



HTML elegant geparst mit dem Html-Agility-Pack

.Net Webseiten lesen lassen



Agenda



- Einleitung: Webservice und mehr
- Das HTML AgilityPack
- Demo 1
- Tool „HtmlVisualiser“
- Demo 2
- Fazit und Links



Webservices sind toll.



- Webservice optimale Lösung um *Applikationen* Daten über das Internet bereitzustellen
- Sehr gute Integration in VS

Was tun wenn kein Webservice geboten wird?



Manuelles HTML-Lesen



Seite wird mit Reader gelesen und über
Schlüsselwörter werden die Daten
„herausgefischt“

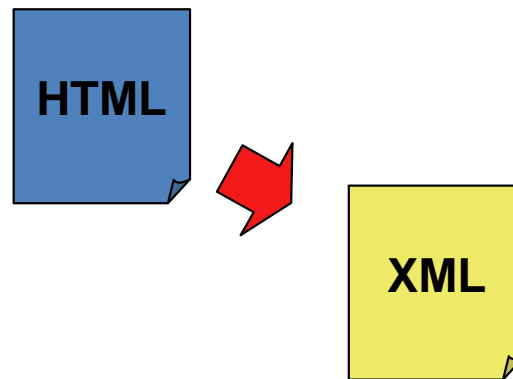
- Aufwendig
- Fall-spezifisch programmiert
- Unflexibel
- Fehleranfällig



HTML-Agility Pack



- Mächtiger HTML-Parser (pure .Net Code)
- „Agiles“ lesen von HTML-Seiten
- Produziert aus HTML-Seite gültiges XML
- Tolerant gegen gängigen (nicht XML-konformen) HTML-Code
- Daten können über XPath adressiert werden



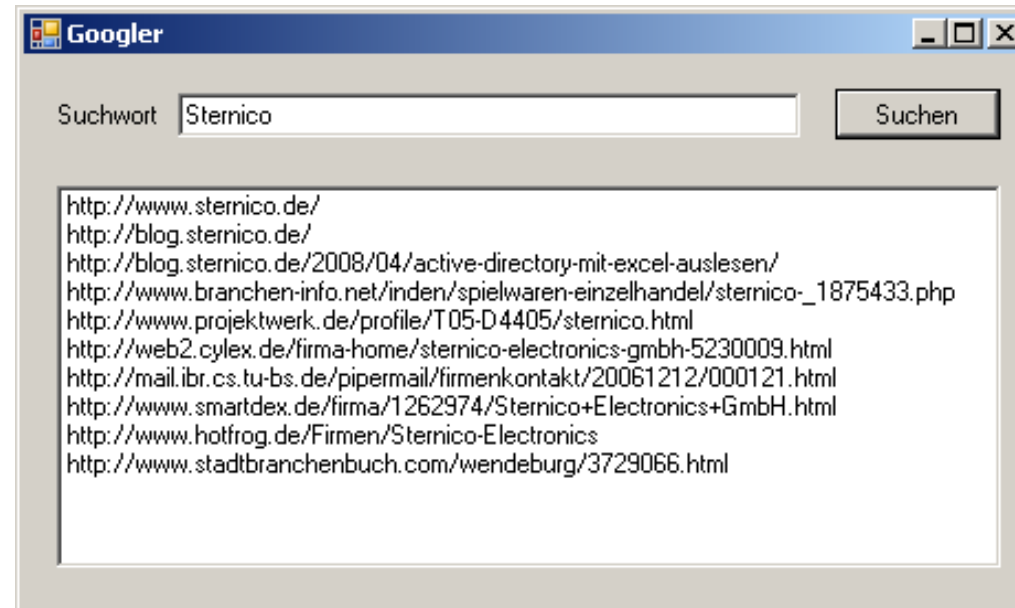
→ <http://www.codeplex.com/htmlagilitypack> ←



DEMO 1



- HTML-Agility-Pack kompilieren und einsetzen
- Beispiel: Google-Abfrage





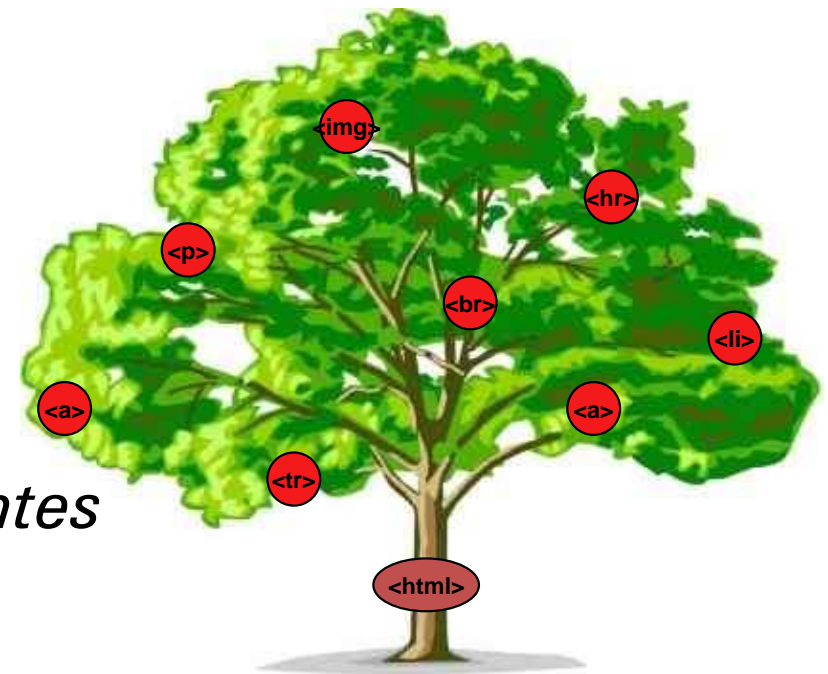
Baumgucker: Tool „HtmlVisualiser“



- Visualisiert den XML-Baum einer Website
- Ermöglicht das Testen von XPath-Abfragen

und... was sind eigentlich XPath abfragen???

„Die XML Path Language (XPath) ist eine vom W3-Konsortium entwickelte Abfragesprache, um Teile eines XML-Dokumentes zu adressieren.“



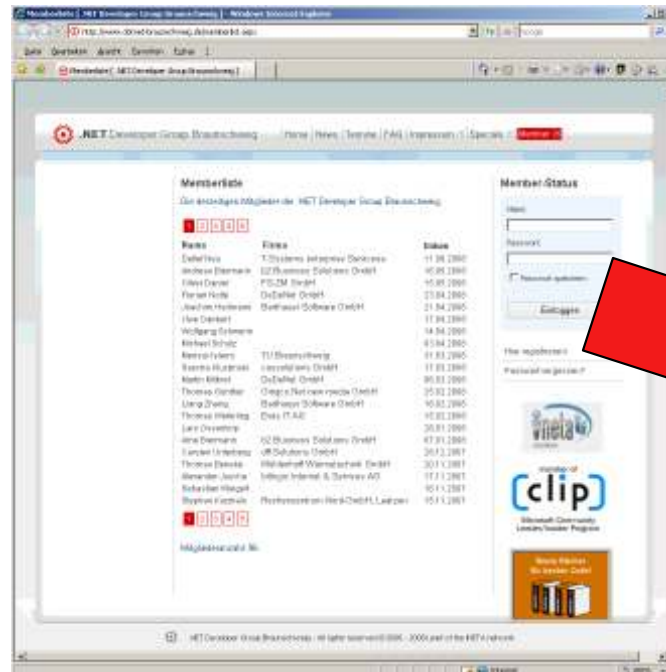


DEMO 2



Aufgabenstellung:

- Die Nutzerliste der Usergroup parsen und in ein DataSet schreiben



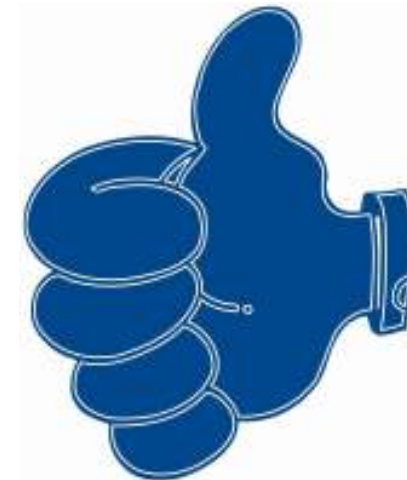
Einlesen		
Name	Firma	Datum
Daniel Neu	T-Systems entep...	11.08.2008
Andreas Biermann	b2Business Solu...	16.05.2008
Divan Daniel	FS-ZM GmbH	15.05.2008
Floian Nolte	DeDeNet GmbH	23.04.2008
Joachim Heilmann	Barthauer Solwa...	21.04.2008
Ulve Gabbart	Enbop:	17.04.2008
Wolfgang Schwe...	Enbop:	14.04.2008
Michael Schulz	Enbop:	09.04.2008
Mariusz Iskani	TU Braunschweig	31.03.2008
Sascha Kluczinski	caprolutions GmbH	17.03.2008
Marin Mibret	DeDeNet GmbH	06.03.2008
Thomas Günther	GingcoNet new...	25.02.2008
Liang Zhang	Barthauer Solwa...	18.02.2008
Thomas Weierling	Evox IT AG	15.02.2008
Ozrenkop	Enbop:	28.01.2008
U. Biermann	b2Business Solu...	07.01.2008
Tim Untertberg	dif solutions GmbH	28.12.2007
Thomas Berneke	MS#248Iberhoff...	20.11.2007



Fazit



- Html-Agility-Pack ist ausgereift, stabil und mächtig
- XPath hilft, Daten in Html zu lokalisieren
- XPath Abfragen können flexibel in Datenbank abgelegt werden





Links



- <http://www.codeplex.com/htmlagilitypack>
→ Offizielle Homepage AgilityPack
- <http://www.zvon.org/xxl/XPathTutorial/>
→ Sehr gutes Tutorial
- http://www.w3schools.com/xpath/xpath_functions.asp
→ Noch mehr zu XPath
- <http://dotnetjunkies.com/WebLog/joshuagough/archive/2006/1/20.aspx>
→ Form-Processor: Inoffizielles Addon für Forms
- <http://blog.sternico.de>
→ ☺ ein bisschen Eigenwerbung ☺