中文界面諮詢委員會

使用造字區的統籌機制

使用造字區的情況

當我們把資料存入電腦時，必須根據某種預先設定的編碼方案將有關資料予以編碼，以不同內碼代表這些不同的文字。就中文資料而言，現時有多種不同的編碼標準，不同內碼標準選取不同的中文常用字作爲它們的標準字。香港大部分中文應用系統均採用「大五碼」編碼標準，這標準是以繁體字爲基礎；內地現時採用的中文編碼標準稱爲「國標碼」，以簡體字爲基礎。「大五碼」與「國標碼」都有一個標準中文字庫，「大五碼」收納了 13,000 個字左右，「國標碼」收納了約 6,700 個字。同時，它們都預留造字區容許「用戶造字」，讓各用戶自行決定非標準字的字形及內碼編碼，例如「大五碼」容許用戶自行決定約 5,800 個造字。

2. 由於任何內碼標準只包含有限數目的字，因此未能完全滿足所有用戶的需求。部份用戶只好因應他們個別的需求，在造字區內自行訂定造字和編碼。這些造字需求大致有以下三類：

- 用戶的造字需求只在獨立的電腦中使用，用戶並不需要以這些造字與其他用戶溝通。

- 用戶需要以這些造字與其他用戶溝通，但這些造字暫時未加入標準內。

- 用戶需要以這些造字與其他用戶溝通，但由於種種原因，這些造字不能加入標準內。

問題

3. 這種做法對獨立運作的電腦而言，並不構成問題。但當這些電腦與其他電腦在網絡相連，或以電子文件傳遞訊息時，這些由各用戶自行訂定的造字就會在通訊及資料交換時產生亂碼問題，因而不能正確地傳輸資料。

4. 爲了避免產生亂碼問題，各用戶需要使用一套相同的造字編碼標準。因此，香港有需要設立一個機制，統籌使用造字區。而該機制亦可使業界能製作對這些造字兼容的產品，和避免不同用戶重覆造字。這套統籌使用造字區的機制應由資訊科技業界、中文文字專家、出版和印刷界、學術界、用戶及政府等共同制訂。

措施及部署

5. 爲使政府可以使用一套相同的造字編碼標準作內部電子通訊，政府已於一九九五年建立了一套名爲「政府通用字庫」（Government Common Character Set - GCCS）的字集，收納在本港通用但並未納入「大五碼」內的中文字。現時「政府通用字庫」有大約 3,000 個中文字，部分屬於香港特有的中文字，例如人名、地址及廣東話口語化用字。這套字集亦被放置於政府網站，讓公眾人士隨意下載。目前，每月從政府網站下載該套字集的次數約 5,000 次。

6. 資訊科技署與法定語文事務署現正合作更新該套字集，以收納更多本港新的中文造字，從而推出「政府通用字庫第二版」（GCCS II）。由於今次的更新已包含了政府內部和從社會各界收集的中文字，所以「政府通用字庫第二版」將符合香港的一般中文字需要。爲了更貼切地形容這套更新後的字集，我們建議把它名爲「香港通用字庫」。不過，由於用戶可隨意使用造字區來訂定他們的造字，「政府通用字庫第二版」所選取的內碼有可能與其他軟件發展商或用戶已經使用的造字編碼有所衝突。

7. 長遠來說，我們的策略是採用「國際標準 10646」作爲香港的中文界面。在國際標準組織的統籌下，我們正與其他政府和機構合作發展一套名爲「國際標準 10646」的國際編碼標準，以單一套通用字集收納各種語文的文字。目前該套標準已收納約 20,000 多個漢字，新的版本估計將會收納超過 60,000 個漢字。我們正積極爭取把香港常用的中文字納入該套國際標準之內。我們相信，這套新的「國際標準 10646」公布後，將會成爲包羅所有我們現時已知編碼方案字庫的超級字庫，並作爲各編碼方案間轉換及交換資料的基礎。目前電腦在處理中文字上所受到的限制，屆時將可獲得紓緩。

8. 不過，由於中文字的數目會因應社會和科技發展等因素而增加，縱使日後的「國際標準 10646」已經包括了香港常用的中文字，造字的需要還是會繼續存在的，我們仍需要一套統籌使用造字區的機制；而在「國際標準 10646」上發展的機制，還要顧及和「大五碼」造字的配合或轉換。

<u>建議</u>

9. 我們建議，以「政府通用字庫第二版」作爲「香港通用字庫」，並盡量減少與現有其他造字衝突的可能，以便利市民使用電腦與政府或其他市民通訊。由於實際的急切需要，我們希望經本委員會討論及確認後，可於本年七月正式公佈這套「香港通用字庫」。現提交「政府通用字庫第二版」初稿（請參閱附件），以便本委員會審議。

10. 我們亦建議本委員會於日後詳細討論有關香港如何統籌使用「國際標準 10646」造字區，以及研究如何將「大五碼」的造字過渡到「國際標準 10646」造字區內等問題。

資訊科技署
一九九九年五月