Chinese XML

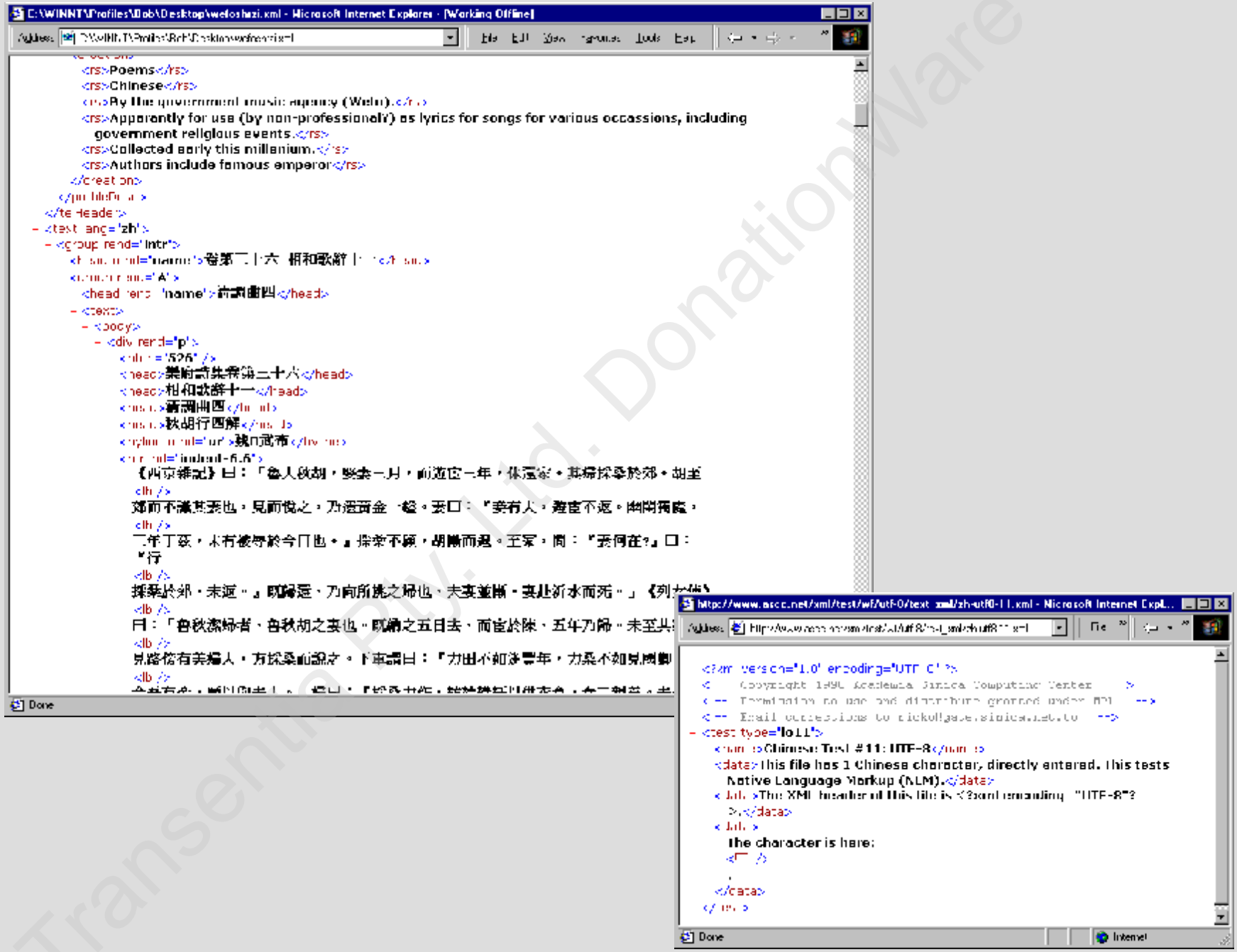
- **Attracting a fair bit of attention**
  - the best reference:
    - *http://xml.ascc.net/*
  - chinese data and markup can be used
    - *encoding problems may arise*
      - in particular Big5 is not good, it seems
        - *uses component characters that aren't allowed in XML names, so can't use as markup*
        - *2nd byte of many characters is ']'; 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…*

July 3, 2009

Examples



xml:lang Attribute

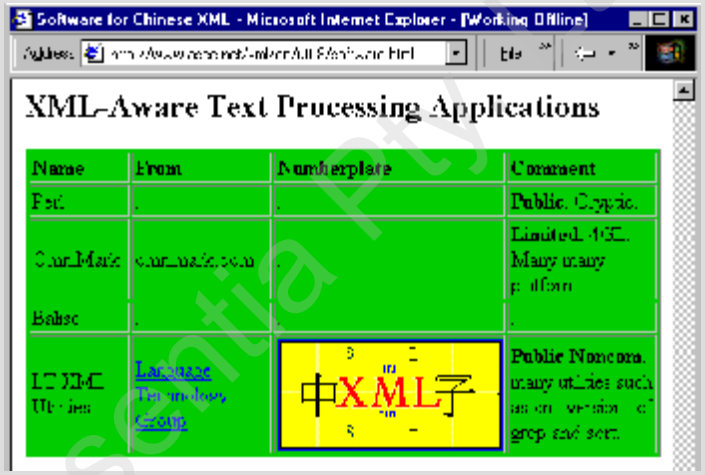
- **Applicable to every XML element**
  - sets the language used in the XML document
  - possible values for Chinese include:
    - *xml:lang="zh" for any Chinese text*
    - *xml:lang="zh-TW" for Chinese text from Taiwan (i.e., traditional characters)*
    - *xml:lang="zh-HK" for Chinese text from Hong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

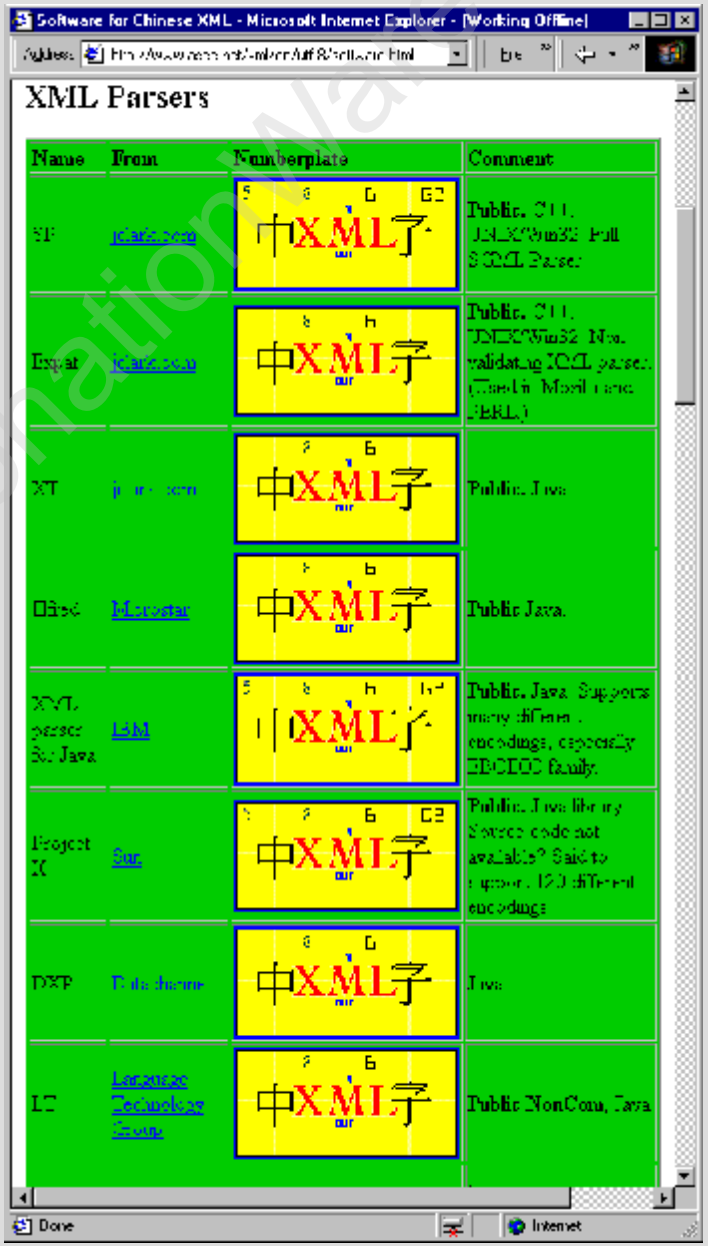
```
<p xml:lang="en-GB">What colour is it?</p>
<p xml:lang="en-US">What color is it?</p>
```

Software

- **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'*





More Software