

Chinese XML



- **Attracting a fair bit of attention**

- the best reference:

- <http://xml.ascc.net/>

- chinese data and markup can be used

- *encoding problems may arise*

- in particular Big5 is not good, it seems

- *uses component characters that aren't allowed in XML names, so can't use as markup*
 - *2nd byte of many characters is 'J'; causes problems in CDATA sections*

- also UniCode has its own problems

- *often uses 3 bytes for chinese -> file 'bloat'*
 - *can be problems translating from other character sets to UniCode*

- *depends on:*

- need to explicitly state the encoding of the XML file

- ```
<?xml version="1.0" encoding="Big5"?>
```

- *if a document ("parseable entity") is not labelled with the correct character set information, an XML processor will reject it; XML moves away from character set guessing (i.e., what HTML does) to explicit markup of character sets*

- application quality

- *the best markup will be useless if the application isn't written to understand it...*

# Examples

```

C:\WINNT\Profiles\Mob\Desktop\wedoshazi.xml - Microsoft Internet Explorer - [Working Offline]
Address: http://www.winnt.com/Profiles/Refr/Desktop/wedoshazi.xml
<?xml version="1.0" encoding="UTF-8" >
<rs>Poems</rs>
<rs>Chinese</rs>
<rs>By the government music agency (Wulin).</rs>
<rs>Apparently for use (by non-professionals) as lyrics for songs for various occasions, including
government religious events.</rs>
<rs>Collected early this millennium.</rs>
<rs>Authors include famous emperor.</rs>
</create>
</put into>
</te-header>
<text lang="zh">
- <group rend="Intro">
<h1 name="name">卷第二十六 相和歌辭</h1>
<h2 name="A">
<head rend="name">前調曲四</head>
- <text>
- <body>
- <div rend="p">
<h3="526" />
<h3>樂府詩集卷第二十六</h3>
<h3>相和歌辭十一</h3>
<h3>前調曲四</h3>
<h3>秋胡行四解</h3>
<h3>魏武帝</h3>
<h3>indented-6.6</h3>
《西京雜記》曰：「魯人秋胡，娶妻一月，而適宦一年，休還家。其婦採桑於郊。胡至
郊而不識其妻也，見而悅之，乃遺黃金一綰。妻曰：『妾有人，難直不返。幽閉獨處，
二年丁亥，未有被尋於今日也。』徐衆不顧，胡慚而退。至家，問：『妾何在？』曰：
『行
採桑於郊，未返。』既歸還，乃向所挑之婦也，夫棄並慚，妻赴沂水而死。』《列女傳》
曰：「魯秋潔婦者，魯秋胡之妻也。既納之五日去，而宦於陳，五年乃歸。未至其
舅路傍有美婦人，方採桑而說之。下車謂曰：『力田不如逢豐年，力桑不如見國卿。』
金綰有金，願因而去。』妻曰：『終為君死，姑待君如君所求，金一綰也。』

```

```

http://www.ascc.net/xml/test/wf/utf-0/text-xml/zh-utf-0-1.xml - Microsoft Internet Expl...
Address: http://www.ascc.net/xml/test/wf/utf-0/text-xml/zh-utf-0-1.xml
<?xml version="1.0" encoding="UTF-8" >
<!-- Copyright 1996 Academia Sinica Computing Center -->
<!-- Permission to use and distribute granted under GPL -->
<!-- Email questions to rickollgase@sinica.edu.tw -->
- <test type="to11">
<name>Chinese Test #11: UTF-8</name>
<data>This file has 1 Chinese character, directly entered. This tests
Native Language Markup (NLM).</data>
<h1>The XML header of this file is <?xml encoding="UTF-8"?
></data>
<h1>
The character is here:
<h1>
</data>
</test>

```

- **Applicable to every XML element**
  - sets the language used in the XML document
  - possible values for Chinese include:
    - *xml:lang="zh" for any Chinese text*
    - *xml:lang="zh-TW" for Chinese text from Taiwan (i.e., traditional characters)*
    - *xml:lang="zh-HK" for Chinese text from Hong Kong (i.e. probably traditional characters)*
    - *xml:lang="zh-CN" for Chinese text from China (i.e., simplified characters)*
    - *xml:lang="zh-SG" for Chinese text from Singapore*
    - *xml:lang="zh-CN-YUH" for Cantonese*

```
<p xml:lang="en-GB">What colour is it?</p>
<p xml:lang="en-US">What color is it?</p>
```

- A growing list
  - see
    - <http://www.ascc.net/xml/en/utf-8/software.html>
  - most XML software is compatible to some extent
    - note the 'numberplates'

Software for Chinese XML - Microsoft Internet Explorer - [Working Offline]

Address: <http://www.ascc.net/xml/en/utf-8/software.html>

### XML-Aware Text Processing Applications







| Name     | From                         | Numberplate | Comment                                                                |
|----------|------------------------------|-------------|------------------------------------------------------------------------|
| Expat    | <a href="#">XML.com</a>      |             | Public, C++.                                                           |
| CoreMark | <a href="#">coremark.com</a> |             | Limited, C++.                                                          |
| Boice    |                              |             | Many many platforms.                                                   |
| LibXML   | <a href="#">Gnome</a>        |             | Public NonCom. many utilities such as an xml version of grep and sort. |

Software for Chinese XML - Microsoft Internet Explorer - [Working Offline]

Address: <http://www.ascc.net/xml/en/utf-8/software.html>

### XML Parsers

| Name                | From                    | Numberplate | Comment                                                                                   |
|---------------------|-------------------------|-------------|-------------------------------------------------------------------------------------------|
| SP                  | <a href="#">XML.com</a> |             | Public, C++.                                                                              |
| Expat               | <a href="#">XML.com</a> |             | Public, C++.                                                                              |
| XT                  | <a href="#">XML.com</a> |             | Public, Java.                                                                             |
| QExpat              | <a href="#">Gnome</a>   |             | Public, Java.                                                                             |
| XML parser for Java | <a href="#">IBM</a>     |             | Public, Java. Supports many different encodings, especially EBCDIC family.                |
| Project X           | <a href="#">Sun</a>     |             | Public, Java library. Source code not available? Said to support 120 different encodings. |
| DXP                 | <a href="#">Gnome</a>   |             | Java.                                                                                     |
| LT                  | <a href="#">Gnome</a>   |             | Public NonCom, Java.                                                                      |

| XML-Aware Applications |                           |                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------|---------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Name                   | From                      | Numberplate                                                                          | Comment                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| ProcessMaker-XML S.S.S | <a href="#">Adobe</a>     |    | The character set the XML character set in which file - we have used in the File-Utilities->Application menu defaults. Supports Chinese input and typesetting. Remember to change the default character set for applications which have glyphs for Chinese, otherwise you will get strange sequences. I don't think selecting character set is as flexible as it is for the Microsoft products, maybe Adobe are making it in something that needs to be set up as part of installation and not a user option.                                                                                         |
| Adapt Editor           | <a href="#">AdaptTech</a> |    | Supports Chinese input and typesetting.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| XML media server       | <a href="#">IBM</a>       |    | Provides XML-based web server on top of IBM's media server.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| DynaEdit               | <a href="#">IBM</a>       |    | DynaEdit (viewer) supports many different character set. The new DynaEdit (editor) which is using technology from General Dynamics TechSight (see next).                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| WebSight               | General Dynamics (GDS)    |   | Seems to be superseded by DynaEdit/IE (see above). This seems to support Chinese character. OK in IE5, but I think the version just accepts strings, and does not understand the character set. (There may be an issue with headings: TechSight seems to uppercase all headings, which corrupts the display of English names which have [a-z] as their second byte; this may just be a stylesheet problem.)                                                                                                                                                                                           |
| Internet Explorer 5    | Microsoft                 |  | The English language version of Internet Explorer 5 (beta 2.00.0910.1200) running on an English Windows 95 was able to display the XML test files correctly. The Chinese language version of the IE5 (same number) running on traditional Chinese Windows 95 could not display the same XML files. We are looking into this. Chinese users might be useful after a bad experience here, do not download English IE beta onto Chinese Windows - it can cause trouble. (Also, the IE5 beta did seem to choose on the 'standalone' attribute in the XML encoding 11 and 'charset' in the stylesheet 22.) |