


From: **Steve Majewski** steve.majewski@gmail.com 
Subject: BIG5 encoding question on Chinese texts
Date: November 30, 2018 at 3:43 PM
To: Anne Kinney aeb2n@virginia.edu, Roper, Jennifer O'Brien (jor2a) jroper@virginia.edu
Cc: Mike Durbin md5wz@virginia.edu



I've almost completed fixes to Chinese Texts, but I am having trouble with one file which has bad BIG5 encoding.
I am unable to edit and save the file, as all of the editors I've tried truncate the file at the bad character.

The most stubborn encoding problem is in the text below (original HTML) and the browser display it as a question mark in a dark diamond:

184. 鶴鳴 HE MING

鶴鳴于九皋、聲聞于野。
魚潛于淵、或在于渚。
樂彼之園、爰有樹檀、其下維蕓。
它山之石、可以為錯。

鶴鳴于九皋、聲聞于天。
魚在于渚、或潛于淵。
樂彼之園、爰有樹檀、其下維❖-c。
它山之石、可以攻玉。

The crane cries in the ninth pool of the marsh ,
And her voice is heard in the [distant] wilds .
The fish lies in the deep ,
And now is by the islet .
Pleasant is that garden ,
In which are the sandal trees ;
But beneath them are only withered leaves .
The stones of those hills ,
May be made into grind-stones .

The crane cries in the ninth pool of the marsh ,
And her voice is heard in the sky .
The fish is by the islet ,
And now it lies hid in the deep .
Pleasant is that garden ,
In which are the sandal trees ;
But beneath them is the paper-mulberry tree ,
The stones of those hills ,
May be used to polish gems .

The XTF version of the text http://xtf.lib.virginia.edu/xtf/view?docId=Chinese/uvaGenText/tei/shi_jing/AnoShih.xml;chunk.id=AnoShih.2;toc.depth=1;toc.id=AnoShih.2;brand=default;query=The%20crane%20cries%20%20in%20the%20ninth%20pool%20of%20the%20marsh#1

appears below:

184. 鶴鳴 HE MING

鶴鳴于九皋、聲聞于野。
魚潛在淵、或在于渚。
樂彼之園、爰有樹檀、其下維蕕。
它山之石、可以為錯。

鶴鳴于九皋、聲聞于天。
魚在于渚、或潛在淵。
樂彼之園、爰有樹檀、其下維蕕。
它山之石、可以攻玉。

The crane cries in the ninth pool of the marsh ☞ ,
And her voice is heard in the [distant] wilds .
The fish lies in the deep ,
And now is by the islet .
Pleasant is that garden ,
In which are the sandal trees ;
But beneath them are only withered leaves .
The stones of those hills ,
May be made into grind-stones .

☞ **The crane cries in the ninth pool of the marsh** ,
And her voice is heard in the sky .
The fish is by the islet ,
And now it lies hid in the deep .
Pleasant is that garden ,
In which are the sandal trees ;
But beneath them is the paper-mulberry tree ,
The stones of those hills ,
May be used to polish gems .

I would think that the XTF xml version is the newer, corrected version, but I wanted to verify if this was the case, or if you have other corrections. (I don't read Chinese, but it looked curious to be that those lines looked identical in Chinese characters, while the translated english text was different.)

There are some other characters in that text that are bad encoding using BIG5 as the encoding, but which seem to be OK if I use one of the other BIG5 variants.

list of BIG5 variants listed below (names on the same line are aliases):

```
UCS-2BE UNICODE-1-1 UNICODEBIG CSUNICODE11
BIG-5 BIG-FIVE BIG5 BIGFIVE CN-BIG5 CSBIG5
BIG5-HKSCS:1999
BIG5-HKSCS:2001
BIG5-HKSCS BIG5-HKSCS:2004 BIG5HKSCS
BIG5-2003
```

Do you know specifically which flavor of BIG5 encoding is used ?

The files which had the charset encoding labeled in HTML used either "x-big5" or "x-x-big5" , and I've changed these to CHARSET="big5" so that they are properly displayed in the browser.

If I use BIG5-2003 or BIG5-HKSCS, for this file, that appears to fix the other encoding errors, so I will assume one of those is the actual encoding unless you have any more specific instructions.

Below are diffs where "|f9|" is replacement character for what is an invalid character in BIG5,

versus the lines where the character is interpreted as valid BIG5-2003 encoding:

```
shijing$ diff AnoShih.subst1 AnoShih.subst3h
1233c1233
< 綠兮衣兮、綠兮黃|f9|堦C<br>
---
> 綠兮衣兮、綠兮黃裏。<br>
3661c3661
< 將仲子兮、無踰我|f9|畹B無折我樹桑。<br>
---
> 將仲子兮、無踰我牆、無折我樹桑。<br>
7784c7784
< 如月之|f9|踞B如日之升。<br>
---
> 如月之恒、如日之升。<br>
```

- Steve Majewski.

PS: I have added a search facet for Chinese Text Initiative files that are in XTF:

<http://xtf.lib.virginia.edu/xtf/search?expand=subject;f1-subject=Chinese%20Text%20Initiative>

I still need to figure out how to get these texts reindexed in Virgo as well as XTF for the search facets to show up there.

