《世載堂雜憶》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中研院史語所漢籍電子文獻工作小組
2008/5/15 陳建安 製作

一、 《世載堂雜憶》一書(世載堂雜憶.xml)使用舊版造字 113 個，字頻 567 次，詳如附件一。這 113 個造字中，92 個可轉成 Windows XP 能顯示的字，字頻 451 次；另外 21 個字必須轉成構字式，字頻 116 次。

二、 附件一的造字分析表說明如下：

甲、 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。
乙、 造字：舊版造字
丙、 字頻(txt)：造字在「.txt」文件的出現次數
丁、 字頻(xml)：造字在「.xml」文件的出現次數
戊、 Big5：造字的 Big5 碼
己、 Unicode：造字所對應的 Unicode 碼
庚、 WinXP：造字在 Windows XP 的對應字形
辛、 構字式：Windows XP 若無對應字形，則改採用構字式
壬、 備註凡例：

1、校對問題，舊版漢籍錯字，可用程式全部取代：在舊版漢籍電子文獻中即存在的錯字，因校對時的疏漏而未更正，持續留存在新版漢籍電子文獻中；若該造字的所有頻次，皆屬於錯誤使用的錯字情形，可以用程式全部取代為正確字形。如編號 923 的「火产」字，原字為「爇」。

2、標記問題：標記問題主要是漢籍電子文獻在舊轉新的過程中，採用了新的標記語言，而在修改標記時，對於某些標題句做了增刪的動作，因為這些標題句的內容包含了缺字,所以這些更改標記的動作造成 txt 檔與 xml 檔缺字頻次不符的情形。如編號 2130 的「」字，xml 檔在 55 頁新增了一個註解，因而造成頻次不同，處理方式為將所有頻次皆修改為「」。

3、異體字問題：新版漢籍考量到使用者檢索及使用時的便利性，將用字原則改為除專詞等特殊情形之外，一律改用標準字呈現。如編號 4132 的「」係「眾」字之異體，故以「眾」字取代。又編號 4067 的「綉」字，係「繡」字之簡體字，亦以標準字的「繡」字取代。

4、Unicode 字型呈現差異：Unicode 字型與舊漢籍造字有些微差異，但只是字體風格差異，實際上仍為同一個字，因此仍取 Unicode 字型。如編號 4158 的「譕」字，Unicode 字型呈現為「譞」，實際上仍為同一字。

5、待造字：Unicode 及漢字構形資料庫皆未收錄的舊漢籍造字，正在等待補造字中，所以「造字」欄空白無法看到字形。如編號 5749 的「片庸」。

三、 Unicode 目前收錄的漢字總數為 70194，分屬於三個不同區段，詳如表一。目前 Windows XP 只支援 CJK 認同表意文字區的 20902 個字，內碼為 4E00-9FFF。所以造字編號 1704 的「  」字，Unicode 編碼為 3BAE，由於 Windows XP 並不支援，仍須使用構字式「木  」。

表一、Unicode 的字數及編碼區段

Unicode	字集子集合	新增字數	新增編碼區段	總字數	WinXP
1.1 版	CJK 認同表意文字區	20902	4E00-9FFF	20902	支援
3.0 版	CJK 認同表意文字擴充 A 區	6582	3400-4DFF	27484	不支援
3.1 版	CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	70194	不支援

四、 txt 檔與 xml 檔頻次不同且造字分析表「備註」欄有備註問題情況者，詳細取代內容可參考附件二的手動取代表。附件二的手動取代表的欄位說明如下：

甲、 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

乙、 造字：舊版造字字形。

丙、 檔案原文摘錄：摘錄檔案(世載堂雜憶.xml)中包含該造字的文句段落，其中紅字底線的部分，為需要手動取代的原文。文句後方標示（全部取代）者，表示執行全部取代的動作，不一一列舉所有原文。

丁、 手動取代結果：變更過後的檔案內容，其中藍字底線的部分，為修改後的結果。文句後方標示（全部取代）者，表示執行全部取代的動作，不一一列舉所有原文。

附件一、《世載堂雜憶》造字分析表

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
296		11	11	FBEC				
297		6	6	FBED				
298		2	2	FBEE				
299		3	3	FBEF				
300		3	3	FBF0				
308	∞	1	1	FBF8			∞	
449	斤	1	1	FCE8			斤	
628	縈	5	5	FDFE			縈	
665	夂	1	1	FE64			夂	
717	足	1	1	FEBA			足	
752	竹	2	2	FEDD			竹	
758	罒	1	1	FEE3			罒	
923		1	1	8EEB			火产兼	爛，校對問題，舊版漢籍錯字，可用程式全部取代。
1167	綱	1	1	90A5	7D97	綱		
1354	褆	1	1	91C3	79A9	褆		
1565	縣	1	1	92F9	7DDC	縣		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
1704	樨	1	1	93E7	3BAE		木挈	
1871	蟲	2	2	94F1	87C1	蟲		
1900	袿	1	1	954F	88BF	袿		
2046	踪	1	1	9644	8E2A	踪		
2050	蹏	1	1	9648	8E4F	蹏		
2130	窻	1	2	96BA	6119	窻		窻，標記問題，人工取代。
2142	睘	1	1	96C6	7758	睘		
2151	卩	1	1	96CF	961D	卩		
2274	鎷	1	1	97AD	9348	鎷		
2374	冢	2	2	9852	51A1	冢		塚，異體字問題。
2570	鴂	1	1	9979	21C08		梟  兀	
2806	勒	1	1	9AEA	976D	勒		
3789	冲	2	2	8154	51B2	冲		沖，異體字問題。
3799	刦	1	1	815E	5226	刦		劫，異體字問題。
3804	効	6	6	8163	52B9	効		
3805	勅	1	1	8164	52C5	勅		
3814	却	9	9	816D	5374	却		卻，異體字問題。
3816	厓	1	1	816F	5393	厓		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
3829	咏	1	1	817C	548F	咏		
3835	噍	1	1	81A4	564D	噍		
3840	坂	2	2	81A9	5742	坂		
3842	塋	1	1	81AB	5803	塋		
3846	堺	8	8	81AF	58FB	堺		Unicode 呈現差異。
3855	孽	4	4	81B8	5B7C	孽		孼，異體字問題。
3858	寃	5	5	81BB	5BC3	寃		冤，異體字問題。
3860	尠	2	2	81BD	5C20	尠		
3864	峯	16	16	81C1	5CEF	峯		峰，異體字問題。
3883	徧	2	2	81D4	5FA7	徧		
3886	悞	2	2	81D7	609E	悞		
3891	慙	1	1	81DC	6159	慙		慚，異體字問題。
3909	擡	2	2	81EE	64E1	擡		
3921	晢	7	7	81FA	6673	晢		晳，異體字問題。
3929	棃	4	4	8243	68C3	棃		
3930	椁	7	7	8244	6901	椁		
3966	烟	30	30	8268	70DF	烟		煙，異體字問題。
3969	煊	19	19	826B	714A	煊		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
3974	牀	8	8	8270	7240	牀		
3990	甎	2	2	82A2	750E	甎		
3996	疎	1	1	82A8	758E	疎		
4000	疴	2	2	82AC	75B4	疴		
4013	盈	1	1	82B9	76CC	盈		
4016	着	15	15	82BC	7740	着		著，異體字問題。
4034	穉	2	2	82CE	7A49	穉		
4036	窯	2	2	82D0	7AB0	窯		
4043	笄	1	1	82D7	7B53	笄		
4067	綉	2	2	82EF	7D89	綉		繡，異體字問題。
4077	罇	4	4	82F9	7F47	罇		
4080	羣	37	37	82FC	7FA3	羣		群，異體字問題。
4081	翺	34	34	82FD	7FFA	翺		
4090	脚	4	4	8347	811A	脚		腳，異體字問題。
4094	舘	1	1	834B	8218	舘		
4098	芺	1	1	834F	82AA	芺		
4105	菓	1	1	8356	83D3	菓		
4111	蕋	25	25	835C	8493	蕋		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
4129	虬	4	4	836E	866C	虬		
4132		63	63	8371			血𠔃巫	眾，異體字問題。
4134	衛	38	38	8373	885E	衛		衛，異體字問題。
4139	袴	1	1	8378	88B4	袴		
4142	褒	2	2	837B	8943	褒		
4146	覩	8	8	83A1	89A9	覩		
4158	譌	2	2	83AD	8B4C	譌		Unicode 呈現差異。
4173	躭	1	1	83BC	8EAD	躭		
4184	迹	11	11	83C7	8FF9	迹		
4193	鈎	3	3	83D0	920E	鈎		
4213	韡	1	1	83E4	97BE	韡		
4220	穎	1	1	83EB	9834	穎		穎，異體字問題。
4234	贏	1	1	83F9	9A58	贏		
4237	髡	1	1	83FC	9AE0	髡		髡，異體字問題。
4242	鬬	3	3	8442	9B2D	鬬		Unicode 呈現差異。
4247	鮎	2	2	8447	9B8E	鮎		
4256	鷄	5	5	8450	9DC4	鷄		
4258	麐	7	7	8452	9E90	麐		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
4259	麕	1	1	8453	9E95	麕		
4267	鼇	1	1	845B	9F08	鼇		
4278	亘	4	4	8466	4E98	亘		
4308	刧	1	1	84A6	5227	刧		劫，異體字問題。
4409	冏	1	1	854C	518B	冏		
4431	遂	2	2	8562			辶㒳象	
4515	呪	4	4	85D8	546A	呪		
4702	塲	1	1	86F6	5872	塲		場，異體字問題。
4768	妬	1	1	8779	59AC	妬		妒，異體字問題。
4940	韵	2	2	88AA	97F5	韵		
4996	斷	2	2	88E2	65B5	斷		
5176		5	5	89F9			氵㠯冗	沉，異體字問題。
5220	冊	2	2	8A66	6BCC	冊		
5290	携	1	1	8ACE	643A	携		攜，異體字問題。
5298	旰	2	2	8AD6	65FF	旰		
5410	枏	1	1	8BA9	67AC	枏		
5437	幷	1	1	8BC4	5E77	幷		并，異體字問題。
5472	厾	3	3	8BE7			厾	瓦，異體字問題。

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
5514	盞	2	2	8C52	25095		執  皿	盞，異體字問題。
5551	賮	3	3	8C77	8CF7	賮		
5556	隸	7	7	8C7C	96B7	隸		隸，異體字問題。
5559	欵	2	2	8CA1	6B35	欵		款，異體字問題。
5666	汚	7	7	8D4D	6C5A	汚		污，異體字問題。
5706	敍	23	23	8D75	654D	敍		敘，異體字問題。
5749		1	1	8DC2			片  庸	待造字。

附件二、《世載堂雜憶》手動取代表

編號	造字	檔案原文摘錄	手動取代結果
2130	鬯	鬧（全部取代）	鬯（全部取代）