Subject: HtmlTools package for U++
Posted by Oblivion on Sun, 18 Jun 2023 23:22:03 GMT
View Forum Message <> Reply to Message

Hi,

I am happy to announce that U++ is about to gain something U++ users have been long missing: Aa very powerful html parser/sanitizer/prettifer: HtmlTools package.

This package is basically a libtidy bindings/wrapper, bringing the power and performance of one of the oldest and widely used html library to U++.

You can find the initial version of the source and example code here.

DONE:
+ Implemented HtmlNode class. (A modification of Upp::XmlNode class)
+ Implemented TidyHtmlParser, TidyHtmlParser::Node classes for traversing the document tree.
+ Implemented ParseHtml and RepairHtml convenience functions.
+ Added a minimal code example, parsing the legacy example.com.

TODO:
- Enable U++'s memory managers in libtidy.
- Add U++ callbacks for libtidy's message queue.
- Refactor buffer allocation code.
- Add Topic++ documentation.
- Add more example code (both console & gui).
- Test the Windows build.
- Cosmetics.

The base example, downloading and parsing the example.com

```
#include <Core/Core.h>
#include <Core/SSL/SSL.h>
#include <HtmlTools/HtmlTools.h>

using namespace Upp;

void PrintHtml(const HtmlNode& node)
{
 for(const HtmlNode& q : node) {
  if(q.IsTag("title"))
   Cout() << q.GatherText();
  else
```

```
  if(q.IsTag("p"))
   Cout() << q.GatherText();
  else
  if(q.IsTag("a"))
   Cout() << "For more information, see: " << q.Attr(0) << EOL;
  PrintHtml(q);
 }
}

CONSOLE_APP_MAIN
{
 StdLogSetup(LOG_FILE);
 HtmlNode n = ParseHtml(
  HttpRequest("https://example.com/").Execute(),
  { { "wrap", 96 } }); // libtidy options...
 PrintHtml(n);
}
```

Note that the package is still experimental.

Feedbacks are welcome.

Enjoy!

Best regards,
Oblivion

Subject: Re: HtmlTools package for U++
Posted by Oblivion on Tue, 20 Jun 2023 21:48:08 GMT
View Forum Message <> Reply to Message

Hi,

HtmlTools package (libtidy wrapper for++) is updated. Hopefully it will be available on UppHub very soon.

It now compiles on Windows too. The library is configured to be statically linked,

API docs are added.

Best regards,
Oblivion

Subject: Re: HtmlTools package for U++
Posted by Oblivion on Tue, 27 Jun 2023 10:43:46 GMT
View Forum Message <> Reply to Message

Hi,

HtmlTools, a libtidy wrapper for U++, is updated:

+ U++'s memory managers are enabled.
+ TidyHtmlParser::Node class gained ToHtmlNode() method. Allows any node to be converted to HtmlNode class.
+ Package is also uploaded to upp-components repo.

Best regards,
Oblivion

Subject: Re: HtmlTools package for U++
Posted by Mountacir on Sat, 01 Jul 2023 08:42:45 GMT
View Forum Message <> Reply to Message

Hi,

I tried to make something similar a couple of months ago, i was aiming for a web scraping plugin like Beautiful Soup. I managed to show amazon.com on XmlView but my code was such a miss i never got the courage to publish it.

Thank you very much Oblivion for this package.