 未包含在当前的 Ext. C1 中。注意到此时（2001 年末至 2002 年初），IRG 基本认为 stroke 应收入 Ext. C1。

Since late 2002, there have been proposals to deal with PUA part of HKSCS and GB 18030 (Goldsmith & Muller, L2/03–411). Subsequently, Michel *et al.* (WG2 N2807R = L2/04–161R2) analyzed the PUA part of HKSCS and pointed out that "Although there are already collections of various CJK fragments (such as CJK Radicals Supplement, Kangxi Radical) and methods to describe their arrangement using the IDCs, these 'stroke' elements stands on their own merit as an interesting mechanism to describe CJK characters and corresponding glyphs [Kushim: Glyphs = actual shapes.]" and "... [Extension C] is not yet mature, but at the same time removing characters from the PUA is urgent. A better solution seems to create a new block containing these CJK Basic Stroke characters ...". Immediately following this Bishop & Cook (L2/04–221) proposed to fill the block with all the strokes to satisfy the needs of Wenlin database. Finally, after IRG #25, Dr. Lu proposed the final version of the stroke block in IRG N1180 = WG2 N3063 = L2/06–212, noting that "Representative forms of some proposed CJK Strokes are similar in appearance to representative forms of some single stroke CJK Ideographs currently encoded in various UCS blocks. However, single–stroke CJK Ideographs do not have the properties of CJK Strokes, and single–stroke CJK Ideographs may in some cases exhibit a range of variation in their representative glyphs which conflates necessary distinctions for the

CJK Strokes block. For example, the proposed CJK Strokes ㇒ vs. ㇓ are conflated in the representative forms currently used for Kangxi Radicals and CJK Ideographs ... The precise definition of the CJK Strokes block and clear differentiation of it from the blocks of CJK Ideographs and Radicals serves an essential purpose in the indexing and collation of encoded and unencoded CJK Ideographs and Radicals ..."
从 2002 年末起，开始有提案处理 HKSCS 和 GB 18030 的 PUA 部分（Goldsmith & Muller, L2/03–411）。随后 Michel 等（WG2 N2807R = L2/04–161R2）在分析 HKSCS 的 PUA 部分时指出“尽管已有 component 的 block 与使用 IDS 描述其构形的方法，但 stroke 仍作为一种有趣的机制描述字形”，“Ext. C 尚未成熟，但从 PUA 中移除字形是当务之急，更好的解决方案似乎是创建一个包含这些基本 stroke 的新 block 并将其从 Ext. C 中移除”。紧接着 Bishop & Cook (L2/04–221) 提出在该 block 中填入全部 stroke 以满足文林数据检索的需求。最终，经 IRG #25，陆老师在 IRG N1180 = WG2 N3063 = L2/06–212 中给出最终版本的 stroke block，指出：“一些 stroke 的字形与现有的多个 block 中出现的单笔画 character 相似，但这些 character 无 stroke 的属性，且 character 的 unification 在 stroke 中可能需要 disunification，如 stroke ‘㇒’ 与 stroke ‘㇓’ 被 unify 成 character ‘㇓’。stroke 的确切定义与合理认同对于 character 的整理与索引起着重要作用。”

The encoding of components has only been documented in IRG #48 and IRG #56, not yet in WG2 (the latest component repertoire has been significantly reduced by the introduction of IDC subtraction), and the encoding history of symbols (e.g., 〇) cannot be verified (it can only be basically determined that it is related to the grasp in original CCS). In summary, the encoding history of radicals and strokes are relatively clear, suggesting that the encoding focuses on indexing (radical appears only in the index, and stroke appears only when describing strokes, and carries the meaning of the component semantics or stroke semantics without any other) and unification differences (the unification process of radical and stroke is different from that of characters).
对 component 的编码仅在 IRG #48 与 IRG #56 上有相关文件，尚未进入 WG2（最新的 component repertoire 因 subtraction 的引入而显著缩减），对 symbol（如“〇”）的编码历史无法核实（仅能基本判断其与编码字符集对这些符号的汉字性的把握相关）。综上所述，radical 与 stroke 的编码历史比较清晰，说明对 radical 和 stroke 的编码注重于索引功能（radical 仅在字书检索部分出现，stroke 仅在说明笔画时出现，且仅携带构件或笔画含义而无其他任何语义）与认同差异（radical 和 stroke 的 unification 过程与 character 存在差异）。

In addition, in GB/T 2312, radicals used for indexing (which can be identified as radicals because their pronunciation is not related to the characters, e.g., *xúnzìtóu*), are scattered in the L2 repertoire. In HKSCS, although the characters for stroke are not scattered in the character part, these strokes are not reproduced in the character part, nor are the single–stroke characters in the stroke part. In TCA–CNS 11643, Kangxi radicals are not scattered in the character part, and are noted in Appendix 3 as “Kangxi radicals”.
此外，GB/T 2312 中，用于索引的 radical（能够确认其为 radical 是因为在 GB/T 2312 的音序索引部分，这些 radical 仅有部首名对应的读音，如“寻字头”）散见于二级字表中。HKSCS 中，虽然作为 stroke 的字符不散见于汉字部分，但这些笔画不在汉字部分重出，单笔画汉字也不在笔画部分重出。TCA–CNS 11643 中，Kangxi radicals 不散见于汉字部分，附录 3 指出其为“康熙部首”。

2 Character under discussion

The second part discusses WS2021–00001, for this purpose we list all the evidences that we can get:
第二部分讨论 WS2021–00001。为此列举全部 evidence：

[The following evidences are provided by SAT Database]

- 慧琳《一切經音義》Vol. 54：“更無  孑。上（姜按：即  ）音惟。鄭箋毛詩：‘  ，忘也。’鄭注：‘禮記云：“遺，猶脱落也。”’說文云：‘從辵  （姜按：實爲賁）聲。’賁，正貴字也。下（姜按：即孑）音結。毛詩：‘靡有孑遺’也。傳曰：‘孑然，遺失也。’說文：‘無右臂，從了，象形。’，聲也，音厥。”
- 湛叡《華嚴演義鈔纂釋》Vol. 57：“丨此爲挑。挑爲趯。須存鋒勢而出。此爲策。斫筆背發而仰收。丿此爲掠。筆鋒左出而須和。”

[The following evidences are provided by Tao Yang]

- 《說文字通》Vol. 14：“孑。無右臂也。從了、，象形。居桀切。”
- 《漢語大字典》：“提。……汉字书法术语。……现代汉字的笔形之一，即挑‘’。”
- 《通易西遊正旨分章注釋》：“心字四筆，不入丨丿乀一丨之内。”
- 《顔李遺書・小學稽業》：“永字八法，側丶勒一努丨趯亅策掠丿啄ノ磔乀。”
- 《玉定金科特宥輯要》：“其他用△以清例。用以分條。概照罰字程規。兹不贅述。”“其長條、中問列款多者。每款用一△圈款完用一勾。”
- 《漢溪書法通解》：“，虎牙勢。虎牙之法，‘𠂇 王’等旁用之。”“，金錐勢。金錐之法，‘扌氵’等處用之。”“盖六書起一成文、衡為一、從為丨、邪為、反為乀、至乀而窮……”

[The following evidences are provided by Eiso Chan]

- 《道法會元》：“又加□，又加座子，又用碧虚至道，復用三，蓋塗之。”

Note that, the form is used as a stroke or symbol in all the evidences except 一切经音义 and 说文字通, and thus these evidences are unacceptable. Therefore, it is necessary to discuss only the evidences given in 一切经音义 and 说文字通, both of which analyze the shape of the Shuowen head character  (WG2 N5209: U+3AADC) in Small Seal script (shape system), and therefore need to discuss the correspondence between Kaishu characters and Small Seal characters in *Shuowen*. The meanings of Shuowen head character in ten versions are shown below.

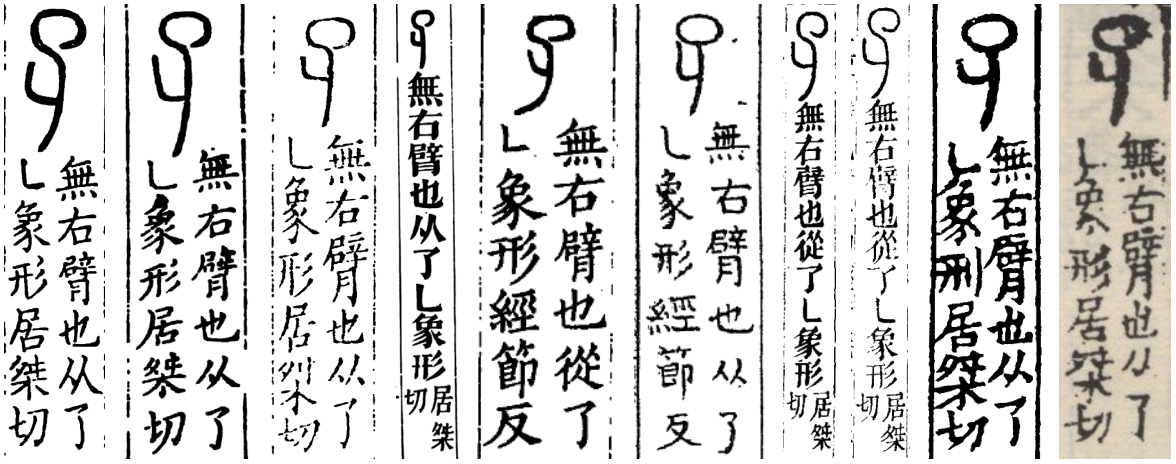
注意到除《一切经音义》与《说文字通》以外的全部 evidence 中形均用作表示 stroke（提、策、挑）或 symbol（三清符），从而这些 evidence 均不可接受。因此只需要讨论《一切经音义》与《说文字通》给出的 evidence。这两份 evidence 都在分析小篆构形系统中的小篆字头 (WG2 N5209: U+3AADC) 的构形，因此需要讨论《说文解字》释文部分中楷书单字与小篆单字的对应关系。各本《说文解字》[1] 给出的小篆字头 对应的释义如下所示。

There are three layers of issues here. First, the character 了 in *Shuowen* explanation part is a Kaishu character, but is the character in Shuowen explanation part a Kaishu character or a Small Seal character? Second, the Kaishu character 了 corresponds to Shuowen head character (WG2 N5209: U+3AADB), but no matter is a Kaishu character or a Small Seal character, is its corresponding Shuowen head character (WG2 N5209: U+3AADB), (WG2 N5209: U+3A4C4), or not included in *Shuowen*? Third, does the fact that *Shuowen* takes the cognition type of as *Xiangxing* rather than *Xingsheng* affect whether (if it is Kaishu) is a character or a stroke?

这里有三层问题。第一，释文部分的“了”为楷书单字，但“”是楷书单字还是小篆单字？第二，楷书单字“了”对应小篆字头 (WG2 N5209: U+3AADB)，但无论“”是楷书单字还是小篆单字，其所对应的小篆字头是 (WG2 N5209: U+3A4BD) 还是 (WG2 N5209:

[1] 說文解字（藤花榭本, THX version），說文解字（汲古閣本），說文解字（孫星衍平津館叢書本），說文解字(陳昌治本），說文解字繫傳（祁寯藻刻本），說文解字繫傳（述古堂景宋寫本），說文解字義證（連筠簃叢書本），說文解字義證（崇文書局本），說文解字（續古逸叢書本），說文解字（汪中藏丁晏跋宋刻元修本）.

U+3A4C4)，还是该字不对应小篆字头？第三，《说文解字》将 𠂇 的理据类型定为象形而非形声，是否影响“𠂆”（如果是楷书单字）是 character 还是 stroke?



The first issue is related to the versions, where Small Seal can appear in the explanation part (e.g., 主), but the existence of two versions using the Kaishu form 𠂆 (e.g., the fifth, seventh and eighth version) is sufficient to confirm the existence of Kaishu character 𠂆 in Shuowen.
第一层问题与《说文解字》的版本相关，释文部分的理据部分可以出现小篆单字（如“主”），但只要存在两版《说文解字》使用楷书形式的“𠂆”（如上图左五、右四、右三）即可确证楷书单字“𠂆”在《说文解字》中的存在性。

The second issue is related to the cognition analysis. One viewpoint is that, *Shuowen* uses “A, contains B” to establish the connection between the head character and the head characters as components, and therefore 𠂆 corresponds to Small Seal head character because the explanation “孑, contains 了𠂆”. According to 说文解字系传, “乚, hook mark … 𠂇 and 成 contains this …”, 𠂆 corresponds to the Shuowen head character 乚. Another viewpoint is that, 𠂇 “has no right arm”, means that if 𠂆 corresponds to head character, its cognition should be “left arm”. Since no head character takes “left arm” as its cognition, it can be said that 𠂆 corresponds to Small Seal components, but it cannot be said that 𠂆 corresponds to Small Seal character.
第二层问题与理据分析相关。一种观点认为《说文解字》使用的“从……”建立了小篆字头与作为构件的小篆字头的联系，因此由“孑”释义“从了𠂆”说明“𠂆”对应小篆字头，采用《说文解字系传》的“乚，鉤識也……𠂇成從此……”观点，“𠂆”对应的小篆字头为 乚。另一种观点认为 𠂇“无右臂”说明“𠂆”若对应小篆字头，则其理据应为“左臂”，由于无小篆字头理据为“左臂”，因而可说“𠂆”对应小篆构件，但不可说“𠂆”对应小篆字头。

The third issue is related to the shape system in *Shuowen*, because:
第三层问题与《说文解字》给出的构形系统相关。由于：

> *Shuowen*, as a self–contained system (explaining *Shuowen* by *Shuowen*), for all the components with clear cognition, whether it is a single character or not, a place will be assigned in the book for its explanation. Most of single–stroke components are not complete characters, do not correspond to words (or morphemes) in the language, and have no actual pronunciation and meaning in use. The explanation of *Shuowen* for these single–stroke components is in fact an explanation of the function when they are contained in other Small Seal characters.

> 《说文》作为一个自足系统（用《说文》解释《说文》），对于每个有明确构意的构件，无论是不是一个字，都会在书中安排一个位置，对其进行说解。绝大多数单笔构件不是一个完整的字，不对应语言中的词（或语素），没有实际使用的音义。而《说文》对于这些单笔构件的意义说解，其实就是对其在组构其他汉字时所体现的构意的说解。
> —— 胡佳佳、樊雨婷《〈说文解字〉对于单笔构件的构意分析》

It is also noted that, the single–stroke components in *Shuowen* ("  ... where refers the ground", "  ... where refers the *Tao*", "  ... where indicates the shape of a hairpin") do not reappear in single–stroke head character, so it should be assumed that *Shuowen* does not distinguish between the Small Seal components and the Small Seal characters, and thus between the Small Seal stroke and the Small Seal character. When there is a clear correspondence between Small Seal and Kaishu, it can also be argued that *Shuowen* does not distinguish between the Kaishu component or Kaishu stroke and Kaishu character, and thus Kaishu character 𠂢 should be regarded as a character.
同时注意到《说文解字》中单笔构件（"  ……一，地也"、"  ……一，道也"、"  ……一以象簪也"）与单笔字头（一）不重出，从而应认为《说文解字》不区分小篆 component 与小篆 character，进而不区分小篆 stroke 与小篆 character。当小篆与楷书有明显对应时，也可以认为《说文解字》不区分楷书 component 或楷书 stroke 与楷书 character。由此楷书单字"𠂢"应视作 character。

At this point, if we analyze the evidences in 一切经音义 and 说文字通, we can see that the status of Kaishu 𠂉 is exactly the same as that of 𠂢 in *Shuowen*, which corresponds to the "left arm" in the head character  , and both of them are broad *Liding* and strict *Liding* of the same Small Seal component. According to the latest unification rules, 𠂉 and 𠂢 or 亻 should be unified in this sense at the level of character.
此时分析《一切经音义》与《说文字通》的 evidence，则可见楷书"𠂉"在《说文解字》中的地位与"𠂢"完全相同，都对应于小篆字头 中的"左臂"，二者是同一个小篆构件的宽式隶定与严式隶定。按最新的 unification rules，"𠂉"与"𠂢"或"亻"在这个意义上应在 character 的层面上进行 unification。

3 Comments on the discourse

The third part discusses all the opinions in the selected material.
第三部分讨论所选材料中的全部意见。

For this purpose, it is important to first clarify part of statements by Eiso CHAN and Arthur NG Hou Man that, "... we feel the encoding principles of component/stroke seem to be needed to clarify ... (1) Reject more abstract CJK stroke in future. (2) For the abstract component, there must be more than one Hanzi which is composed by the component. (3) If any expert can provide the running text as the evidence for one component and it is commonly used in some areas, we can consider treating it as a special CJKUI, but it is not encouraged. (4) It should be used rarely in the daily life."
为此须首先廓清 Eiso CHAN 与 Arthur NG Hou Man 的部分论述，"对于 component 和 stroke 的编码原则应为：（1）拒绝编码任何 CJK abstract stroke；（2）应编码出现于一个之上汉字中的 abstract component；（3）若 abstract component 在 running text 中单独且普遍使用，则可视作 character 但不鼓励；（4）abstract component 应为罕用的。"

This statement was not immediately clear to participants in IRG #60, and thus it is important to note here that, the "abstract stroke" and "abstract component" here are those with specific functions, specific usage scenarios, and specific principles of unification rules as described in **Part 1** of this document, and if WS2021–00001 is analyzed as a single–stroke character (or encoded as a encoded stroke), it is outside the scope of these principles. Thus, the principles for components and strokes has little to do with whether WS2021–00001 should be analyzed as a character or a stroke.
在 IRG #60 中参会者未能立即明晰地把握这一论述，从而须在此指出，abstract stroke 与 abstract component 是本文第一部分所述的具有特定功能、特定使用场景和特定认同原则的 stroke 和 component。而若 WS2021–00001 被分析为 single–stroke character，则其不在这些原则的范围中。因此对于 component 和 stroke 的编码原则的设立与将 WS2021–00001 分析为 character 还是 stroke 关系不大。

Then, we discuss the opinions that are not related to Small Seal head character and cognition analysis. For one, Toshiya Suzuki claims that it is doubtful for 慧琳 to claim 𠂉 is a phonetic part, because the more common expression should be "contains A, pronounced as B" rather than "takes A shape, B is phonetic part. In this regard Wang Yifan claims that, in 慧琳 analysis, the *Xiangxing* function does not conflict with the *Xingsheng* function, c.f., 主，履，盾." We agree with Wang Yifan that it is common for *Xiangxing* components to behave in *Xingsheng* situation.
然后讨论与小篆字头和理据分析无关的意见。其一，Toshiya Suzuki 称："慧琳称'𠂉'为声符可疑，因为一般体例为'从 × × 声'而非'从 × 象形，× 声也'。"对此 Wang Yifan 称："在慧琳分析中，构件的象形功能与形声功能不冲突，类似如'主'、'履'、'盾'等。"认同 Wang Yifan 说，象形构件亦声的情形是常见的。

Secondly, Ken Lunde said that, the encoding of single–stroke character in Extension B should not be regarded as valid precedents for encoding a stroke in CJKUI, but should be regarded as mistakes. Any character can have a reading / name, just they cannot be documented in Unihan database, their readings / names are provided in *Core Specification* Appendix F. To this, Wang Yifan said, the logic lies in first recognizing it as a character rather than a stroke, and then encoding it as a character. According to 说文解字系传, 𠂉 is the variant of 𠃌, and please note the distinction between reading and pronunciation, the reading / name of stroke 𠂉 is *tí*, and the pronunciation of character 𠂉 is *jué* (according to 慧琳音义). The same example is 一 *héng* and *yī*. We agree with Wang Yifan, and pointing out that Ken Lunde's "treating them as mistakes" has a premise that examples in Extension B like 𠃍 and 𠃌 should be regarded as abstract strokes, but have been incorrectly encoded as single–stroke characters. In fact, 𠃍, 𠃌, etc. were included in 汉语大字典, 康熙字典, etc., because of their variant relationship (e.g., 𠂇 ~ 肱, 𠃍 ~ 隐, etc.), or because they follow the treatment of *Shuowen* (𠃌 is used as a marker, and 𠃊 is the inverted tip of a hook, both of them are cognition / function). Thus, there may be incorrect encoding decisions in Extension B, but they should be unrelated to these single–stroke characters.
其二，Ken Lunde 称："Ext. B 中包含 single–stroke character 不能视作将 stroke 收入 CJKUI 的先例，而应将其视作错误。stroke 也可以有 reading / name，只是不能记录在 Unihan 中，其 reading / name 参见 *Core Spec* Appendix F。"对此 Wang Yifan 称："逻辑在于先认定为 character 而非 stroke，再将其作为 character 收入。采《说文解字系传》，'𠂉'是'𠃌'的 variant。另请注意区分 reading 与 pronunciation，作为 stroke 的'𠂉'的 reading / name 是'提'，但作为 character 的'𠂉'的 pronunciation 依据慧琳《音义》为'厥'。如同作为 stroke 的'一'与作为 character 的'一'。"认同 Wang Yifan 说，并指出 Ken Lunde 的"将其视作错误"说法本身可能带有前提，即认为 Ext. B 中的示例如"𠃍"、"𠃌"等仅是 abstract stroke，但被错误地把握为 single–stroke character，进而被编码。事实上"𠃍"、"𠃌"等是因其见于《汉语大字典》、《康熙字典》等而被收入的，

这些字书收入“乚”、“亅”等或因为其有异体关系（如“厶”同“肱”、“乚”同“隐”等），或因为其从《说文解字》的体例（如“亅”为用作标记的符号，“冂”为钩的倒尖，均为构义）。因此 Ext. B 中可能确有错误的编码决策，但应与这些 single–stroke character 无关。

Thirdly, WANG Yifan informatively states that the design for character and stroke should be different. This is correct, but it should be noted that this is not mandatory, and that the design in Kaishu may be the same.
其三，WANG Yifan informatively 称：“character 与 stroke 的设计应不同。”无误，但应指出这并非强制性的做法，在楷体 typeface 中 character 与 stroke 的设计很可能相同。

Fourthly, Eiso CHAN informatively stated that, the Script attribute of U+31C0 is Common, but the Script_Extensions attribute is Han. Eiso CHAN and Arthur NG Hou Man informatively stated that, the behavior of stroke in the literature is exactly the same as character. And there is a problem with Microsoft's handling of stroke's line feed." We agree with these opinions, and consider that setting the Script attribute of CJK Strokes to Common is doubtful. Setting CJK Strokes to Common is used to include the strokes of other ideographs (Tangut script, Khitan Large script, Khitan Small script, Jurchen script, etc.), but the strokes specific to other ideographs are not accepted for inclusion in CJK Strokes (as far as we can tell). The block name CJK Strokes seems to be bound to CJK ideographs, so we think the Script attribute of CJK Strokes should be changed to Han instead of Common.
其四，Eiso CHAN informatively 称：“U+31C0 的 Script 属性是 Common，但 Script_Extensions 属性是 Han。”Eiso CHAN 与 Arthur NG Hou Man informatively 称：“文献中 stroke 的 behavior 与 character 完全相同。Microsoft 对 stroke 的移行处理有问题。”无误。我们还认为 CJK Strokes 中字符的 Script 属性为 Common 是可疑的。将 CJK Strokes 设定为 Common 是用于囊括其他表意文字（西夏文字、契丹大字、契丹小字、女真文字等）的 stroke，但其他表意文字特有的 stroke 又不被接受编入 CJK Strokes 中（印象中）。且其 block name 为 CJK Strokes 似与 CJK ideograph 绑定，从而我们认为 CJK Strokes 的 Script 属性可能应修改为 Han，而非 Common。

The discussion then moves on to the discourses related to the Small Seal head character and cognition analysis. There are three layers of issues here.
接着讨论与小篆字头和理据分析相关的众多论述。这里有三层问题。

The first issue is that Toshiya SUZUKI claims that, the cognition of 亅 can be analyzed as the “left arm”, but the Kaishu 𠂉 cannot be analyzed as the “left arm” of 孑, so 慧琳 is analyzing the cognition of Kaishu 𠂉. But other discussions think 慧琳 is borrowing Kaishu 𠂉 to cite the cognition of Small Seal head character  . So the issue is, when 慧琳 cites *Shuowen* to describe the cognition of 孑, is 慧琳 discussing the cognition of Kaishu 孑 or Small Seal head character ? We believe that the rationale for the single character 孑 in the Regular Script was indeed lost when it was developed from , so we could be lenient and assume that 慧琳 was citing the cognition of the Small Seal head character .
第一层问题是，Toshiya SUZUKI 称“小篆‘亅’的理据可分析为‘左臂’，但楷书‘𠂉’不能分析为楷书‘孑’的左臂”时，说明其在分析作为楷书构件的“𠂉”在相关的楷书构件“孑”中的理据。而其他有关论述均认为慧琳在借楷书字形“𠂉”引用小篆字头 的理据。因此当慧琳引用《说文解字》描述“孑”的理据时，其是在论述楷书单字“孑”的理据，还是小篆字头 的理据？我们认为 发展到楷书单字“孑”时理据确实有所丢失，不如宽容地认为慧琳在引用小篆字头 的理据。

The second issue is that Conifer TSENG claims that, 慧琳 describes 𠂉 as phonetic marker, so 𠂉 is only a marker and not a character. So, the issue is, does the *Xiangxing* or *Xingsheng* determine whether 亅 is a character or not? Here TCA argues that the *Xiangxing*, *Zhishi* or *Xingsheng* function suggests that 𠂉 is not a character. Using the analysis in **Part 2** of this document, we point out that *Shuowen*, as a self–contained

system, takes all the components / head characters as characters, even if the pronunciation and explanation is constructed according to the cognition, and that when there is a clear correspondence between Small Seal and Kaishu, it can be argued that the Kaishu correspondence to each Small Seal head characters should be regarded as characters too.
第二层问题是，Conifer TSENG 称“慧琳称‘ ⺀ ’为声符，因此‘ ⺀ ’仅为 marker 而非 character。”因此采信  的理据的象形说和形声说是否说明“  ”仅为 marker 而非 character？这里 TCA 认为采象形理据认为“”是象形构件（或指示构件），或采形声理据认为“”是声符，均只能使得“”不应被视作 character。利用第二部分的分析，我们指出《说文解字》作为 self–contained 的构形系统，即使每一个小篆字头的音义是按照构字理据增补的，这些小篆字头都应被视作 character。当小篆与楷书有明显对应时，也可认为每个小篆字头对应的楷书单字也应视作 character。

The third issue is whether  corresponds to a Small Seal head character. SUZUKI–san thinks that does not correspond to a Small Seal head character, while WANG Yifan takes the viewpoint of 说文解字系传. We prefer the viewpoint in 说文解字系传, because, according to the analysis in the **Part 2**, the fact that single–stroke components and single–stroke head characters do not overlap in *Shuowen*, which suggests that the explanation part might not have listed all the possible cognition, and that “contains  and ” can be regarded as a statement to split one head character into other head characters. However, we also believe that this can be regarded as an academic controversy.
第三层问题是，“  ”是否对应小篆字头？SUZUKI 认为“”不对应小篆字头，而 WANG Yifan 认为采《说文解字系传》观点“”对应于小篆字头  。我们倾向于采《说文解字系传》观点，因为利用第二部分的分析，《说文解字》中单笔构件与单笔字头不重出，说明小篆字头的释文部分本就可能不罗列该小篆字头可能的所有理据功能，且“从了”可视作将一个小篆字头拆解成其他小篆字头的声明。但我们也认为这可视作一学术性争议。

Some of the remaining questions include: (1) Should we adopt the *Xiangxing* or *Xingsheng* cognition? We prefer *Xiangxing*, but we also think that this can be regarded as an academic controversy. (2) Is *Xingsheng* trustworthy? We prefer to think that it is not trustworthy, purely because in Middle Chinese, all the characters 孑 have the phonetic component 赦, but the Middle Chinese status for these characters do not include the Middle Chinese status of 孑. But we also think that this can be regarded as an academic controversy. (3) If the cognition of “left arm” is not independent, should we consider that should not be regarded as character? We think that for the character  , if  (shape like a human) disappears, then the cognition “the ground” for  is not independent either. And if we think corresponds to the head character , then it has another independent cognition “hook mark”. However, this can be regarded as an academic controversy derived from the third issue.
余下的一些问题包括：（1）应该采纳象形说还是形声说？我们倾向于采纳象形说，但我们也认为这可视作一学术性争议。（2）形声说本身是否可信？我们倾向于认为不可信，纯粹因为中古地位除“孓”外其余字的间接声旁均为“赦”，而以“赦”为声旁的字的中古地位的小韵不含“孑”所处的小韵。但我们也认为这可视作一学术性争议。（3）“”作为“左臂”的理据无独立性，进而是否应认为“”不应视作 character？我们认为，的理据若无“象人形”的在，则其 的“地也”的理据似也不独立。且如果认为“”对应于小篆字头 ，则其具有另一独立理据“钩识也”。但继承第三层问题，这可视作一学术性争议。

4 Regular expression

In IRG N2581 SAT Feedback, the author(s) argued that a search tool that utilizes only Script attribute and not Script_Extensions attribute is not able to match the stroke, and therefore encoding the character as a stroke should be avoided. In order to make it possible to encode 𠂉 as a stroke, it is necessary to provide a concise regular expression that can match the stroke.
在 IRG N2581 SAT Feedback 中，作者认为仅利用 Script 而不使用 Script_Extensions 的搜索工具无法选中 stroke，因此应避免将 character 按 stroke 编码。因此，为了使“𠂉”按 stroke 编码成为可能，有必要给出能够选中 stroke 的简明正则表达式。

Take Python 3 as an example, the program framework is:
以 Python 3 为例，程序框架为：

```
import regex

string = "this is the string"
pattern = r"[this is the pattern]"
findall = regex.findall(pattern, string, regex.VERSION1)
```

The key is to design the appropriate pattern so that the characters filtered using the pattern contain CJKUIs, strokes, and radicals, but not punctuation. Here are some alternative patterns.
关键在于设计合适的 pattern，使得利用 pattern 筛出的字符包含 CJKUI、strokes 和 radicals，而不包含 punctuation。以下是一些可供选择的 pattern。

```
pattern = r"[\p{scx=Hani}--\p{P}--\p{N}]"                          (1)
pattern = r"[\p{sc=Hani}||\p{Block=CJK_Strokes}]"                  (2)
pattern = r"[\p{scx=Hani}&&\p{L}||\p{Block=CJK_Strokes}]"          (3)
pattern = r"[\p{scx=Hani}&&[\p{S}||\p{L}]]"                        (4)
```

In summary, we believe that if 𠂉 is treated as a character, it follows the convention of *Shuowen* and should be unified to the encoded character. If 𠂉 is treated as a stroke, then the encoded stroke should be used, and a recommendation for character searching is provided in **Chapter 4**.
综上，我们认为，若将“𠂉”视作 character 编码，则其遵从《说文解字》的惯例，而应与已编码字符认同。若将“𠂉”视作 stroke 编码，则应使用已编码 stroke，第 4 章为相关的字符搜索提供了建议。

(End of Document)