Every XML document should have an encoding declaration as part of its XML declaration. The encoding declaration tells the parser what character set the document is written in. When a parser reads a document, it translates characters from the document's native encoding as specified by the encoding declaration, into Unicode. Not all parsers know how to convert all encodings, but most major parsers can handle most character sets.

The encoding declaration can be omitted if and only if the document is written in the UTF-8 or UTF-16 encodings of Unicode. UTF-8 is a superset of ASCII, so pure 7-bit ASCII files can be legal XML documents without an encoding declaration.

Most XML processors understand other legacy encodings. However, XML processors are only required to support UTF-8 and UTF-16, and not the hundreds of different legacy encodings used around the world.

ISO Character Sets

ASCII(ISO-646 or ISO-ASCII)

7-bit ASCII uses characters 0-127.

ISO-8859-1(Latin-1) Note: all ISO-8859 sets are 8-bit

ASCII plus the accented characters and other characters needed for most Latin-alphabet Western European languages, including Danish, Dutch, Finnish, French, German, Icelandic, Italian, Norwegian, Portuguese, Spanish, and Swedish.

ISO-8859-2(Latin-2)

ASCII plus the other characters needed to write most Latin alphabet Central and Eastern European languages, including Czech, English, German, Hungarian, Polish, Romanian, Croatian, Slovak, Slovenian, and Sorbian.

ISO-8859-3(Latin-3)

ASCII plus the accented letters and other characters needed to write Esperanto, Maltese, and Turkish.

ISO-8859-4(Latin-4)

ASCII plus the accented letters and other charactees needed to write most Baltic languages, including Estonain, Latvian, Lithuranian, Greenlandic, and Lappish. Now deprecated – use 8859-10 or 8859-13 instead.

ISO-8859-5

ASCII plus the Cyrillic alphabet used for Russian and other languages of the former Soviet Union and other Slavic countries, including Bulgarian, Byelorussian, Macedonian, Serbian, and Ukrainian.

ISO-8859-6
ASCII plus Arabic, not including Farsi and Urdu.

ISO-8859-7
ASCII plus modern Greek. This set does not have the extra letters and accents necessary for ancient and Byzantine Greek.

ISO-8859-8
ASCII plus Hebrew script used for Hebrew and Yiddish.

ISO-8859-9(Latin-5)
Like Latin-1, except six letters used in Icelandic have been replaced with six letters used in Turkish.

ISO-8859-10(Latin-6)
ASCII plus accented letters and other characters needed to write most Baltic languages, including Estonian, Icelandic, Latvian, Lithuanian, Greenlandic, and Lappish.

ISO-8859-11
ASCII plus Thai.

ISO-8859-13(Latin-7)
Similar to Latin-6, except with some question marks.

ISO-8859-14(Latin-8)
ASCII plus Celtic languages, including Gaelic and Welsh.

ISO-8859-15(Latin-9, Latin-0)
A revived version of Latin-1 that replaces symbols such as ¤, with extra French and Finnish letters. Includes the Euro currency symbol €.

Unicode
Unicode is an international standard character set that can be used to write documents in most languages. The latest version, 3.0.1, contains 49,194 characters. Unicode covers the Latin alphabet, Greek-derived scripts(including ancient and modern), Cyrillic scripts. Unicode also covers ideographic scripts, including the Han character set used for Chinese and Japanese, the Korean Hangul syllabury, and phonetic representations of these languages, including Katakana and Hiragana. It covers right-to-left Arabic and Hebrew scripts. It covers scripts of the Indian subcontinent, including Devanagari, Thai, Bengali, and Tibetan. Unicode can potentially hold more than a million characters.

UCS-2 and UTF-16
UCS-2(**Universal Character System-2**), also known as **ISO-10646-UCS-2**, represents each character as a two-byte, unsigned integer between 0 and 65,535. A document that uses UCS-2

plus “surrogate pairs” is in **UTF-16** encoding. UTF stands for **Unicode Transformation Format**.

UTF-8

UTF-8 is a variable-length encoding of Unicode. Characters 0 through 127(the ASCII character set) are encoded in 1 byte each. Pure ASCII files are also acceptable UTF-8 files. UTF-8 represents characters 128 to 2047 in 2 bytes each. The remaining characters, mostly from Chinese, Japanese, and Korean, are represented in three bytes each. When characters with code points(numeric values) above 65, 535 are added to Unicode, they will be encoded in four bytes.

UTF-8 is the most broadly supported encoding of Unicode. It is how Java .class files store strings, and it is the default encoding an XML processor uses.

Unicode code charts are at:
www.unicode.org/charts

ISO 639:1988
"Code for the representation of names of languages".

aa Afar
ab Abkhazian
af Afrikaans
am Amharic
ar Arabic
as Assamese
ay Aymara
az Azerbaijani

ba Bashkir
be Byelorussian
bg Bulgarian
bh Bihari
bi Bislama
bn Bengali; Bangla
bo Tibetan
br Breton

ca Catalan
co Corsican
cs Czech
cy Welsh

da Danish
de German
dz Bhutani

el Greek
en English
eo Esperanto
es Spanish
et Estonian
eu Basque

fa Persian
fi Finnish
fj Fiji
fo Faeroese
fr French
fy Frisian

ga Irish
gd Scots Gaelic
gl Galician
gn Guarani
gu Gujarati

ha Hausa
hi Hindi
hr Croatian
hu Hungarian
hy Armenian

ia Interlingua
ie Interlingue
ik Inupiak
in Indonesian
is Icelandic
it Italian
iw Hebrew

ja Japanese
ji Yiddish
jw Javanese

ka Georgian
kk Kazakh
kl Greenlandic
km Cambodian
kn Kannada
ko Korean
ks Kashmiri
ku Kurdish
ky Kirghiz

la Latin
ln Lingala
lo Laothian
lt Lithuanian
lv Latvian, Lettish

mg Malagasy
mi Maori
mk Macedonian
ml Malayalam
mn Mongolian
mo Moldavian
mr Marathi
ms Malay
mt Maltese
my Burmese

na Nauru
ne Nepali
nl Dutch
no Norwegian

oc Occitan
om (Afan) Oromo
or Oriya

pa Punjabi
pl Polish
ps Pashto, Pushto
pt Portuguese

qu Quechua

rm Rhaeto-Romance
rn Kirundi
ro Romanian
ru Russian
rw Kinyarwanda

sa Sanskrit
sd Sindhi
sg Sangro
sh Serbo-Croatian
si Singhalese
sk Slovak
sl Slovenian
sm Samoan
sn Shona
so Somali
sq Albanian
sr Serbian
ss Siswati
st Sesotho
su Sundanese
sv Swedish
sw Swahili

ta Tamil
te Tegulu
tg Tajik
th Thai
ti Tigrinya
tk Turkmen
tl Tagalog
tn Setswana
to Tonga
tr Turkish
ts Tsonga
tt Tatar
tw Twi

uk Ukrainian
ur Urdu
uz Uzbek

vi Vietnamese
vo Volapuk

wo Wolof

xh Xhosa

yo Yoruba

zh Chinese
zu Zulu