

# CS595 Intro to Web Science, Assignment #1

Valentina Neblitt-Jones

September 12, 2013


## 1 cURL Exercise

**Demonstrate that you know how to use "curl" well enough to correctly POST data to a form. Show that the HTML response that is returned is "correct" (e.g., save it to a file and then view that file in a browser and take a screen shot).**

Finding a web form that did not require me revealing my credentials and actually used the POST method was quite difficult. I used W3C Markup Validator Service [http://validator.w3.org/#validate\\_by\\_input](http://validator.w3.org/#validate_by_input). It was being extra fussy about the special characters (`-urlencode` was not solving the problem) so so I had to substitute the appropriate URL encoding using [http://www.w3schools.com/tags/ref\\_urlencode.asp](http://www.w3schools.com/tags/ref_urlencode.asp). My cURL statement follows.

```
curl --data "fragment=%3Cp%3ECurrently I am 40 years old and an academic  
systems librarian. My passions are books, movies, and television. My ultra  
-passions include Star Wars, LEGO, The Simpsons, and Harry Potter.%3C%2Fp  
%3E&prefill=1&doctype=Inline&fbid=1&prefill_doctype=html401&group=0&ss=1&st  
=1&outline=1&No200=1&verbose=1" --url http://validator.w3.org/check -o  
debug.html
```


The four screenshots below illustrate the difference in how the page was rendered when using the browser versus cURL to fill out the form. Figures 1 & 2 show the browser version and Figures 3 & 4 show the cURL version. Although the stylesheet and images are missing from the cURL version, the same information is present.


**Markup Validation Service**  
Check the markup (HTML, XHTML, ...) of Web documents


[Jump To:](#)
[Notes and Potential Issues](#)
[Congratulations - Icons](#)
[Source Listing](#)

**This document was successfully checked as HTML 4.01 Strict!**

<b>Result:</b>	Passed, 1 warning(s)
<b>Source:</b>	<pre>&lt;!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd"&gt; &lt;html&gt; &lt;head&gt; &lt;title&gt;I AM YOUR DOCUMENT TITLE REPLACE ME&lt;/title&gt; &lt;meta http-equiv="Content-Type" content="text/html; charset=UTF-8"&gt; &lt;meta http-equiv="Content-Style-Type" content="text/css"&gt; &lt;/head&gt; &lt;body&gt; &lt;div&gt; &lt;p&gt;Currently, I am 40 years old and an academic systems librarian. My passions are books, movies, and television. My ultra-</pre>
<b>Encoding:</b>	utf-8 (detect automatically)
<b>Doctype:</b>	HTML 4.01 Strict (detect automatically)
<b>Root Element:</b>	html



The W3C validators rely on community support for hosting and development.  
[Donate](#) and help us build better tools for a better web.

4369


**Options**

☒ 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

**Notes and Potential Issues**

The following notes and warnings highlight missing or conflicting information which caused the validator to perform some guesswork prior to validation, or other things affecting the output below. If the guess or fallback is incorrect, it could make validation results entirely incoherent. It is *highly recommended* to check these potential issues, and, if necessary, fix them and re-validate the document.

 Using Direct Input mode: UTF-8 character encoding assumed

Unlike the "by URI" and "by File Upload" modes, the "Direct Input" mode of the validator provides validated content in the form of characters pasted or typed in the validator's form field. This will automatically make the data UTF-8, and therefore the validator does not need to determine the character encoding of your document, and will ignore any charset information specified.

If you notice a discrepancy in detected character encoding between the "Direct Input" mode and other validator modes, this is likely to be the reason. It is neither a bug in the validator, nor in your document.

Figure 1: This is the top half of the screen representing output when the browser was used to fill out the form

## Congratulations

The uploaded document was successfully checked as HTML 4.01 Strict. This means that the resource in question identified itself as "HTML 4.01 Strict" and that we successfully performed a formal validation of it. The parser implementations we used for this check are based on [OpenSP](#) (SGML/XML).

### "valid" Icon(s) on your Web page

To show your readers that you have taken the care to create an interoperable Web page, you may display this icon on any page that validates. Here is the HTML you could use to add this icon to your Web page:

 `</img>`

A [full list](#) of icons, with links to alternate formats and colors, is available: If you like, you can download a copy of the icons to keep in your local web directory, and change the HTML fragment above to reference your local image rather than the one on this server. See also our help items related to documents transferred over secure protocols for these [icons](#) and the [uri=referer](#) feature.

### Validating CSS Style Sheets

If you use [CSS](#) in your document, you can check it using the [W3C CSS Validation Service](#).

[TOP](#)

## Source Listing

Below is the source input I used for this validation:

```
1. <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN"  
2. "http://www.w3.org/TR/html4/strict.dtd">  
3. <html>  
4. <head>  
5. <title>I AM YOUR DOCUMENT TITLE REPLACE ME</title>  
6. <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">  
7. <meta http-equiv="Content-Style-Type" content="text/css">  
8. </head>  
9. <body>  
10. <div>  
11.  
12. <p>Currently I am 40 years old and an academic systems librarian. My passions are books, movies, and television. My ultra-passions include Star Wars, LEGO, The Simpsons, and Harry Potter.</p>  
13.  
14. </div>  
15. </body>  
16. </html>
```

[TOP](#)


[Home](#) [About...](#) [News](#) [Docs](#) [Help & FAQ](#) [Feedback](#) [Contribute](#)



COPYRIGHT © 1994-2012 W3C® (MIT, ERCIM, KEIO). 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.



Figure 2: This is the bottom half of the screen representing output when the browser was used to fill out the form



# Markup Validation Service

Check the markup (HTML, XHTML, ...) of Web documents

- [Jump To:](#)
- [Notes and Potential Issues](#)
- [Congratulations - Icons](#)
- [Source Listing](#)
- [Outline](#)

**This document was successfully checked as HTML 4.01 Strict!**

**Result:** Passed, 1 warning(s)

**Source:**

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html>
<head>
<title>I AM YOUR DOCUMENT TITLE REPLACE ME</title>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
<meta http-equiv="Content-Style-Type" content="text/css">
</head>
<body>
<div>
<p>Currently I am 40 years old and an academic systems librarian. My passions are books, movies, and

```

**Modified:**

**Server:**

**Size:**

**Content-Type:**

**Encoding:** utf-8 (detect automatically)

**Doctype:** HTML 4.01 Strict (detect automatically)

**Root Element:** html



The W3C validators are hosted on server technology donated by HP, and supported by community donations.

[Donate](#) and help us build better tools for a better web.

4369

[Flash](#)

**Options**

☒ 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](#)

Figure 3: This is the top half of the screen representing output when cURL was used to fill out the form

If you notice a discrepancy in detected character encoding between the "Direct Input" mode and other validator modes, this is likely to be the reason. It is neither a bug in the validator, nor in your document.

### Congratulations

The uploaded document was successfully checked as HTML 4.01 Strict. This means that the resource in question identified itself as "HTML 4.01 Strict" and that we successfully performed a formal validation of it. The parser implementations we used for this check are based on [OpenSP](#) (SGML/XML).

#### "valid" Icon(s) on your Web page

To show your readers that you have taken the care to create an interoperable Web page, you may display this icon on any page that validates. Here is the HTML you could use to add this icon to your Web page:



```
<p>
  <a href="http://validator.w3.org/check?uri=referer"></a>
</p>
```

A [full list](#) of icons, with links to alternate formats and colors, is available: If you like, you can download a copy of the icons to keep in your local web directory, and change the HTML fragment above to reference your local image rather than the one on this server. See also our help items related to documents transferred over secure protocols for these [icons](#) and the ["uri=referer" feature](#).

#### Validating CSS Style Sheets

If you use [CSS](#) in your document, you can check it using the W3C [CSS Validation Service](#).

[↑ Top](#)

#### Source Listing

Below is the source input I used for this validation:

```
1. <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN"
2. "http://www.w3.org/TR/html4/strict.dtd">
3. <html>
4. <head>
5. <title>I AM YOUR DOCUMENT TITLE REPLACE ME</title>
6. <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
7. <meta http-equiv="Content-Style-Type" content="text/css">
8. </head>
9. <body>
10. <div>
11.
12. <p>Currently I am 40 years old and an academic systems librarian. My passions are books, movies, and television. My ultra-passions include Star Wars, LEGO, The Simpsons, and Harry Potter.</p>
13.
14. </div>
15. </body>
16. </html>
```

[↑ Top](#)

#### Document Outline

Below is an outline for this document, automatically generated from the heading tags (<h1> through <h6>.)

If this does not look like a real outline, it is likely that the heading tags are not being used properly. (Headings should reflect the logical structure of the document; they should not be used simply to add emphasis, or to change the font size.)

If you want to examine the semantic structure of your documents, beyond the outline, try the [Semantic data extractor](#).

[↑ Top](#)

Figure 4: This is the bottom half of the screen representing output when cURL was used to fill out the form

## 2 Python Exercise

Write a Python program that: (1) Takes one argument, like "Old Dominion" or "Virginia Tech" (2) takes another argument specified in seconds (e.g., "60" for one minute) (3) takes a URI as a third argument: `http://scores.espn.go.com/ncf/scoreboard?confId=80&seasonYear=2013&seasonType=2&weekNumber=2` OR `http://scores.espn.go.com/ncf/scoreboard?confId=80&seasonYear=2013&seasonType=2&weekNumber=1` OR `http://scores.espn.go.com/ncf/scoreboard?confId=80&seasonYear=2012&seasonType=2&weekNumber=1` etc. and (4) downloads the URI, finds the game corresponding to the team argument, prints out the current score (e.g., "Old Dominion 27, East Carolina 17), sleeps for the specified seconds, and then repeats (until control-C is hit).

### 3 Web Graph Structure Exercise

Consider the "bow-tie" graph in the Broder et al. paper (fig 9): <http://www9.org/w9cdrom/160/160.html> Now consider the following graph:

```
A --> B
B --> C
C --> D
C --> A
C --> G
E --> F
G --> C
G --> H
I --> H
I --> J
I --> K
J --> D
L --> D
M --> A
M --> N
N --> D
```

For the graph above, give the values for: IN, SCC, OUT, Tendrils, Tubes, and Disconnected.

Table 1: Nodes in Each Category

IN	
SCC	
OUT	
Tendrils	
Disconnected	

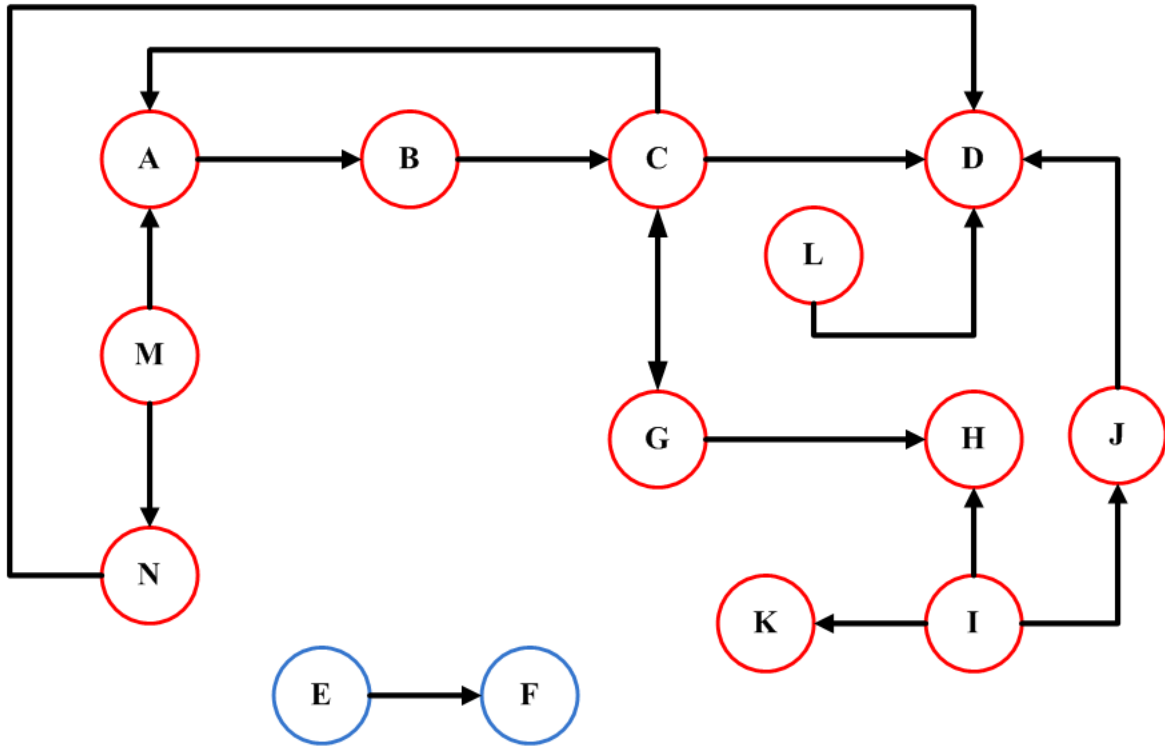


Figure 5: Visual representation of the connecting nodes