Internationalization

Introducing Character Sets and Encodings

intended audience: anyone who is new to internationalization and needs guidance on topics to consider and ways to get into the material on the site.

Updated 2009-05-01 09:44 (http://www.w3.org/blog/International/tag/gs-characters/)

This page provides some orientation for newcomers to Web internationalization who don't really know where to start. The aim is to ease you gently into some of the material on the site.

You can find a selection of more detailed articles using the links to the right. Once you get some ideas from this page, you will probably just use the techniques index (/International/technique-index) , the topic index (/International/resource-index) , or the site search.

What's it about?

A character set is a collection of letters and symbols used in a writing system. For example, the ASCII character set covers letters and symbols for English text, ISO-8859-6 covers letters and symbols needed for many languages based on the Arabic script, and the Unicode character set contains characters for most of the living languages and scripts in the world.

Learn more...
Character encodings for beginners (/International/questions/qa-what-is-encoding) explains some of the basic concepts about character encodings, and why you should care.

Character encodings: Essential concepts (/International/articles/definitions-characters/) provides explanations of terminology such as Unicode, character sets, coded character sets, character encodings, the document character set, and character escapes.

Characters in a character set are stored as one or more bytes in a computer. Each byte or sequence of bytes represents a given character. A character encoding is the key that maps a particular byte or sequence of bytes to particular characters that the font renders as text.

There are many different character encodings. If the wrong encoding is applied to the bytes in memory, the result will be unintelligible text. It is therefore important, if people are to read your content, that you correctly label the character encoding used.

Abo this artic

This artic was publ with publ revie If ther are thing that nee addi plea send us feed usin the link near the bott of the page

Choosing an encoding

Everyone developing content, whether content authors or programmers, should use the UTF-8 character encoding, unless there are very special reasons for using something else. (If you decide to not use UTF-8, you must choose one of the few encodings that are interoperably implemented across all browsers.)

Learn more...

HTML & CSS authors
Choosing and applying a character encoding (/International/techniques/authoring-html#choosing)

Spec developers
Choosing character encodings (/International/techniques/developing-specs#char_choosing)

Server setup
Choosing a character encoding (/International/techniques/server-setup#choosing)

Declaring and applying an encoding

Content developers and programmers must ensure that the character encoding used for a document or page is declared in the right way.

You must also ensure that your data is saved in the encoding you have chosen, it is not sufficient to just label it.

(Note that with XHTML, encoding declarations are not always straightforward; they require an understanding of 'standards' vs. 'quirks' modes (/International/articles/serving-xhtml/) , and the impact of the XML declaration.)

Content developers and webmasters may also need to ensure that the *server* delivers content with the correct character encoding declarations, since server settings can override in-document declarations.

Learn more...

HTML & CSS authors
Declaring the character encoding for HTML (/International/techniques/authoring-html#indoc)
Declaring the character encoding for a CSS style sheet (/International/techniques/authoring-html#css)

Spec developers
Identifying character encodings (/International/techniques/developing-specs#char_identifying)

Server setup
Setting the HTTP charset parameter (/International/techniques/server-setup#setting)
Setting character encoding information using .htaccess (/International/techniques/server-setup#htaccess)

Escapes

Escapes are a way of representing a character using only ASCII text. They provide a way of representing characters that are not available in the character encoding you are using, or a way of avoiding the use of the character for other reasons (such as when they may conflict with syntax). You should be clear on when and how these escapes should be used.

Learn more...

HTML & CSS authors
Using escapes to represent characters (/International/techniques/authoring-html#escapes)

SVG authors
Using escapes to represent characters (/International/techniques/authoring-svg#escapes)

XML authors
Using escapes to represent characters (/International/techniques/authoring-xml#escapes)

Spec developers
Designing character escapes (/International/techniques/developing-specs#char_escapes)

Web addresses

Web addresses can also include non-ASCII characters. The user does little other than click on the appropriate link or enter the text as they see it, the heavy lifting is done by the user agent, but you may be interested to know how this works.

Specification developers should design their specifications so that non-ASCII web addresses can be used.

Learn more...

HTML & CSS authors
Using non-ASCII web addresses (/International/techniques/authoring-html#iris)

By: Richard Ishida, W3C.

Content first published 2006-01-16. Last substantive update 2009-05-01 09:44 GMT. This version 2015-01-01 05:32 GMT

For the history of document changes, search for *gs-characters* in the i18n blog.

Copyright © 2006-2015 W3C® (MIT, ERCIM, Keio, Beihang), All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.