论证“和谐体”的理论优势，科学规劃数码汉字

朱一星*

关键词：和谐体　数码汉字　汉字同步

简目：

0. 前言

本文运用汉字单位，即汉字的理论值概念，分析论证香港中国语文学会倡议和谐体提出的如下两个基本诉求的重要意义：一是统一汉字并不意味着偏废特定的字体字形，即“可简用简，该繁用繁”[1]；二是统一汉字的关键是必须消除两岸的汉字规範在理论值上的不一致，即“一一对应才是重中之重”[2]。

笔者近二十年跟踪观察中国和日本的汉字规範政策以及汉字编码领域的动向，深感汉字系统的整体优化和汉字符号理论是难分难解的整体工程。事到如今，“简单地论定汉字的优劣、简繁、难易的只言片语，已经不能解决问题”。“诸如此类的问题，都要在深入、科学的汉字理论研究中来解决。”（王宁 2009）本文将举出许多理由，说明汉字的规範是不可能在现有的汉字观念中踏碎步了，而应该在理论上大胆提升，技术上不断突破，政策上一步到位。从这一立场出发，本文试图对“书同文”赋予合乎现代符号理论，契合信息化社会要求的解释，提出应该以“汉字同步”的概念，代替多年来“汉字统一”的主张。

许多围绕汉字规範化看起来似乎难以沟通的难题，比如最近在《语讯》106期围绕苏培成、张书岩提出的有关汉字类推简化的问题，其实也只有在“汉字同步”的理论框架中才是一个具有现实意义的合理诉求。

---

* 朱一星先生，日本 京都 京都外国语大学。

1. 编辑部：“识繁写简”不如“可简用简，该繁用繁”《语文建设通讯》92期（下称《语讯》92）。
2. 同注1。

笔者曾经就汉字符号属性做过一些论述，说明了在符号理论意义上如何思考汉字的同一性。（朱一星 2013）。本文可说是姊妹篇，表明笔者讨论汉字符号理论的意图在於解释一些现实当中迫切需要解决的课题。

笔者在已发表的文章中用“汉字单位”表示抽象的汉字符号，这一术语可以解释为汉字作为符号系统的一个共时性的单一取值。“汉字非一一对应”的现象，从汉字符号理论的立场来看，其实就是地区间规範汉字单一符号取值非同步现象的一部分。而现行国际统一码（Unicode）中显露大面积的汉字字符取值非同步的事实，也是汉语使用者在全球範围进行数码信息交换时的障碍。所以说，“非对应”问题，既是一个汉字规範问题，也是一个汉字符号的理论问题和实用技术问题。

在进入正题以前，笔者认为讨论一下相关的理论背景是有益的。因为这有助於我们理解为什么要坚持推进汉字符号系统的整体优化立场。

1．明确定位汉字和语言的关係

近三年前本人曾经针对汉字和语言的关係做了一些思考，但在今天看来，还必须做不小的修正[3]。

当时笔者尝试将汉字实质上视为一个能够标示语义的视觉表达形式，犹如说汉字就是语音语言的另一个可视化的“侧面”。其理由是主张语言行为并非仅限於依靠听觉，当我们看到汉字有赖於视觉如同手语一般时，把汉字视为属於语言的视觉形式似乎才顺理成章。笔者因此主张现代语言学研究领域也能够包括汉字研究。

然而，今天笔者却要主张，汉字系统应该被视为和语言不同的另一个符号系统，而不是语言的另一个侧面。应当说汉字符号系统并列、且服务於语音语言符号系统，这样才能够解释汉字对应不同的语言系统的客观事实。应该说，将文字和语言分别视为两个不同性质的符号系统，更符合近代符号科学的原理原则。

明确这一点并不只是为了契合索绪尔的一句话“对汉人来说，文字就是他们的第二语言”。而是通过更多的思索和实事求是的分析。

文字和语言的最根本区别，是人类语言（母语）的自然性，和文字的人工性。笔者先前虽然也看到了这一点，却完全忽视了其重要性。今天重新检点两者之间的本质性不同，促使笔者有充分的理由认为在讨论语言符号时正确区别这两个符号系统，避免将两者混为一谈。也促使自己能够以截然不同的态度和研究方法来对待语言和文字这两套符号系统。

在人类史上，任何民族部落都拥有属於自己的，族群成员之间一开口便能互相认同的天然母语。任何人都会在幼年不知不觉之中获得这种自然语言。这一点就连听障者也不例外，只要身处一个使用手语的环境，听障者就会在不知不觉中掌握自然手语。

同时，我们也知道并非任何民族，或族群都拥有自己的文字，这是因为文字符号系

3. 刊登在京都外国语大学《研究论丛》No.81 2013，香港中国语文学会《语文建设通讯》105期转载时，没有刊出这一部分。

统并非任何族群都能够自然获得并成为权威性的记载工具系统的，这是人类文明歷史告诉我们的事实。从每个个人体验来说也同样如此，人们不可能在无意之中习得文字及熟练的书写技术。按照乔姆斯基的观点说，语言（母语）是自然“获得”的，而文字则是需要後天“学习”的。所以，将文字视为文明社会的工具产品，而将语言视为人类的自然属性的观点受到普遍承认，应该是令人信服的。

索绪尔对现代语言学提出的一个重要主张，或者说在语言学讲座上教授们常挂在嘴边的一个方法立场是：语言学研究应该是“描写性”的，而不应该是“规定性”的。但是如果我们把这种态度也用来对待汉字，对於这种人工创造的产品性符号系统也只进行“描写性”的研究，只能被动地接受前人留下的符号现象而不能对其进行规範整理发展创新，那就等於是对汉字系统“产品性”的否定，也是对汉字系统可优化思路的否定，那么，也就毫无疑问地等於在终止汉字符号系统的生命力。

明确地认识上述道理，我们就能够推导出这样的结论：文明社会的不断发展，书写工具的不断改变，都会进一步要求汉字这一符号系统不断完善它的工具性和合理性，就会要求其更有效地表现语言（无论在标意性功能方面；还是在标音性功能方面）。这就必然产生对汉字系统进行加工改造的需求和动机。

致力於普及语言文字地位规劃和本体规劃理论的冯志伟指出：“我们在这裡把‘语言规劃’叫做‘语言文字规劃’，多加了‘文字’这个词，是因为在我国通行的汉字是一种非常複杂而重要的文字，在语言规劃中，必须给以特别的重视。”（冯志伟 2000）

语言本体规劃理论，对欧洲语言来说其实是不会在标示语音的字母文字系统上做过多深刻思考的。然而，比拼音文字複杂得多的汉字，则需要適时地进行规劃和整理。况且汉字系统自古以来就是跨方言口语的广域信息交流工具，本质上也需要不断维护保养，不断建设的产品性符号系统。

重新定位汉字符号和语言符号的关係，明确地指出两者间的非同质性，就能赋予我们讨论汉字规範问题的强烈动机，增强加速调整汉字符号整体系统性规範性的信念。

世界著名文明起源地都是以有文字记载而为人类所记忆，并得以积累和延续的，人类文明歷史离不开文字工具。戈登·柴尔德主张人类文明进程的第一步是新石器革命。我们今天还可以说，考虑到人类发现文字概念以及创造文字所具有的决定性意义，称文字概念和文字系统的创造是人类文明史上最伟大，最早期的信息革命都不为过。没有文字，人类科学技术史就无从谈起，也不会有今天（实质上的）第二次信息革命。

下面回到本文正题：求证“和谐体”在汉字符号理论意义上的优势。

2. 和谐体的原则之一：可简用简，该繁用繁

笔者以前分析过，立足於符号理论的立场审视汉字系统时，汉字单位本质上是不存在绝对音值和绝对形值的抽象的概念。也就是说当我们讨论某个汉字的理论上的取值时，实际上不仅包含这个汉字的各种字体。共时性地看当下汉字的话，我们不言而喻地还指称这个汉字单位的各种异体或异形。当然如果这个汉字曾被简化（不论在哪个地

区，不论是被认可的规範字还是手头字），就自然的还包括这个汉字的简化字形。

和谐体的“可简用简，该繁用繁”的初衷是指当多对一简化字[4] 容易引起误会时使用繁体字[5]。

胡百华在《语讯》106期上的说明，简要地表明了和谐体的明确立场：“我们倡议的简繁和谐体，是基於需要，原则上接受不同体系的汉字，……（略）约略地说，在和谐体建议恢復的‘有价值的字’中，恢復使用後需要类推简化的有以下所列举的例字（略）。”可见和谐体的“用繁”是带条件的，是有可能进一步（类推）简化的。这样的“用繁”实质上就是指使用汉字的理论值（不拘泥实际字形）。这样的解释也许在某些人看来有点不得要领：当前尚不存在的（类推）字形如何去要求实现呢?

为了把问题说明清楚，在这裡，还必须理解我们正处在一个什么样的时代——

古代的文字革命和今天的信息革命有一个共同之处，那就是都伴随着符号记载工具的发明创新。在古代是文字概念的形成以及相关的纸笔墨等一系列书写工具的技术发明。而今天，众所周知是伴随着数码符号的新概念形成以及电子计算机、数据联网等相关的信息工具的技术发明。我们今天讨论规範汉字问题，有一个非常重要的前提，就是必须思索汉字在信息时代应该如何使用，必须把我们的研究对象放在信息时代文字发生了什么变化这一背景下来讨论。

什么是信息时代的文字特点呢？这就要从理解数码化的文字信息是如何运作开始。

信息时代的文字符号，其最大的变化就是文字从笔墨书写或排版印刷的图形状态，转化为无数“＋”和“－”的电磁状态，分别用“1”和“0”来表示，这就是文字的数码化，或叫数字化。当还是笔墨文字的时代，书写者必须决定文字的图形性形状，即决定“笔迹”。笔迹也是文字的图形性形状的初值。同时，传递文字信息的载体（兽骨兽皮，石碑木简，纸张布帛等）决定文字的初值基本能够保证在储存和流通过程中维持原样。我们可以说文字革命的符号特点之一就是符号初值的非挥发性，符号的非挥发性使其图形性初值得以半永久地留存。

但在信息时代，数码文字的书写者所指定的符号则是一连串看不见的电磁性信号，我们通过终端机用户平台所选择的一个电脑字模来理解确认我们选择的目标字符。当这个字符一旦离开显示屏，任何图形性痕迹就荡然无存，因为整套行为过程中压根就不曾存在“笔迹”这回事。信息时代对文字符号性质的改变，彻底推翻了传统意义的文字概念，称其为文明史上的第二次文字革命都不为过。

如此看来，一系列工具的进化和技术的成熟所造成的文字概念变化，首先就是“笔迹”概念的丧失，这意味着文字符号的运用者已经不能如同昔日那样自如地掌控字符的图形性形状。当今信息时代的文字一旦脱离“书写”者，进入各种硬盘软盘、有线无线，网上网下，其符号的图形性形状就再也不属於特定的用户平台。而当我们需要读取

4. 严格意义上简化字只是大陆规範汉字的一部分，以下除特别语境，尽量使用“大陆规範汉字”
5. “繁体字”的说法，原指大陆内地的繁体选用字新字形，有时也用来指台湾现行规範字体，两者指称的对象和字形是不同的，这裡指後者。本文以下尽量称“台湾正体字”

这一信息时，就必须通过终端设备特定的用户平台来实现。这时候的“无形”文字才由电脑字模赋予相应的图形性。需要注意的是，在数码环境中，企望文件阅读者的用户平台和该文件原书写者的完全一致，是不切实际的。大多数的用户平台都会有属於自己的默认设置或偏好选择。

一方面，工程师们也会让我们有许多方法能够在某种意义上保持书写者对文字形体的初值“意图”，一般被称作“标记语言”（或“置标语言”“标识语言”）技术。常见的比如有 HTML、XML、XHTML 等等。然而请读者留意，标记语言的工作原理是针对未曾标记的字符（纯文本）而设计的。也就是说，为了在形形色色的用户平台上实现看起来保持原汁原味的字体或状态，就必须保证数码文字是非图形性的纯文本。简言之，标记语言保证的是文字图形性“意图”，而非并不存在的文字图形本身。

数码化文字的重要功能在於异地交换，以及流通在各种终端设备之间，保持可检索等多种用途。第一需求是必须“通行无阻”，其次才是“可视性”。在笔墨时代，文字符号即等于图形性“痕迹”，永远依附於赖以存在的载体上，字形就如同语音语言中的方言方音，每个文字都有不尽相同的笔迹初值。然而数码时代的文字符号，则不是图形性的“痕迹”，并不依附於任何物质性载体，只以电磁信号的形式存在。这就是数码化汉字不仅应该能够兑现成所有可能出现的字体字形，同时又应该能够通行无阻的原理。因为这样才能让信息在瞬息之间跨域转移以及计算机高速检索和各种信息的利用加工。体现出数码文字符号的抽象性超越笔墨文字具体性的优势。

这就是数码时代文字符号现象的一个基本事实：数码汉字的初值不是图形性的，而只是一连串无形的电磁信号，回顾笔墨文字的时代，表现文字初值的物质性纸张文件被称作“原件”或“正本”。但是在信息时代，“正本”很可能仅仅是一串电磁信号，只是在必要时才按照需要印製成有形物质形式。这时候的纸张文件就只不过是相对於电磁性正本的一个副本而已。

这种被称作“无纸化”的信息保存和流通方式，已经在现代社会不知不觉地成为普遍的新常态。比如上海、深圳的股市证券就都已经实施无纸化，证券持有者再也不必像从前那样把股票放在柜子裡或保险箱内。需要证券原件时，只需让证券公司开具一份证明单即可。日常生活中，我们购物，坐车，简单的信函……处处在无纸化的浪潮冲击下。日本在户籍管理领域，已经从上世纪末开始逐步地、作为推进电子政府计劃之一部分，分期分批实施居民户政管理的无纸化。市民通过网络从家裡，或者通过24小时营业的便利商店就能办理取证，登记，纳税等多种手续。

确认以下两点：第一，汉字是工具产品性符号系统，汉字使用者不仅有义务也需要对汉字符号不断进行系统性优化，和设计上的改进，才能称得上是继承了古老的汉字。第二，讨论信息时代的汉字已经不能单从汉字的表层现象来理解，而必须从汉字符号的本质上领会文字的功能和特性。理解承担汉字功能的不再是附着在物质载体上文字痕迹，而是一个抽象的、无形的理论值，即汉字单位。

有足够想象力的文字学家会理解：汉字，这一本质上不具备绝对音值和绝对形值的

标意文字，由於他本身的单字独立性和稳定性（没有连写变异问题），与其说在信息时代面临淘汰，不如说是適得其所更恰当。準确认识新一轮文字革命的本质，我们就能够顺应时代变化，主动把握正确发挥汉字符号系统功能的机会。这一点，笔者的结论和某些“汉字符号学”论者对汉字在信息时代的命运持悲观态度的观点截然相反。

经过如上讨论，再来看看和谐体的“可简用简，该繁用繁”原则，我们就能以较高的境界去认识，和谐体的原则并非是走中间道路，也不是在文字使用层面上把“识”与“写”分离。按照笔者的理解，是在数码时代弱化字体字形概念的一种体认，是积极的超越。

上个世纪，当埃尔温·薛定谔和爱因斯坦讨论恼人的量子力学诠释问题时，提出了一个被称作“薛定谔的猫”的假想试验，用来质疑现实世界中不可能的，死猫和活猫同时存在的“叠加态”是否具有现实性。这一假想试验不但在量子力学领域引发热议，也成了一个让科学家们议论纷纷的哲学命题。

薛定谔猫的假想试验对於今天我们来理解数码汉字的不确定性时，无疑是一个再合適不过的比喻了：当数码汉字处在电磁状态下时，符号理论值无法用具体的图形表示，而当我们想要看看这个符号到底是什么样的那一个瞬间，符号才呈现图形性状态（显示在屏幕上，或打印出来）。从理论上说，如果该汉字单位兼具大陆规範字和台湾正体字的两种标準，那么就可以说这个数码汉字如同薛定谔的猫，是两种状况的叠加态。

3. 和谐体的原则之二：一一对应才是重中之重

前面讨论过，从实质性意义上说，和谐体并非主张恢復某些汉字的图形性“字形”，而是主张恢復一些大陆汉字和少数台湾汉字当中，单独归并了的“理论值”。可惜这一主张的实质内容目前尚未得到学界重镇的理解。比如苏培成明确指出“需要恢復为繁体的只限於那些在转换中容易出现差错的字”（苏培成 2004）。显示对於“一一对应”这个原则并不完全认同。苏培成认为：“简繁转换时常出错，主要是因为从事转换的人不掌握汉字简体和繁体的对应知识。把全部一对多都改为一对一，既无必要又无可能”（苏培成 2011）。

的确，撇开数码化汉字的交换问题，大陆内地多年来的文字生活并不缺少这些字。而且汉字单个字形负载多项音义本来就不是特殊现象，何必难为个别简化字呢？笔者认为这裡还是必须通过对信息社会数码汉字的认识来理解。在这裡我们还必须把问题同另一个较普遍的、今天仍然颇有市场的误解联繫起来讨论。这个误解，就是寄希望於电脑软件的提升来实现汉字的繁简转换。王宁的如下一段话可以说是很有代表性的。

> 目前“一简对多繁”产生的问题，主要是在简繁自动转换和文言文印刷中。前者可以通过语言文字研究与计算机技术的改进得到解决，後者可以通过政策的適当调整，製定文言文印刷用字规範来解决。同时可以尽快启动对简化字优化条件的研究，以便在条件成熟时有计劃地进行全面调整。（王宁 2011）

对繁简汉字“自动转换”存在的幻想，已经成为今天阻碍汉字整体优化的一个心理障碍。还在笔墨文字的时代，人们通过汉字的图形性信息用大脑来识读。经过训练的识读者能够始终如一地正确判断各种字体或字体变异在语意功能上的同一性，甚至包括规範的和不怎么规範的字形。汉字的“笔迹”本来就因人而异，我们如果过分拘泥於细节，就无法利用汉字实现书面交流。

然而，信息时代的字符概念是电磁性信号，而不是显示屏上依靠电脑字模的图形性字形。在同一个设计合理的，编了码的字符集内部，数码字符是不可能在对应问题上，即字符的同一性问题上出错的。需要特殊对待的文本，只是在不同的编码系统之间才成为需要。这种时候，事故的机制大致表现如下：

本来，按数据库的基本原理，某一个关係模式中的任何属项都必须是一个不可分割的单项值，即一个数据属项的定义不可能既是 A，又是 B。这是数据关係模式的最基本的要求（第一範式，或叫第一正规化）。大陆规範汉字和台湾正体字如果完全保持一一对应，那么即使分别属於两个编码字符集，数据的交换也毫无问题。可是当两个字符集的汉字目录（这裡指汉字理论值清单）字数不一致时，就会出现某些属项的定义违反第一範式，这些失去同一性的汉字字符就在数据关係模式中没有着落，使数据库无法运作。这时候汉字编码字符集作为关係模式原本需要的绝对可靠性丧失殆尽。

当今的繁简转换事实上是一个要以脱序字符的方式来交换的人工智能处理程序。比如需要把简化字“复”转换到繁体字（复→複／復）时，处理程序就必须根据上下文可预测範围进行判断。但由於常有不可预测或原本独立的语境，这时处理程序就会束手无措，转换成功率自然也就无法百分之百地保证，这有悖於自动转换本身的存在意义。

以笔者的观点来看，认为两岸的汉字文件交流理所当然地必须求助於语境分析，依靠转换引擎的思维定势，犹如麻醉剂般正在潜移默化地向许多学者的头脑植入了一个毫无根据的“前提”：繁简汉字无法兼容是天经地义的事。尽管这个“前提”是极具误导性的假象，也是一个原本值得克服也能够克服的文字整理规劃範畴的课题。然而许多人至今仍未能勇敢地去挑战数码汉字的系统规範化，宁愿以眼前暂时混乱的假象为前提进行思考。

这就是多年来在汉字编码问题上的一个重大偏误，学者们对於汉字系统本质的理解被汉字看得见的图形性外表字形所绑架，许多人无法透过现象去捕捉数码汉字的本质，并通过基於数码汉字的抽象理論值的思维去探索建立一个正确的汉字编码框架。

如果我们勇敢地放弃对字体字形的刻板认识，转而注重对汉字理论值的追求，前景就将会更加乐观。因为汉字的字体字形虽然庞杂，然而汉字单位则是相当有限的。如果把对汉字的整体规劃聚焦在有限的汉字单位上的话，两岸汉字规範的分歧就小的多。一个实现了统一的汉字单位的数码汉字字符集，就再也不需要複杂而不可靠的转换程序，因为大陆规範汉字也好，台湾正体字也好，都不过是用户偏好设置管辖的任务、属於汉字的字模显示问题。

读者也许已经能够理解，和谐体所诉求的一一对应原则，说到底就是在维护简化汉

字成果的前提下，谋求汉字的实质意义上的统一。多少人长年梦寐以求的书同文目标，其实并非是字形的统一，而是汉字单位（符号理论值）的统一。笔者在2010年提出国际汉字编码应该导入“汉字单位”概念，与和谐体追求的目标是一致的。

笔者认为应该将狭义的、遥遥无期的、着眼於汉字外表字形层面的“汉字统一”目标，修改为与信息时代数码技术吻合的、着眼於汉字实质性理論值层面一致的“汉字同步”。这样就既能保留汉字原本丰富多彩的图形价值，又不失去準确无误地承载传递信息的文字符号功能。这裡不存在文化偏向或政策立场问题，纯粹是一个工程学意义上的科学性合理性问题。

和谐体的“可简用简，该繁用繁”和“一一对应”理念正是实践着让不同汉字使用社会能够基本上保持各自的规範字体字形，延续各地区目前为止被广大民众所接受的，已经习以为常的书写习惯。实现这一目标不仅最大限度地符合所有汉字使用地区的最广大民众利益，也印证了汉字理论值概念在数码时代的重要性，和高度的文明社会要求汉字本体规劃的必要性。

4. 实现和谐体目标的路线图

许多人或许觉得实现上述目标恐怕是“说起来容易做起来难”，下面我们对实现这一目标的可行性做一简单描述。

因为我们正在讨论的对象是数码汉字，这就理所当然地涉及到汉字的编码框架问题，涉及到当前汉字编码的总体设计。这其中有一个颇为专业的技术性问题：如何既保持汉字的理论值不变，同时又满足在不同地区保证显示当地的规範汉字，甚至一些“非规範的”异体字的要求?

文字的编码原理，即是将某个码位分配给一个特定的文字符号。这意味着权威编码机构经审核同意赋予某个文字在数码世界“落户”。原则上当然一个字符只能报进一个户口（符号唯一性原则）。为了达此目的，就需要对数码世界的户口（码位）所能够接纳落户（数码化）的文字资格做精确定义。这种定义显然不能是因人而异，因表面的字形而异的。这就是所谓“字符，而非字形”（Characters, not glyphs）原则，意思是说，能够赋予码位的只能是抽象的字符而不是具体的字形[6]。 然而，对於中 日 韩统一字符集来说，汉字的哪一个层面才能相当於字符，长久以来一直议论纷纷而没有在理论上形成明确的答案，显露出汉字现代符号学理论研究大大滞後於信息革命的先天不足。

如今也许很少有人知道，当年各国编码专家，在汉字问题上陷入困境之中时，日本的两名参与编码工作的工程师则致力於摸索如何从技术上使用“附加标识”向同一码位区别不同字形字体的技术攻关。这一被称作“汉字异形选择符（Ideographic Variation Selector）”的主意，技术上类似於将异体字编号的建议[7]，日後获得标準化组织认可，

6. 参见统一码“第二章：总结构”，参见网页：http://www.unicode.org/versions/Unicode8.0.0/
7. 将异体字编号的建议参见姚德怀 2007。

成为统一码的一个标準技术[8] 而获得正式通过[9]。姑且不管目前统一码协会官方对这一技术的理解如何，这些实实在在的努力指出了汉字的字符抽象性，部分地弱化了人们对纯文本汉字字形的执着，是值得高度评价的。

为了让读者初步了解早年编码实践中，人们如何对待汉字的不同字形，如何试图在编码设计上摸索“繁简对应”，我们在这裡不妨对早年的汉字编码做一简单回顾。

（1）中文资讯交换码 (CCCII)（台湾早期的汉字编码）

说起中文汉字编码歷史，就不得不特别提及台湾的由国字整理小组[10] 研发的中文资讯交换码 (CCCII)。该字符集在1980年当时是中文汉字领域唯一的一个大型编码。它将编码範围从一开始就瞄準“世界汉字”的概念，不仅是台湾正体字，也包括大陆的简化字体，加上日本的简体汉字、国字，以及韩国的汉字。想当年大陆简化字在台湾被视为“匪字”，再想想近年大陆内地还有人提议退简还繁，中文资讯交换码知难而进的精神和远见卓识足以让人钦佩[11]。

最值得注目的是其编码框架设计之科学性，就是将台湾正体字和大陆规範汉字、以及其他同位异体字都从结构上考虑安排在对应的码位上。这样便能够在不同规範字体间交叉检索。这一点就连《中日韩越信息处理》的作者小林剑（Ken Lunde）也称道说其结构的逻辑性是台湾“考虑最周到”的一个编码[12]。中文资讯交换码的设计成功，得到北美图书馆业界的认可，多年来一直是北美许多图书馆和美国国会图书馆对东亚汉字书籍编目的主要工具。

虽然中文资讯交换码因为在编码模式上採用的框架耗费码位（造成许多空档码位），因而缺乏技术优势。另外其大陆的简化字也有不準确之处，但是在当年的客观条件下，应该说已经是难能可贵的了。

（2）大陆汉字编码字符集的基本集（GB 2312-80）和辅助集（GB/T 12345-90）

大陆内地的《信息交换用汉字编码字符集－基本集》（下称 GB0）只是一个简化字的字符集，随後又增加了针对大陆繁体字的《信息交换用汉字编码字符集－辅助集》（下称 GB1）。比照 GB0 和 GB1，可以分析出如下三个不同側面。

第一. GB1 宣称完全对应 GB0，是其繁体字版，然而事实上大部分没有简化的非简非繁汉字实质上在 GB1 中是重複编码。

第二. GB1 对简化字相应码位上安排了大陆规範繁体字。大家知道大陆的规範繁体字是根据《第一批异体字整理表》淘汰了若干异体字後的选用繁体字，又是根据《印刷通用汉字字形表》规範了的新字形。这样就不仅和台湾正体字在字形上不尽相同，而

8. Variation Selector 又称 Variation Database 即 UTS #37。参见网页：http://www.unicode.org/reports/tr37/
9. 参见小林龙生 (2011)。不过因为汉字本身基於符号理论的研究在整体上严重滞後於信息理论的现状依然如故，“字符，而非字形”的编码原则问题直到今天仍处於没有终解的状态。
10. 由张仲陶、谢清俊、黄克东、杨键樵等科学家领导的台湾民间团体。
11. 参见网页：https://zh.wiktionary.org/wiki/
12. Ken Lunde (2002) p.98

且字符的理论定义（所涵盖的异体字）方面也会跟同一个字形的台湾正体字有出入[13]。

第三. 因《第一批简化字表》被替代的汉字，由於在 GB0 中减少了字数，就必须在 GB1 中还原其原来被替代的字，因此整套编码汉字的清单在 GB0 和 GB1 的字数是不一致的。相应码位容不下 GB1 的汉字表，於是出现了把“溢出”的汉字集中放到末尾增补区域的结果[14]。

上述第一点说明客观上造成了针对同一个汉字存在两个标準（GB0 和 GB1）的矛盾结果。第二点则可以认为 GB1 在设计上和台湾的中文资讯交换码取基本相同的思路，即二者都认为：繁体字和简化字虽然字形表达不同，实际上是同一个抽象符号，所以才在设计上将其放在完全对应的码位上。这裡，我们可以很清楚的看到，早期汉字编码的理论框架，不论是台湾还是大陆内地，都是採取繁简对应的编码思路的。

第三点，即繁体字版本的溢出字符中，有的字形上没有区别，只是字符理論值和 GB0 的同一个字不尽相同，因为“释放”了另一个被替代的字。GB1 把一部分“同形同值繁体字”（姑且这么称呼）放在 GB0 相应字的码位上，把另一个繁体字放在增补区域。而在另外一些非对应字组中则把“异形同值繁体字”放在 GB0 相应字的码位上，把“同形同值繁体字”挪到了增补区域。如表 1 所显示的部分例字那样。

GB0（基本集）的码位		GB1（辅助集）同码位	GB1 增补的 88区 - 89区	
曲 39-90	→	曲 39-90	麴 88-96	
系 47-21	→	系 47-21	係 88-82	繫 88-83
云 52-38	→	雲 52-38	云 88-93	
只 54-27	→	祇 54-27	只 89-03	隻 89-02

表 1: 从部分例字看GB1是如何对应GB0的：灰色栏安排了“溢出”的繁体汉字。

若只是为了满足“显示和打印”繁体字，那么可以把 GB1 看做犹如是一个“外字集”，对应 GB0 的重複编码（第一点）就毫无必要。所以反过来证明 GB1 的初衷是为了繁简之间的自动切换，大概是没有疑问的。但是，若真的为了达到繁简汉字“一键切换”的目的，那么就陷入 GB1 溢出字符在 GB0 没有对应码位的结果。所以结论很清楚，GB1 对 GB0 的繁简切换对应是不成功的。

笔者手头就有按照 GB0 和 GB1 编码标準的电脑字模，只需在 GB 码的状态下切换字模就能获得繁体字文件。这样看起来挺方便，其实很多大陆内地不熟悉繁体字的人，不知不觉地就会在一对多汉字面前跌跟斗。《语讯》各期中言及此类“笑话”的文章就不少，社会上对简化字有偏见的非难中也少不了这类“举例说明”。可以说其中相当一部分原因其实是繁简汉字系统理论框架缺陷的显露。

13. 这部分是源於《第一批异体字整理表》的一对多汉字。
14. 这些溢出的汉字就是因简化而非对应的一对多汉字。

鉴於 GB0 和 GB1 都是大陆内地的工业标準，而大陆内地的选用繁体字又不同於台湾正体字。所以严格地说，在大陆規範汉字和台湾正体字“一一对应”课题之前，其实早已经存在大陆内地的简化字“一一对应”大陆内地选用繁体字的问题了。从纯粹的文字符号理论看，“繁简对应”问题，实质上是一个“繁・繁・简对应”的问题。

GB0 和 GB1 都準确地反映了汉字整理规範的科研成果这一功绩是值得称道的，遗憾的是， 其失算之处多年来竟然很少有人从编码理论的根源上进行分析， 来说明 GB 编码结构上的非对应汉字直接危害到繁简同位切换的设计初衷。如果当初能有这样的分析报告，将汉字系统编码结构性问题的严重性反馈给有关部门，那么这一信息无疑会对汉字的系统性規範工作提供一个重要的思路和线索，为汉字的整理規範做一些贡献。

一方面，汉字规範部门，或汉字应用问题的研究部门也一直没有从编码框架上追究原因，反而寄希望於繁简汉字的电脑程序转换软件的进一步完美。

5. 当前汉字编码框架成为汉字同步和优化的拦路虎

目前大陆内地的标準编码 GB 18030-2000，以及升级版 GB 18030-2005的编码框架说明编码部门一开始就完全依赖利用转换程序。“统一”的繁简汉字依照字形分别安置码位，可见已经在编码结构上和早年谋求汉字符号理論意义上同一性的中文资讯交换码和 GB0／GB1 的设计意图完全不同，放弃了繁简汉字对应切换的思路。

更搗乱的是，由於“统一码”从起步时就添设了一个“源码分离原则”，规定即使是原本可以“认同”编码的两个字形接近的汉字，如果在某个地区已经分开编码的话，就有权利在新标準中仍然维持独立码位以保证新标準的向後兼容性。比如“户”字必须拥有［户／戶／戶］三个码位，就是因为作为源码之一的台湾官方字符集（CNS 码）已经把它们分别编上了码，因而在“统一码”中就不得统合，必须分别编码[15]。

听起来颇有道理的“源码分离原则”，实质上极容易让我们为了片面追求向後兼容而被诱入一个将错就错的歧途，尽管这一低级原则冲撞“字符，而非字形（Characters, not glyphs）”的高级原则。事实上，许多汉字认同的紊乱因“源码分离原则”从区域性问题扩散成世界性问题，勾销了多年来汉字整理的成果。（见表2）

同一汉字的异形字	台湾官方码 CNS 11643		统一码 Unicode		大陆官方码 GB 18030
戶	01 37 34	→	U+6236	→	91F4
户	03 01 69	→	U+6237	→	BBA7
戶	03 01 70	→	U+6238	→	91F5

表2: 以字形编码（异形异码，互不能检索）现象无原则地经由统一码扩散到大陆

曾经担任台湾教育部国语推行委员会主任委员的李鍌曾经说明过台湾教育部十分重

15. ［戶户戶］这组字来自於台湾的标準字符集 CNS 11643-1992 的第一面和第三面。

视汉字的整理工作(参见李鍔 2006)。笔者由衷希望，如果李鍔和他当年的属下或接班人果真是认真的和言行一致的话，是否应该把汉字整理工作的成果反映到数码化领域，努力消除这些字形泛滥（汉字单位重複编码）的瑕疵呢？初步估计这类源於台湾 CNS 码的无原则编码不下 160 组[16]，这一数字超过了狭义的繁简汉字非对应字符组的数量。

许多读者或许会认为数码汉字的问题本该由统一码协会来解决。然而，从结论上说，统一码编码框架的确是在当年汉字符号理论严重滞後的情况下匆匆而就，所以它的汉字编码思路不尽完美，且经过多次增补，更是杂乱无章[17]。但若想“返工”则是萬萬使不得的。因为稍有所了解的人都知道统一码另有一个不可逾越的“稳定性原则”，意思是已经被编了码的符号就不得再次分配码位。因为这毕竟也是经由各个官方代表和编码专家们讨论妥协的结果，不能因为思路的改变就在一夜之间抛弃原先的编码。从维护网际汉字通讯秩序的角度说，也不能随性而来。

由此看来，唯一能够摆脱目前种种束缚[18]的“合理途径”，就自然是两岸文字规範部门率先跨出统合汉字单位这一步，也就是本文开头所描述的走“汉字同步”道路了。这正是和谐体呼吁实现的“一一对应”，而且必须由两岸相关部门共同协商才做得到。

毕竟，统一码编码的基本要求中，“通用性”原则也是必不可少的前提条件，汉字的编码亦不能例外。实现两岸汉字单位的同步，既是全球中文汉字应用的大事，也是对汉字编码提供解决技术难题的契機，为發展世界汉字无障碍通讯做出应有的贡献。

6. 两岸学术名词统一事业是汉字同步的可借鉴範例

自上世纪九十年代以来，海峡两岸即开展各种学术名词的交流活动，力图互相沟通并逐步消除由於歷史上一段隔绝期间所造成的差异。1993年首次汪辜会谈签订协议的第四项“文教科技交流”中，就有“交换科技研究出版物以及探讨科技名词统一与产品规格标準化问题”的内容。此一事业在1996年开始具体落实，经过数百名学术工作者多年的努力，已基本完成。这为我们推动汉字单位统一提供了值得借鉴的範例[19]。

一方面，由於文字标準问题的重要性和技术上相应的要求，也就必然关係到产业界能否理解並顺利吸收汉字编码框架及认同标準的变革。解决问题的方法途径就必然需要充分论证後，协同产业界配合，技术标準一次到位，以此最大限度地减少损失[20]。

行文至此，我们可以把本文讨论所涉及到的，必须直面的汉字问题大体归纳如下：

一. 汉字本体规劃：首先是汉字自身是否应该追求系统上的一致性这一语言文字本体规劃问题，其实质即是全局型，未来型思维的汉字整理问题。

16. 信息来源：ISO/IEC 10646:2003 Annex S。
17. 参见杨宝忠 2008 ISO-10646 国际编码字符集存在的问题， 第五届两岸四地中文数字化论坛，合肥
18. 国际标準组织表意文字工作小组的汉字认同原则告诉我们：《通用规範汉字表》公布的新类推简化字不可能取代原繁体字的码位（必须另行分配编码区域），所以今后所有的类推汉字都意味着增加字符。这说明现在的统一码编码框架和认同原则是和中国继续优化汉字的要求相悖的。
19. 参见全国科学技术名词审定委员会官方网站 http://www.cnctst.cn/Home/HX
20. 目前使用的编码，仍然可以作为纪录古典文献时用。

二. 汉字符号理论深化及概念更新：数码汉字如何规劃的理论前提是认识汉字本身的性质、属性，正确表述汉字的符号同一性。以新的理论为汉字应用提供科学依据。

三. 汉字的工程学技术课题：如何运用数码化技术实现汉字符号的跨区域信息交换，需要在编码原则、编码框架及认同原则上重建工程学方法所支持的设计方案。

四. 文字公共政策：推进汉字科学的研究及本体规劃，统领汉字技术建设的主体，无疑应该经由所有汉字使用区的公共事业管理部门共同协商，统一解决。

以上这些问题长期以来互相牵掣，纠结不清，使得人们总也看不清楚解决难题应该从何处着手。因此笔者还需强调，汉字同步，即当今的书同文事业，绝不可能是某个区域、某些部门的单边行为，而是相关国家和地区共同参与，多边协作才能到达的目标。但共同使用中文的两岸文字规範部门不跨出第一步，一起推动实现汉字单位兼容多种字体字形的字符集编码框架，东亚就绝无可能实现汉字“同步”。

本文罗列了多重理由表明：香港中国语文学会所主张的和谐体原则：“可简用简，该繁用繁”加上“一一对应”无疑是实现现代汉字同步（书同文）的最佳选择。

当笔者还是个大孩子时，一天家裡的一个灯泡突然坏了，便自告奋勇上街去买来新的。没料到买来的灯泡无法用，因为家裡的电压是110伏，自己错买来220伏的。多年後自己才懂得，上海因旧时租界割据的关係，由英国人掌控的公部局和由法国人独揽的公董局曾各自为政，多家电力公司皆以本国资本利益为重，而不顾市民是否方便，致使形成两种电压规格的供电格局。

如今的上海，已不復当年，市民早已不用为电压规格的不同而苦恼了。然而在汉字同步问题上，虽然问题要複杂得多，但是两岸居民却还仍然在忍受汉字非同步的不便，成千上萬的人仍有可能稍不留意就因汉字非对应的规範而蒙羞。笔者相信，关键部门一定能理解数码汉字如何规劃的本质是一个单纯的统一度量衡性质的问题。

昨天，有关汉字书同文的理论问题和解决办法似乎还朦胧且纠缠，但是今天，柳暗花明，迈向书同文目标的途径就在我们自己的脚下。

［本文采用和谐体，这是无论转成标準简体字版或标準繁体字版都无需手工调整的唯一版本。］

〔参考文献〕

柯少斋译 1987 美国国会图书馆电脑编目，《图书馆学与资讯科学》13期，(1987-10) pp.210-223

张鼎锺 1993 中文资讯交换码与中文图书资料自动化之回顾，《鼎锺文集》，秀威資訊，pp.43-51

黄克东 1994“书同文”应向前迈进一步，《中文信息》中国科学技术协会，1994年第5期 pp.3-10

[日]加藤弘一 2000 《電脳社会の日本語》，文艺春秋

冯志伟 2000 论语言文字的地位规劃和本体规劃，《中国语文》，2000年04期

[美]Ken Lunde 2002 《CJKV Computing》，（日文版）小松章・逆井克己译，O’Reilly Japan

苏培成 2004 重新审视简化字，史定国主编《简化字研究》，商务印书馆，pp.63-81
李　鎏 2006 從學術觀點看「正體字」與「簡化字」，《语文建设通讯》99期，pp.8-13
姚德怀 2007 寻求和谐的语文生活——文字改革三大任务50周年反思之一：文字问题，《语文建设通讯》88期，pp.1-3
杨宝忠 2008 ISO-10646 国际编码字符集存在的问题，第五届两岸四地中文数字化论坛
王　宁 2009 汉字研究的新时代，《语言文字应用》2009.3，pp.17-20
[日]松冈荣志 2010《漢字・七つの物語　中国の文字改革一〇〇年》，三省堂
苏培成 2011 论《通用规範汉字表》的修订和完善，《汉语教学与研究》第一辑，pp.33-42
[日]小林龙生 2011《ユニコード戦記 ―文字符号の国際標準化バトル》，东京电机大学出版局
陈明然 2014 汉字规範和汉字信息处理技術，《语文建设通讯》，第105期 pp.29-31
朱一星 2010 漢字の国際コード規格をどう考えるべきか，漢字コード規格の根拠となる 「国際漢字単位」の提言，京都外国语大学《研究論叢》， No.76, pp.211-223
朱一星 2013 如何界定汉字的理论单位 ― 试论汉字的性质及同一性问题，京都外国语大学,《研究論叢》 2013, No.81, pp.195-207（部分转载於《语文建设通讯》第105期 pp.21-28）

［相关资料］

Unicode协会：Unicode 8.0.0　网页: http://unicode.org/versions/Unicode8.0.0/
Unicode协会：《Unicode 5.0 标准》（中文版），孙伟峰，李德龙译 2009，清华大学出版社
Unicode协会：UTR#17: Character Encoding Model，网页：http://www.unicode.org/reports/tr17/
Unicode协会：UTS #37:Unicode ideographic variation database，网页：http://unicode.org/ reports/tr37/
Unicode协会：Unicode Character Encoding Stability Policy［UNICODE 编码稳定性策略］，网页：http://unicode.org/policies/stability_ policy.html　□

（原文刊在香港中國語文學會《語文建設通訊》第110期，2015年12月）
（本文件在原文基础上稍作了词句错误的修订，2016年2月22日）