Dodatek – Technologie internetowe http://pl.wikipedia.org/wiki/UTF-8

1. UTF-8 wg
2. Adresy URL

Dodatek – Technologie internetowe
http://pl.wikipedia.org/wiki/UTF-8

1. UTF-8

Zalety i wady

Zalety

1. Każdy tekst w **ASCII** jest tekstem w UTF-8.
2. Żaden znak spoza ASCII nie zawiera bajtu z ASCII.
3. Zachowuje porządek sortowania **UCS-4** (UTF-32)
4. Typowy tekst ISO-Latin-X rozrasta się w bardzo niewielkim stopniu po przekonwertowaniu do UTF-8.
5. Nie zawiera bajtów 0xFF i 0xFE, więc łatwo można go odróżnić od tekstu **UTF-16.**
6. O każdym bajcie wiadomo czy jest początkiem znaku, czy też leży w jego środku
7. Nie ma problemów z kodowaniem od najstarszego bajtu z lewej strony do prawej lub z prawej do lewej.

Wady

1. Znaki z języków: **chiński**, **japoński**, **koreański,** zajmują po 3 bajty zamiast 2 w kodowaniach narodowych.
2. Znaki alfabetów niełacińskich zajmują po 2 bajty zamiast jednego w kodowaniach narodowych.
3. W chwili obecnej (2006 rok) większość zastosowań w **Internecie** (**poczta elektroniczna, usenet, HTML**) wymaga deklarowania UTF-8 zgodnie ze standardem **MIME.** Dopiero w **XHTML** UTF-8 jest kodowaniem domyślnym.
4. UTF-8 nie używa przesunięć zasięgów, co stanowi dodatkowe utrudnienie dla implementacji UTF-8 (szczegóły dalej)

Sposób kodowania

Mapowanie znaków Unikodu na ciągi bajtów:

- 0x00 do 0x7f - bity **0**xxxxxxx, gdzie iksy to bity od najwyższego licząc
- 0x80 do 0x7FF - bity **110**xxxxx **10**xxxxxx
- 0x800 do 0xFFFF - bity **1110**xxxx 1**0**xxxxxx **10**xxxxxx
- 0x10000 do 0x1FFFFF - bity **11110**xxx **10**xxxxxx **10**xxxxxx 10xxxxxx
- 0x200000 do 0x3FFFFFF - bity **111110**xx **10**xxxxxx **10**xxxxxx **10**xxxxxx **10**xxxxxx
- 0x4000000 do 0x7FFFFFFF - bity **1111110**x **10**xxxxxx **10**xxxxxx **10**xxxxxx **10**xxxxxx **10**xxxxxx

Oznacza to, że ten sam znak można zapisać na kilka sposobów.

Przykładowo znak **ASCII** / (ukośnik **00101111**) można zapisać jako:

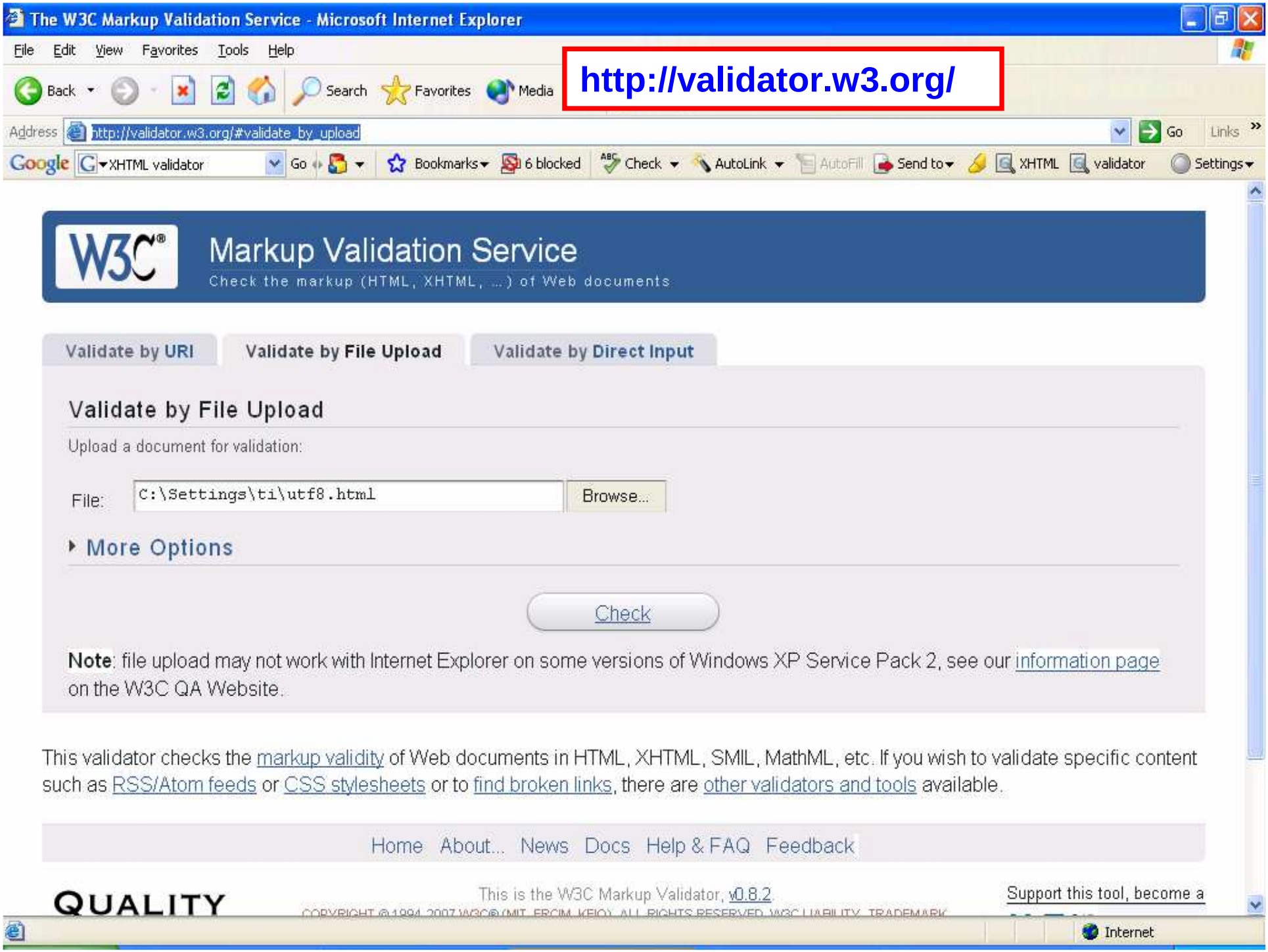
00**101111**

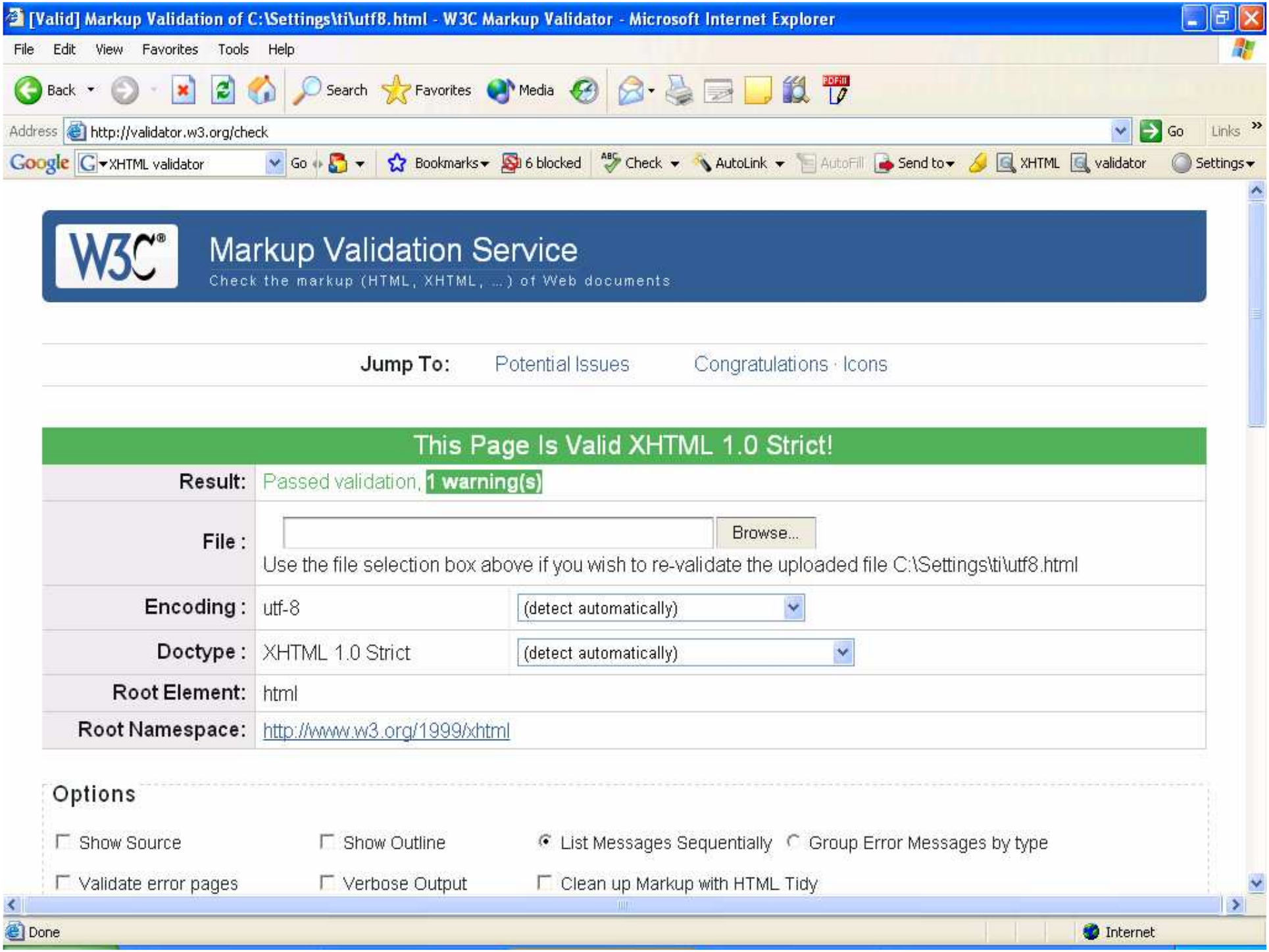
11000000 **10101111**

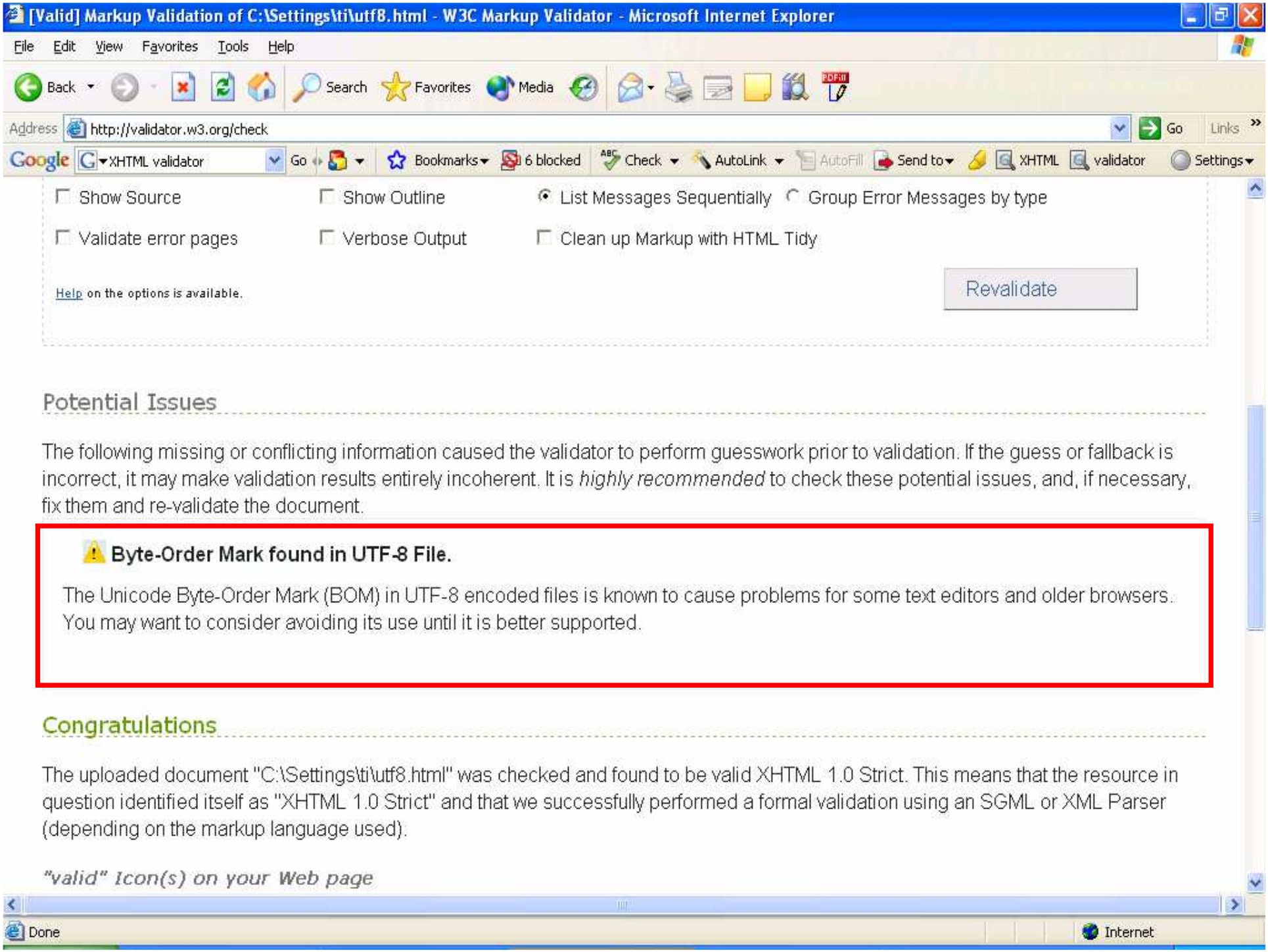
11100000 **10**000000 **10101111** itd.

Znaki polskie kodowane w UTF-8

Kod dużej litery	znak	Kod małej litery	znak
&#260;	Ą	&#261;	ą
&#262;	Ć	&#263;	ć
&#280;	Ę	&#281;	ę
&#321;	Ł	&#322;	ł
&#323;	Ń	&#324;	ń
&#346;	Ś	&#347;	ś
&#377;	Ź	&#378;	ź
&#379;	Ż	&#380;	ż
&#211;	Ó	&#243;	ó


The W3C Markup Validation Service - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
http://validator.w3.org/
Address http://validator.w3.org/#validate_by_upload
Go Links
Google XHTML validator Go Bookmarks 6 blocked Check AutoLink AutoFill Send to XHTML validator Settings
Markup Validation Service
Check the markup (HTML, XHTML, ...) of Web documents
Validate by URI
Validate by File Upload
Validate by Direct Input
Validate by File Upload
Upload a document for validation:
File:
C:\Settings\ti\utf8.html
Browse...
More Options
Check
Note: file upload may not work with Internet Explorer on some versions of Windows XP Service Pack 2, see our information page on the W3C QA Website.
This validator checks the markup validity of Web documents in HTML, XHTML, SMIL, MathML, etc. If you wish to validate specific content such as RSS/Atom feeds or CSS stylesheets or to find broken links, there are other validators and tools available.
Home About... News Docs Help & FAQ Feedback
QUALITY
This is the W3C Markup Validator, v0.8.2
Support this tool, become a
Internet


[Valid] Markup Validation of C:\Settings\ti\utf8.html - W3C Markup Validator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://validator.w3.org/check
Go Links
Google XHTML validator Go Bookmarks 6 blocked Check AutoLink AutoFill Send to XHTML validator Settings
Markup Validation Service
Check the markup (HTML, XHTML, ...) of Web documents
Jump To: Potential Issues Congratulations · Icons
This Page Is Valid XHTML 1.0 Strict!
Result: Passed validation, 1 warning(s)
File : Browse...
Use the file selection box above if you wish to re-validate the uploaded file C:\Settings\ti\utf8.html
Encoding : utf-8 (detect automatically)
Doctype : XHTML 1.0 Strict (detect automatically)
Root Element: html
Root Namespace: http://www.w3.org/1999/xhtml
Options
Show Source
Show Outline
List Messages Sequentially
Group Error Messages by type
Validate error pages
Verbose Output
Clean up Markup with HTML Tidy
Done
Internet

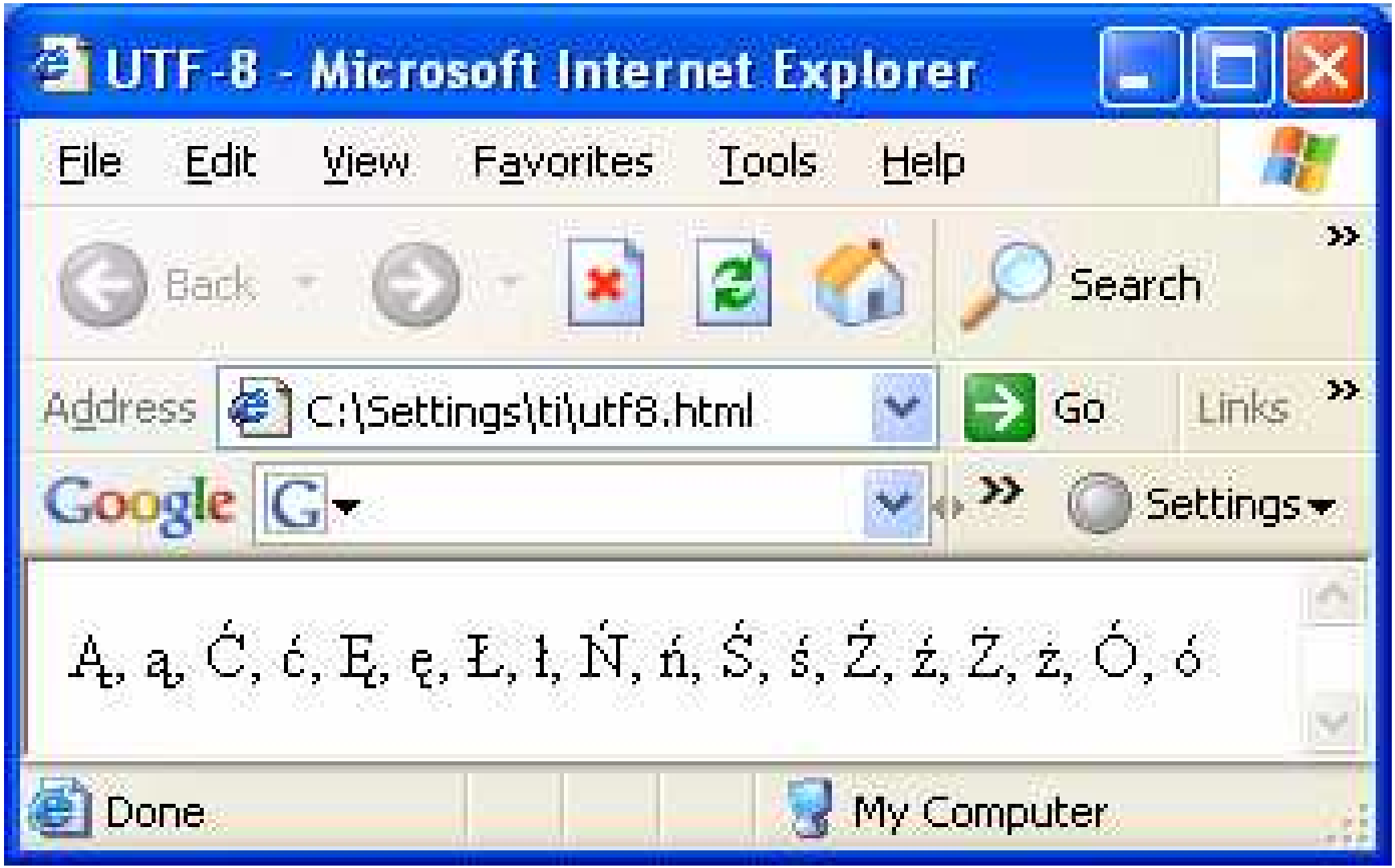

[Valid] Markup Validation of C:\Settings\ti\utf8.html - W3C Markup Validator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://validator.w3.org/check
Go Links
Google XHTML validator Go Bookmarks 6 blocked Check AutoLink AutoFill Send to XHTML validator Settings
Show Source
Show Outline
List Messages Sequentially
Group Error Messages by type
Validate error pages
Verbose Output
Clean up Markup with HTML Tidy
Help on the options is available.
Revalidate
Potential Issues
The following missing or conflicting information caused the validator to perform guesswork prior to validation. If the guess or fallback is incorrect, it may make validation results entirely incoherent. It is highly recommended to check these potential issues, and, if necessary, fix them and re-validate the document.
Byte-Order Mark found in UTF-8 File.
The Unicode Byte-Order Mark (BOM) in UTF-8 encoded files is known to cause problems for some text editors and older browsers. You may want to consider avoiding its use until it is better supported.
Congratulations
The uploaded document "C:\Settings\ti\utf8.html" was checked and found to be valid XHTML 1.0 Strict. This means that the resource in question identified itself as "XHTML 1.0 Strict" and that we successfully performed a formal validation using an SGML or XML Parser (depending on the markup language used).
"valid" Icon(s) on your Web page
Done
Internet

Przykład kodowania bezposredniego znaków polskich w kodzie UTF-8

```
<!--Komentarz-->
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="pl" lang="pl">
 <head>
<meta http-equiv="content-type" content="text/html;charset=utf-8"/>
<title> UTF-8 </title>
</head>
<body>
<p>&#260;, &#261;, &#262;, &#263;, &#280;, &#281;, &#321;, &#322;,
&#323;,
&#324;, &#346;, &#347;, &#377;, &#378;, &#379;, &#380;, &#211;,
&#243;
</p>
</body>
</html>
```

Efekt kodowania bezpośredniego



Dodatek – Technologie internetowe
http://pl.wikipedia.org/wiki/UTF-8

1. UTF-8
2. Adresy URL

Uniform Resource Locator -URL

1. **URL** (ang. ***Uniform Resource Locator***) oznacza ujednolicony format adresowania zasobów (informacji, danych, usług), stosowany w Internecie i w sieciach lokalnych.
2. **URL** najczęściej kojarzony jest z adresami stron **WWW**, ale ten format adresowania służy do identyfikowania wszelkich zasobów dostępnych w Internecie. Większość przeglądarek internetowych umożliwia dostęp nie tylko do stron WWW, ale także do innych zasobów w Internecie, po wpisaniu do przeglądarki poprawnego adresu URL danego zasobu.
3. Standard URL opisany jest w dokumencie **http://tools.ietf.org/html/rfc1738**

Część zależna od rodzaju usługi zwykle przybiera jedną z postaci:

- **W przypadku zasobów będących plikami:**
  //adres_serwera:port/sciezka_dostępu
  jeżeli port jest standardowy dla danego rodzaju zasobu, jest pomijany i stosuje się formę uproszczoną:
  //adres_serwera/sciezka_dostępu
  Niekiedy może być wymagane podanie nazwy użytkownika i hasła:
  //nazwa_użytkownika:hasło@adres_serwera/sciezka_dostępu
  ale najczęściej zarówno nazwa_użytkownika, jak i hasło nie są wymagane i mogą być pominięte.
- **W przypadku zasobów nie będących plikami (konta shellowe, adresy email itp.):**
  nazwa_uzytkownika@adres_serwera

Często oprogramowanie, szczególnie przeglądarki internetowe, akceptuje także niepoprawne formy adresów – pominięty separator // czy określenie protokołu http://, np.:
adres_serwera/sciezka_dostepu

Przykładowy URL:
http://www.wikipedia.com/wiki/URL
gdzie:http – protokół dostępu do zasobu
www.wikipedia.com – adres serwera
wiki/URL – ścieżka dostępu do zasobu