

史語所「考古資料數位典藏資料庫」中倚天字問題

資訊所 文獻處理實驗室

2007/10/15 鄧賢瑛

一、問題描述：

2007年9月12日，由史語所中原考古庫房、資訊所數位典藏技術發展組以及文獻處理實驗室一同開會討論「考古資料數位典藏資料庫」中的缺字及顯示亂碼問題，史語所中原考古庫房同仁在會中反應有若干特定的字，雖然在個人電腦使用的輸入法中可以打得出這些字形，但若將包含這些字形的資料以批次轉入或線上著錄的方式，輸入「考古資料數位典藏資料庫著錄系統」之後，這些字形不但會以「？」呈現，同時也會影響其上下文，使該字形之後出現一連串的亂碼。如以「銹」字為例：

【舉例一】

登錄號：T0030384

欄位名稱：遺物形制描述

原本正確資料為：器表銹蝕嚴重

資料呈現亂碼為：器表？襍k嚴重

考古資料數位典藏系統
Archaeological Data

遺物描述

形制	描述	與T0030379同型式但小些，略弧狀平底，無圈足，器表？襍k嚴重，素面無紋。
	其他形制描述	
	分類	碗

【舉例二】

登錄號：T0030389

欄位名稱：遺物形制描述

原本正確資料為：銅銹分佈不一

資料呈現亂碼為：銅？蚌壩洵虫@



二、原因及解決方式

由史語所中原考庫房同仁在 2007 年 9 月 12 日的「考古資料數位典藏資料庫」缺字討論會中提出此一問題，會中決議由資訊所數位典藏技術發展組同仁將現行「考古資料數位典藏資料庫」中的錯誤資料轉成 excel 資料表，並交由中原考古庫房同仁以人工方式一一作字形確認。在經過歸納分析之後發現，除了前例「鏽」字以外，還會產生問號及亂碼的字形還有「裏」字（參見附件一）以及「嫺」字（參見附件二）。而這三個字正是當初倚天中文系統針對 Big5 收錄字形不足，而另行加入的四十一個「倚天字」中的字形，這四十一個「倚天字」及其在 Big5 中的字碼可如下表所示：

字形	Big5 字碼
碁	F9D6
鏽	F9D7
裏	F9D8
墻	F9D9
恒	F9DA
粧	F9DB
嫺	F9DC
𠂔	F9DD
𠂕	F9DE
𠂖	F9DF
𠂗	F9E0
𠂘	F9E1
𠂙	F9E2

𠂚	F9E3
𠂛	F9E4
𠂜	F9E5
𠂝	F9E6
𠂞	F9E7
𠂟	F9E8
𠂠	F9E9
𠂡	F9EA
𠂢	F9EB
𠂣	F9EC
𠂤	F9ED
𠂥	F9EE
𠂦	F9EF
𠂧	F9F0

𠂨	F9F1
𠂩	F9F2
𠂪	F9F3
𠂫	F9F4
𠂬	F9F5
𠂭	F9F6
𠂮	F9F7
𠂯	F9F8
𠂰	F9F9
𠂱	F9FA
𠂲	F9FB
𠂳	F9FC
𠂴	F9FD
■	F9FE

由於會產生亂碼問題的字形都與「倚天字」有關，可推知在「考古資料數位典藏資料庫」中產生亂碼的字形，應與當初資訊所數位典藏技術發展組開發該資料庫的第一個版本時，未處理「倚天字」的字碼有關。

由於資訊所數位典藏技術發展組於 2003 年 4 月所開發「考古資料數位典藏資料庫」著錄系統的最初版本時，因不曉得有「倚天字」的問題，也並未對「倚天字」的字碼作特殊處理。所以當中原考古庫房同仁將包含「倚天字」的資料輸入「考古資料數位典藏資料庫」著錄系統時，因資料庫著錄系統無法辨識「倚天字」的字碼，便會將原本由二組 ASCII 組成的字碼誤轉為由三組 ASCII 組成的字碼，並連帶影響下面字碼的取得，而造成一連串的錯誤字碼，如以「鏽」為例，原本正確的資料如下：

字形	背	面	鏽	塊	大	而	遍	佈	器	表
字碼	AD49	ADB1	F9D7	B6F4	A46A	A6D3	B94D	A747	BEB9	AAED

因「鏽」字屬於「倚天字」之一，故資料庫著錄系統會將「鏽」字的字碼由 F9D7 轉成 A148D7，並且連帶影響了之後的字碼：

字形	背	面	?	祛	櫓	j	而	遍	佈	器	表
字碼	AD49	ADB1	A148	D7B6	F4A4	6A	A6D3	B94D	A747	BEB9	AAED

資訊所數位典藏技術發展組於 2003 年 12 月更新「考古資料數位典藏資料庫」系統版本時，已發現並處理了「倚天字」字碼的問題，所以自 2003 年 12 月之後著錄的資料並不會有此一情況發生，然而原先已存在於資料庫第一版本中的錯誤字碼則一直未作修訂。

直至 2007 年 9 月 12 日，史語所中原考古庫房、資訊所數位典藏技術發展組以及文獻處理實驗室一同討論「考古資料數位典藏資料庫」中的缺字問題後，決議將錯誤的資料由資料庫中轉出，再經過歸納與分析，發現問題的原因是來自於倚天字碼，目前正在逐一修訂當中。