



ZoloPages

© 2008 ZoloPages

Title page 1

Use this page to introduce the product

by enter value here

This is "Title Page 1" - you may use this page to introduce your product, show title, author, copyright, company logos, etc.

This page intentionally starts on an odd page, so that it is on the right half of an open book from the readers point of view. This is the reason why the previous page was blank (the previous page is the back side of the cover)

ZoloPages

© 2008 ZoloPages

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of the publisher.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Printed: April 2011 in (wherever you are located)

Publisher

...enter name...

Managing Editor

...enter name...

Technical Editors

...enter name...

...enter name...

Cover Designer

...enter name...

Team Coordinator

...enter name...

Production

...enter name...

Special thanks to:

All the people who contributed to this document, to mum and dad and grandpa, to my sisters and brothers and mothers in law, to our secretary Kathrin, to the graphic artist who created this great product logo on the cover page (sorry, don't remember your name at the moment but you did a great work), to the pizza service down the street (your daily Capricciosas saved our lives), to the copy shop where this document will be duplicated, and and and...

Last not least, we want to thank EC Software who wrote this great help tool called HELP & MANUAL which printed this document.

Table of Contents

| | |
|--|----------|
| Foreword | 1 |
| Part I Introduction | 3 |
| 1 ZoloPages Suite | 3 |
| ZoloPages Extractor (ZPG) | 4 |
| Extractor - Surf and Collect | 5 |
| Extractor - Select and Save | 8 |
| ZoloMask Creator (ZMC) | 10 |
| Step 0 | 11 |
| Step 1 | 12 |
| Step 2 | 14 |
| Next Page | 17 |
| Options | 19 |
| Wild Cards | 21 |
| ZPG Structure..... | 22 |
| Index | 0 |

Foreword

This is just another title page
placed between table of contents
and topics

Top Level Intro

This page is printed before a new
top-level chapter starts

Part



1 Introduction

ZoloPages: Web Scraping for Dum-Dums!



ZoloPages is a Name extractor, Address extractor, Phone extractor, Fax extractor, URL extractor, Email extractor for any web page with deployed data. Fully customizable, it enables you to develop your own ZoloMasks to carry out tedious data mining tasks, such as (but not limited to) retrieving data from the white pages or pink, yellow, green pages in almost any country in the world today. You can grab data from the web in no time! ZoloPages can then save the data you've selected in Microsoft Excel CSV (text) format.

What's cool about ZoloPages is that the ZPG extraction templates it uses are incredibly simple to produce.

As easy as ABC, in three simple steps (from 0 to 2)...



Visit <http://www.zolopages.com> for more recent versions of this software and manual/help file.

Alfred Zolo is an Independent Developer, member of ASP (<http://www.asp-software.org/>^[8] - the World's #1 Trade Organization for Independent Software Developers and Vendors) and OISV (<http://www.oisv.com/> the Organization of Independent Software Vendors)

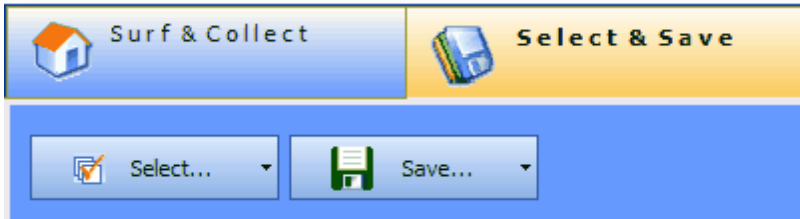
1.1 ZoloPages Suite

ZoloPages is a suite with two main applications :

1. [ZoloPagesExtractor](#)^[4]

also called [ZoloPages Extraction Engine](#)^[4] allows the capture of data from online services. The extractor is supplied with no guarantee implied, and no templates except the few found in the "**My Documents/Zolo**" folder. If you're looking for specific ZPG templates for your own country, you will have

to resort to third party web sites (including, but not limited to: <http://www.zolomask.com> in Asia).



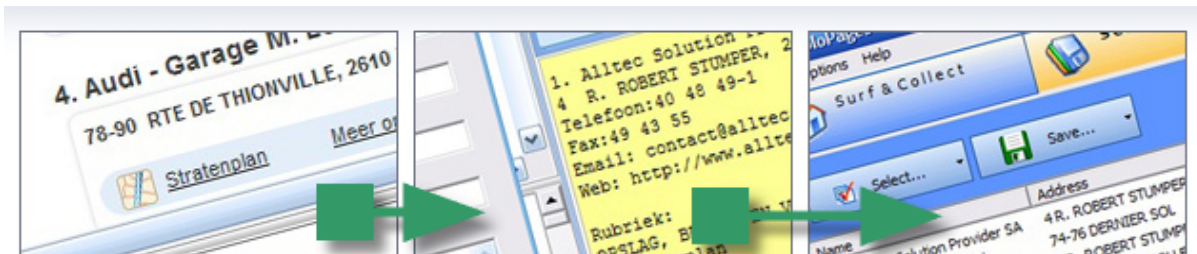
2. [ZoloMask Creator](#) ^[10]

also called [ZoloMask Editor](#) ^[10] - allows the creation of filters specific to some determined web services, prior to extraction.

| | | | | |
|----|--------------------------|-----|-------------------|---------------------|
| 18 | <input type="checkbox"/> | AOH | Single Extraction | href="http://yellow |
| 19 | <input type="checkbox"/> | AIT | Next Page by Text | Next |

Both applications are, in many ways, **extremely limited**. For more professional web scraping solutions, please visit our other web sites (<http://www.pageraptor.com> and <http://www.catchapage.com> ^[8]) and test our professional products PageRaptor and CatchaPage for free.

1.1.1 ZoloPages Extractor (ZPG)





The Extractor application features two main screens, or tabs:

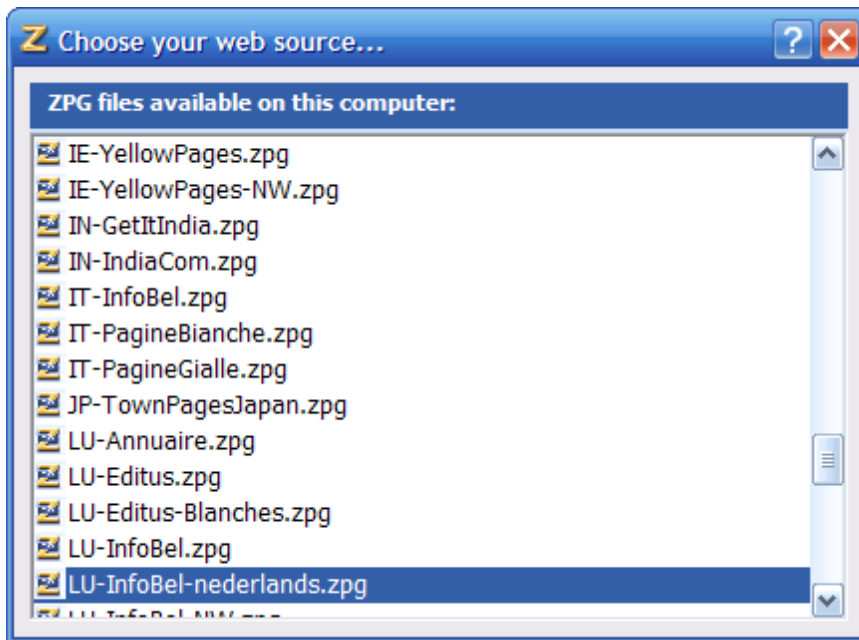
1. [Surf and Collect](#) ⁵
2. [Select and Save](#) ⁸

Visit <http://www.zolopages.com> for more recent versions of this software and manual/help file.

1.1.1.1 Extractor - Surf and Collect



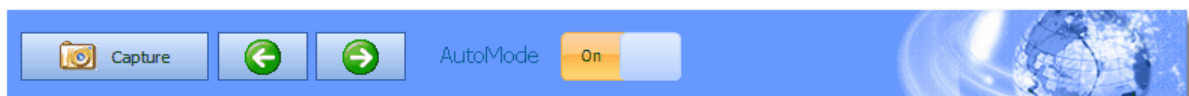
Click on FILE then OPEN in the top menu. By default this will open the "My Documents/Zolo" folder.



Select a web source. The corresponding web page will be displayed in the integrated browser.



Check AUTOMODE if you wish to capture more than one page automatically.
The two green buttons will let you go forwards (right arrow) and backwards (left arrow) one page.



Warning: AutoMode is not available in the freeware version!

For more details on the [Next Page](#)¹⁷⁾ feature, click on the link: [Next Page](#)¹⁷⁾
Then click on the CAPTURE button when you are satisfied with the data displayed on the page.



ZoloPages is now ready to capture data. It will save that data in the second tab: [Select and Save](#)

8

1.1.1.2 Extractor - Select and Save



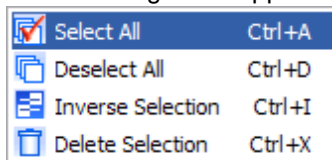
This page holds the data captured from the web page, stored in a regular grid. Please not (below) that no data has been selected at this point. No checkbox is actually selected.

| H3mmwehtsz* | Address | zipval | Pbusiness-text | SPANphonebxt |
|---|---|--------|-----------------------|---------------------------------------|
| <input checked="" type="checkbox"/> B J's Place * | 710 W Jessamine St, "Fort Worth", "TX", "76110 | 76110 | | (817) 923-2334 |
| <input checked="" type="checkbox"/> Big Apple Cafe * | 14200 Trinity Blvd, "Fort Worth", "TX", "76155 | 76155 | | (817) 572-7753 · (817) 545-5935 (fax) |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewery | 17503 Ih 10, "San Antonio", "TX", "78256 | 78256 | | (210) 690-2600 · (210) 690-3710 (fax) |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewery | 515 W Bay Area Blvd, "Webster", "TX", "77598 | 77598 | | (281) 316-3037 · (281) 316-1852 (fax) |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewery | 2231 Highway 6, "Sugar Land", "TX", "77478 | 77478 | | (281) 242-0400 |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewery | 7637 Fm 1960 Rd W, "Houston", "TX", "77070 | 77070 | | (281) 477-8777 · (281) 477-6131 (fax) |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewery | 7637 Fm 1960 Rd W, "Houston", "TX", "77070 | 77070 | | (281) 477-8777 · (281) 477-6131 (fax) |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewhouse | 1101 N Central Expy, "Plano", "TX", "75075 | 75075 | | (972) 424-4262 |
| <input checked="" type="checkbox"/> Bj's Restaurant & Brewhouse | 2609 S Stemmons Fwy, "Lewisville", "TX", "75067 | 75067 | | (972) 459-9700 |
| <input checked="" type="checkbox"/> Boston's the Gourmet Pizza * | 1827 N Loop 1604 E, "San Antonio", "TX", "78232 | 78232 | | (210) 576-1300 |
| <input checked="" type="checkbox"/> Bronx Zoo Bar & Grill * | 700 Carroll St, "Fort Worth", "TX", "76107 | 76107 | | (817) 870-0008 |
| <input checked="" type="checkbox"/> Caprock Cafe * | 3405 34th St, "Lubbock", "TX", "79410 | 79410 | Cold Beer, Big Bur... | (806) 784-0300 |
| <input checked="" type="checkbox"/> City Buzz * | "", "", "", "" | | Citybuzz.com: T... | |
| <input checked="" type="checkbox"/> Club Rodeo * | 702 Edwards Dr, "Harker Heights", "TX", "76548 | 76548 | The only country c... | (254) 213-5885 |
| <input checked="" type="checkbox"/> Fox Hole Lounge * | 794 N Harvey Mitchell Pkwy, "Bryan", "TX", "77807 | 77807 | | (979) 823-0542 |
| <input checked="" type="checkbox"/> George's Restaurant & Catering * | 1925 Speight Ave, "Waco", "TX", "76706 | 76706 | | (254) 753-1421 |
| <input checked="" type="checkbox"/> Gordon Biersch Brewery Restaurant | 8060 Park Ln, # 125, "Dallas", "TX", "75231 | 75231 | | (214) 369-2739 · (214) 361-8589 (fax) |
| <input checked="" type="checkbox"/> Gordon Biersch Brewery Restaurant - ... | 8060 Park Ln, Ste 125, "Dallas", "TX", "75231 | 75231 | | (214) 369-2739 |
| <input checked="" type="checkbox"/> Jackpot Saloon Sports Bar & Grill * | 704 Edwards Dr, "Harker Heights", "TX", "76548 | 76548 | Comfortable atmo... | (254) 449-9157 |
| <input checked="" type="checkbox"/> Katz Steakhouse & Club 21 * | 317 N Mesquite St, "Corpus Christi", "TX", "78401 | 78401 | | (361) 884-1221 |
| <input checked="" type="checkbox"/> Leapin' Lizard Pub * | 302 E Commerce St, "San Antonio", "TX", "78205 | 78205 | | (210) 271-9494 |
| <input checked="" type="checkbox"/> Little O's Patio Bar & Grill * | 4650 Little Rd, "Arlington", "TX", "76017 | 76017 | | (817) 561-0000 |
| <input checked="" type="checkbox"/> No Frills Grill * | 4914 Little Rd, "Arlington", "TX", "76017 | 76017 | | (817) 478-1766 |
| <input checked="" type="checkbox"/> Oxbow Steakhouse And BBQ * | Belton, TX 76513 zip code | 76513 | | (254) 939-0588 |
| <input checked="" type="checkbox"/> Paradise Billiards * | 5141 Oakhurst Dr, "Corpus Christi", "TX", "78411 | 78411 | | (361) 774-7401 |

Select the data you wish to save to disk: either individually, or several at a time by *right-clicking* on the SELECT button.



The following menu appears:



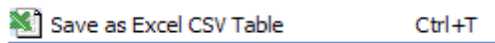
With these menu items you can thus SELECT, UNSELECT all records. You can also inverse your selection and even delete part or all of it.

e

Save the data you've selected by *right-clicking* on the SAVE button.

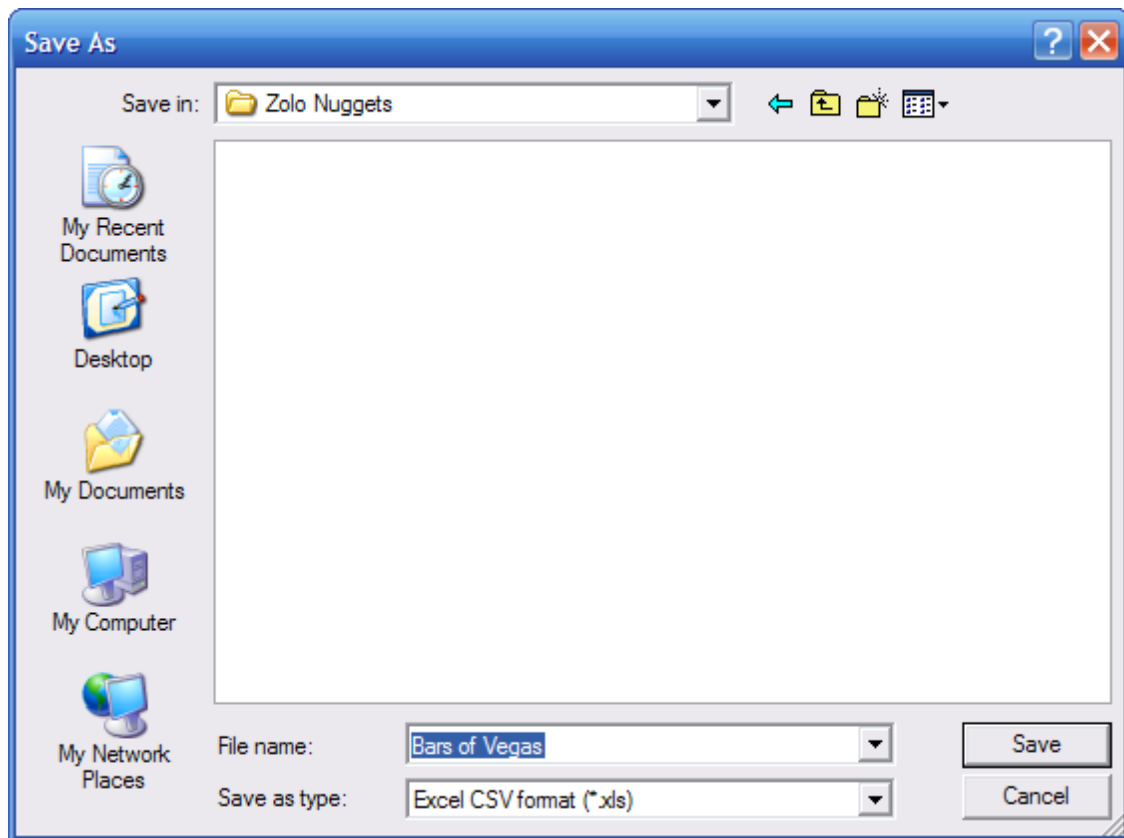


The following menu appears:



You can save to CSV (text-based) for Excel and Access.

Click on the item of your choice. The following dialog will pop up.



Give the new file the name of your choice and save it where you see fit. That's it!

You can now open the same file in Excel (Open as text file/CSV). Here's the result, below:

Microsoft Excel - ZoloPages Data

File Edit View Insert Format Tools Data Window Help

Type a question for help

100%

10 B

Import ODF Export ODF

A12 Caprock Cafe *

| | A | B | C | D | E | F | G |
|----|--------------------------|--------------------------|----------------|----|-------|-------|--------------------------------|
| 1 | B J's Place * | 710 W Jessamine St | Fort Worth | TX | 76110 | 76110 | (817) 923-2334 |
| 2 | Big Apple Cafe * | 14200 Trinity Blvd | Fort Worth | TX | 76155 | 76155 | (817) 572-7753 · (817) 545-593 |
| 3 | Bj's Restaurant & Brew | 17503 Ih 10 | San Antonio | TX | 78256 | 78256 | (210) 690-2600 · (210) 690-37 |
| 4 | Bj's Restaurant & Brew | 515 W Bay Area Blvd | Webster | TX | 77598 | 77598 | (281) 316-3037 · (281) 316-18 |
| 5 | Bj's Restaurant & Brew | 2231 Highway 6 | Sugar Land | TX | 77478 | 77478 | (281) 242-0400 |
| 6 | Bj's Restaurant & Brew | 7637 Fm 1960 Rd W | Houston | TX | 77070 | 77070 | (281) 477-8777 · (281) 477-61 |
| 7 | Bj's Restaurant & Brew | 7637 Fm 1960 Rd W | Houston | TX | 77070 | 77070 | (281) 477-8777 · (281) 477-61 |
| 8 | Bj's Restaurant & Brew | 1101 N Central Expy | Plano | TX | 75075 | 75075 | (972) 424-4262 |
| 9 | Bj's Restaurant & Brew | 2609 S Stemmons Fwy | Lewisville | TX | 75067 | 75067 | (972) 459-9700 |
| 10 | Boston's the Gourmet F | 1827 N Loop 1604 E | San Antonio | TX | 78232 | 78232 | (210) 576-1300 |
| 11 | Bronx Zoo Bar & Grill * | 700 Carroll St | Fort Worth | TX | 76107 | 76107 | (817) 870-0008 |
| 12 | Caprock Cafe * | 3405 34th St | Lubbock | TX | 79410 | 79410 | Cold Beer, (806) 784-0300 |
| 13 | City Buzz * | | | | | | --Citybuzz.com: The video insi |
| 14 | Club Rodeo * | 702 Edwards Dr | Harker Heights | TX | 76548 | 76548 | The only c (254) 213-5885 |
| 15 | Fox Hole Lounge * | 794 N Harvey Mitchell Pk | Bryan | TX | 77807 | 77807 | (979) 823-0542 |
| 16 | George's Restaurant & | 1925 Speight Ave | Waco | TX | 76706 | 76706 | (254) 753-1421 |
| 17 | Gordon Biersch Brewer | 8060 Park Ln, # 125 | Dallas | TX | 75231 | 75231 | (214) 369-2739 · (214) 361-85 |
| 18 | Gordon Biersch Brewer | 8060 Park Ln, Ste 125 | Dallas | TX | 75231 | 75231 | (214) 369-2739 |
| 19 | Jackpot Saloon Sports | 704 Edwards Dr | Harker Heights | TX | 76548 | 76548 | Comfortabl (254) 449-9157 |
| 20 | Katz Steakhouse & Clu | 317 N Mesquite St | Corpus Christi | TX | 78401 | 78401 | (361) 884-1221 |
| 21 | Leapin' Lizard Pub * | 302 E Commerce St | San Antonio | TX | 78205 | 78205 | (210) 271-9494 |
| 22 | Little O's Patio Bar & G | 4650 Little Rd | Arlington | TX | 76017 | 76017 | (817) 561-0000 |
| 23 | No Frills Grill * | 4914 Little Rd | Arlington | TX | 76017 | 76017 | (817) 478-1766 |

ZoloPages Data/

Draw AutoShapes

Ready Sum=158820 NUM

1.1.2 ZoloMask Creator (ZMC)



ZoloMask Editor is a tool to create filters usable by the ZoloPages extraction program (the extractor).

What you need to do is to select data from the page, and save the template elements to disk in three easy steps.

Creating a new ZoloMask - Step by step process:

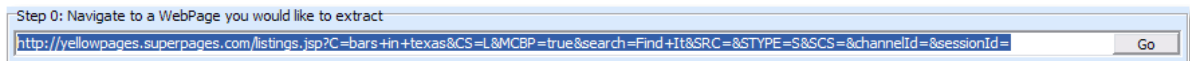
[Step 0](#)¹¹: Finding the the data relevant to your query (on the page)

[Step 1](#)¹²: Selecting what needs to be captured, applying a mask

[Step 2](#)¹⁴: Saving your ZPG template to disk

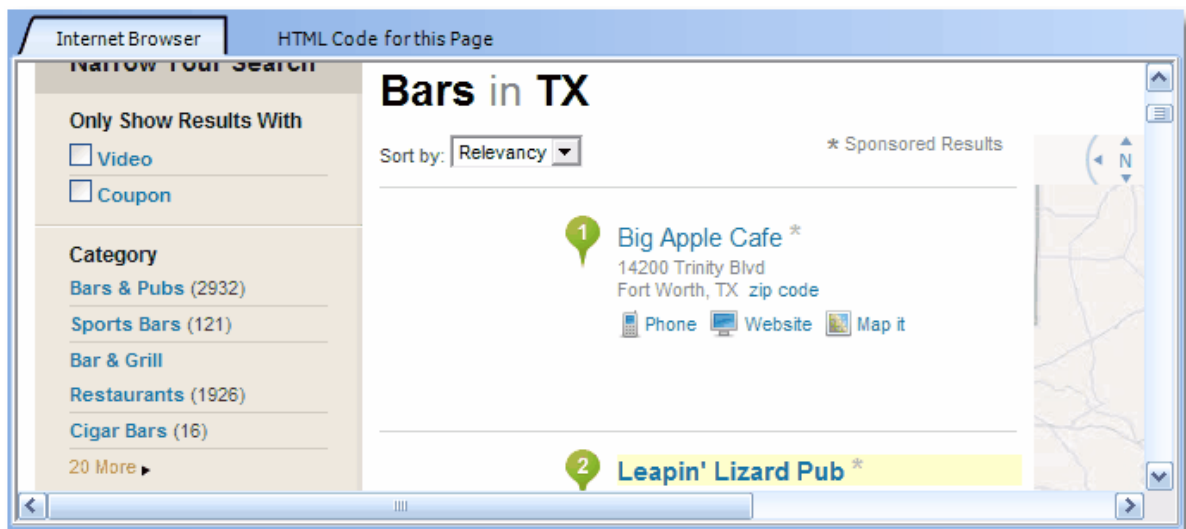
1.1.2.1 Step 0

STEP 0: Navigate to a any web page you'll need to capture data from (first tab in left pane). Enter URL and press ENTER.



As you can see we have chosen to visit the US SuperPages.

Wait for the page to display properly below. This may take some time depending on the source.



For all intents and purposes the HTML Code for the current page is also visible in a memo when clicking on the second tab. However we won't use it as such. It just might prove useful in case you need to inspect a clickable link on the page.

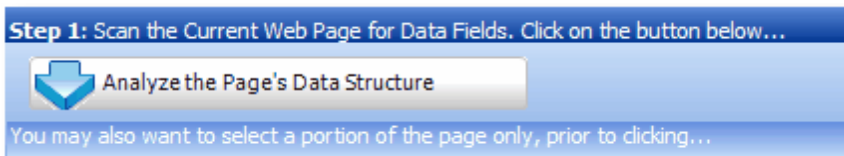
```

Internet Browser  HTML Code for this Page
1 <DIV class=master sizcache="4" sizset="0"><!--header result page-->
2 <SCRIPT language=JavaScript type=text/javascript src="http://yellowpages.superpages.com/cc
3
4 <SCRIPT language=JavaScript type=text/javascript src="http://yellowpages.superpages.com/pr
5
6 <SCRIPT type=text/javascript>var cookieDomain = ".superpages.com";</SCRIPT>
7
8 <DIV id=city_image class="masthead " sizcache="4" sizset="0"><A class=logoc title="Yellow E
9 <DIV class=sociallinks sizcache="4" sizset="1">
10 <UL sizcache="4" sizset="1">
11 <LI class=hdtwitter sizcache="4" sizset="1"><A title="Follow us on twitter" href="http://w
12 <LI class=hdfacebook><IFRAME style="BORDER-RIGHT-WIDTH: 1px; WIDTH: 110px; BORDER-TOP-WIDT
13 <DIV id=header_link class=hdaccount-links sizcache="4" sizset="2">
14 <UL sizcache="4" sizset="2">
15 <LI class=left sizcache="4" sizset="2"><A title="Advertise with Superpages.com" href="http
16 <LI id=spSignin class=center sizcache="4" sizset="3"><A id=spSigninRegistration title="Sig
17 <LI id=spFbSignin class=right sizcache="4" sizset="4"><A class=facebookhd title="Facebook

```

1.1.2.2 Step 1

STEP 1: Click on the "Analyze Page Data Structure" button



What you can see in the bottom part of the screen is roughly this:

| | | | | |
|-----|--------------------------|--------------------------|-------------------|---|
| 250 | <input type="checkbox"/> | DIVresult rtitle | Single Extraction | Leapin' Lizard Pub * 3 reviews Write a Review 302 E Commerce St San |
| 251 | <input type="checkbox"/> | DIVmptop1top | Single Extraction | Leapin' Lizard Pub * 3 reviews Write a Review 302 E Commerce St San |
| 252 | <input type="checkbox"/> | DIVmpbtm1btm dear | Single Extraction | Leapin' Lizard Pub * 3 reviews Write a Review 302 E Commerce St San |
| 253 | <input type="checkbox"/> | DIVlisting2info pin2 | Single Extraction | Leapin' Lizard Pub * 3 reviews Write a Review 302 E Commerce St San |
| 254 | <input type="checkbox"/> | DIVvlistingdiv 1 | Single Extraction | Leapin' Lizard Pub * 3 reviews Write a Review 302 E Commerce St San |
| 255 | <input type="checkbox"/> | H3nmwehtszdf nmwehtdrred | Single Extraction | Leapin' Lizard Pub * |
| 256 | <input type="checkbox"/> | A1 | Single Extraction | http://www.superpages.com/bp/San-Antonio-TX/Leapin-Lizard-Pub-L203920 |
| 257 | <input type="checkbox"/> | DIVmapreviews | Single Extraction | 3 reviews Write a Review |
| 258 | <input type="checkbox"/> | SPAN | Single Extraction | 3 reviews Write a Review |
| 259 | <input type="checkbox"/> | A | Single Extraction | http://www.superpages.com/bp/San-Antonio-TX/Leapin-Lizard-Pub-L203920 |
| 260 | <input type="checkbox"/> | Reviewlink | Single Extraction | Write a Review |

You will now proceed to select the page elements that are key to your extraction. In the example above you will thus check the box located @ **line 255** in the grid, which represents the name element of the data entry you wish to capture. As the **first element of the data** you are capturing it **needs to be checked first in your grid!**

| | | | | |
|-----|-------------------------------------|---------------------------|-------------------|----------------------|
| 255 | <input checked="" type="checkbox"/> | H3nmwehtszdf nmwehtclrred | Single Extraction | Leapin' Lizard Pub * |
|-----|-------------------------------------|---------------------------|-------------------|----------------------|

A Wild Card: *

But as you will soon find out the sequence "H3nmwehtszdf nmwehtclrred nmwehtstybckclr" is not present in every line on the US SuperPages.

However the initial block "H3nmwehtszdf" is. We will therefore edit this cell in the grid, leave out the end part of the tag, and truncate it with the "*" (star) wild card, like in the image below:

| | | | | |
|-----|-------------------------------------|---------------|-------------------|--------------------|
| 255 | <input checked="" type="checkbox"/> | H3nmwehtszdf* | Single Extraction | Leapin' Lizard Pub |
|-----|-------------------------------------|---------------|-------------------|--------------------|

Our extractor will thus now capture every tag starting with "H3nmwehtszdf".

Continue selecting all the tags you want captured.

| | | | | |
|-----|-------------------------------------|----------------------|-------------------|---|
| 262 | <input checked="" type="checkbox"/> | Address | Single Extraction | 302 E Commerce St**San Antonio, TX 78205 zip code |
| 263 | <input checked="" type="checkbox"/> | SPANzipVal1zipval | Single Extraction | 78205 |
| 264 | <input type="checkbox"/> | SPANzipLabel1ziplebl | Single Extraction | zip code |
| 265 | <input type="checkbox"/> | A | Single Extraction | javascript:showZip('1'); |
| 266 | <input checked="" type="checkbox"/> | DIVsphoneLink1 | Single Extraction | (210) 271-9494 |

CAREFUL! You will notice that line 263 contains a tag that is also a reference to the number of lines present on the web page and the order of the data entry on the page: "SPANzipVal**1**zipval". What you can do in this case is simply edit the cell and keep only the very last part of the tag, thus "zipval".

| | | | | |
|-----|-------------------------------------|--------|-------------------|-------|
| 263 | <input checked="" type="checkbox"/> | zipval | Single Extraction | 78205 |
|-----|-------------------------------------|--------|-------------------|-------|

But if you are starting to understand the logic of our extractor, you might already have guessed that the wild card * would work too: SPANzipVal*zipval"

Single vs multiple extraction

In most cases you will need to capture one page element or tag, and **one tag only**. Then you won't have to change **cell 3** at all.

| | | | | |
|-----|-------------------------------------|----------------------|-------------------|---|
| 262 | <input checked="" type="checkbox"/> | Address | Single Extraction | 302 E Commerce St**San Antonio, TX 78205 zip code |
| 263 | <input checked="" type="checkbox"/> | SPANzipVal1zipval | Single Extraction | 78205 |
| 264 | <input type="checkbox"/> | SPANzipLabel1ziplebl | Single Extraction | zip code |
| 265 | <input type="checkbox"/> | A | Single Extraction | javascript:showZip('1'); |
| 266 | <input checked="" type="checkbox"/> | DIVsphoneLink1 | Single Extraction | (210) 271-9494 |

Sometimes you might need to repeat a capture. This is often the case for phone/fax numbers. Then please edit column 3 so as to select **Multiple Extraction** instead of Single Extraction.

| | | | | |
|-----|-------------------------------------|--------------|---------------------|----------------------|
| 335 | <input checked="" type="checkbox"/> | SPANphonetxt | Multiple extraction | (381) 807-3002 |
| 336 | <input checked="" type="checkbox"/> | SPANphonetxt | Single Extraction | (281) 855-5213 (fax) |

In the example above, even if you check both lines # 335 and # 336, the extractor would only scrape the phone number, and not the fax number at all. With *Multiple Extraction* selected in the cell, both will be captured.

Applying a mask (A and 0)

Take a look at address line # 262 above. What you see in the cell is correct, but inelegant:

302 E Commerce St.-San Antonio, TX 78205 zip code

The address shows on one line, without any clearly delineated fields. In order to give this line some clarity we will edit the value cell, and make it look like this:

{*}--{*}, {AA} {00000}

By the use of the { marker the extractor is told to use a mask, which is a simple expression with a distinct pattern.

The expression here reads like this:

1. Take any character {*} followed by two --
2. followed by any character before a comma {*},
3. followed by any two capital letters (represented by "A") {AA}
4. followed by any five figures (represented by "0") {00000}

Concretely the mask here will match

{302 E Commerce St}--{San Antonio}, {TX} {78205}

or, symbolically

{StreetAddress}--{City}, {State} {ZipCode}

Line # 262 now looks like this:

| | | | | |
|-----|-------------------------------------|---------|-------------------|------------------------|
| 262 | <input checked="" type="checkbox"/> | Address | Single Extraction | {*}--{*}, {AA} {00000} |
|-----|-------------------------------------|---------|-------------------|------------------------|

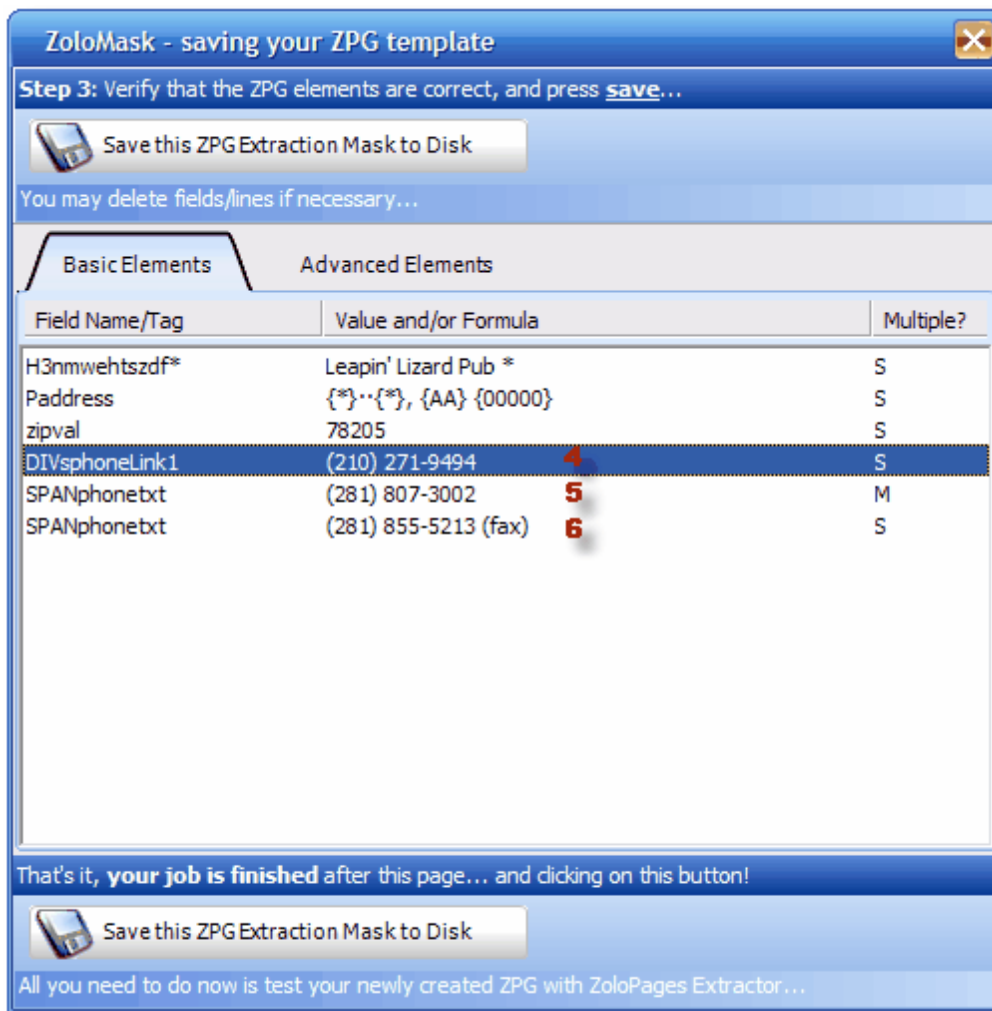
You are now ready to save your ZPG extraction template!

1.1.2.3 Step 2

STEP 2: Review your template elements and wild card masks, and save.

Basic elements

After clicking on the Save Button, a dialog box (below) pops up.



What you need to do at this point is to make sure all the elements needed for the extraction are present in the ZPG template. In the list above, you have the name taken care of, the address also (with a wild card mask in the second column), and the telephone number.

What's wrong?

Nothing much: you simply have too many elements listed. In fact you'd probably need 4 such elements instead of 6.

Lines 4 and 6 are not indispensable. Why?

Because:

1. Line 4 with the tag `DIVsphoneLink1` is a repetition of line 5, based on the same pattern. Besides the "1" at the end of the tag looks ominously like a data line number (to be avoided!)
2. Line 6 (the fax number) will not be captured as such, because it holds the same tag line as line 5: `SPANphonetxt`. Only the phone number will be scraped because it comes first in the order of the page.

Whereas:

- Line 5 can take care of both the phone number and the fax number by turning the `SINGLE EXTRACTION`

(S) option to `MULTIPLE EXTRACTION (M)`.

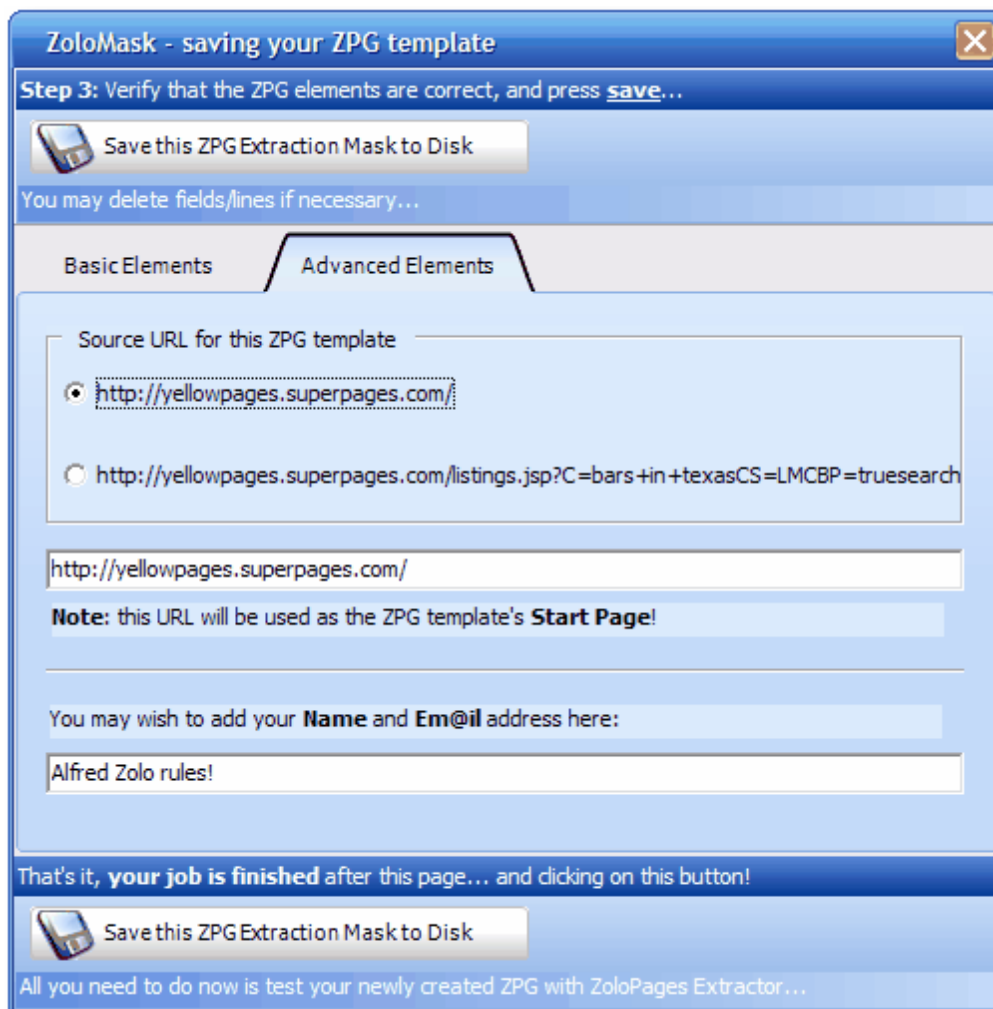
What you thus need to do is select and delete lines 6 and 4. And keep line 5 alive.

Close the dialog box if you need to make other changes without saving.

Advanced elements

All you need to do to complete your task is to select a start page for the ZPG template. Typically you will choose the root of the web server (thus <http://yellowpages.superpages.com/> in our example).

But you might like to choose the current page instead: <http://yellowpages.superpages.com/listings.jsp?C=bars+in+texas&CS=L&MCBP=true&search=Find+It&SRC=&STYPE=S&SCS=&channelId=&sessionId=>



Finally, if you are very proud of yourself for creating such a great ZPG file, you might want to sign it with your name (and email address?).

It is really up to you... Whatever you choose, don't forget to click on the "**Save ZPG Extraction Mask to Disk**".

Your work is finished. Now you can close the application, and start using the ZoloPages extractor with your newly created ZPG file!

1.1.2.4 Next Page

This topic deals with the **NextPage** and **AutoMode** features.

In it you will learn how to teach ZoloPages which link to click to turn to the next page; and you will also scan a portion of the web page and analyze its HTML properties.

Warning: AutoMode is not available in the freeware version!

Finding the next-page link

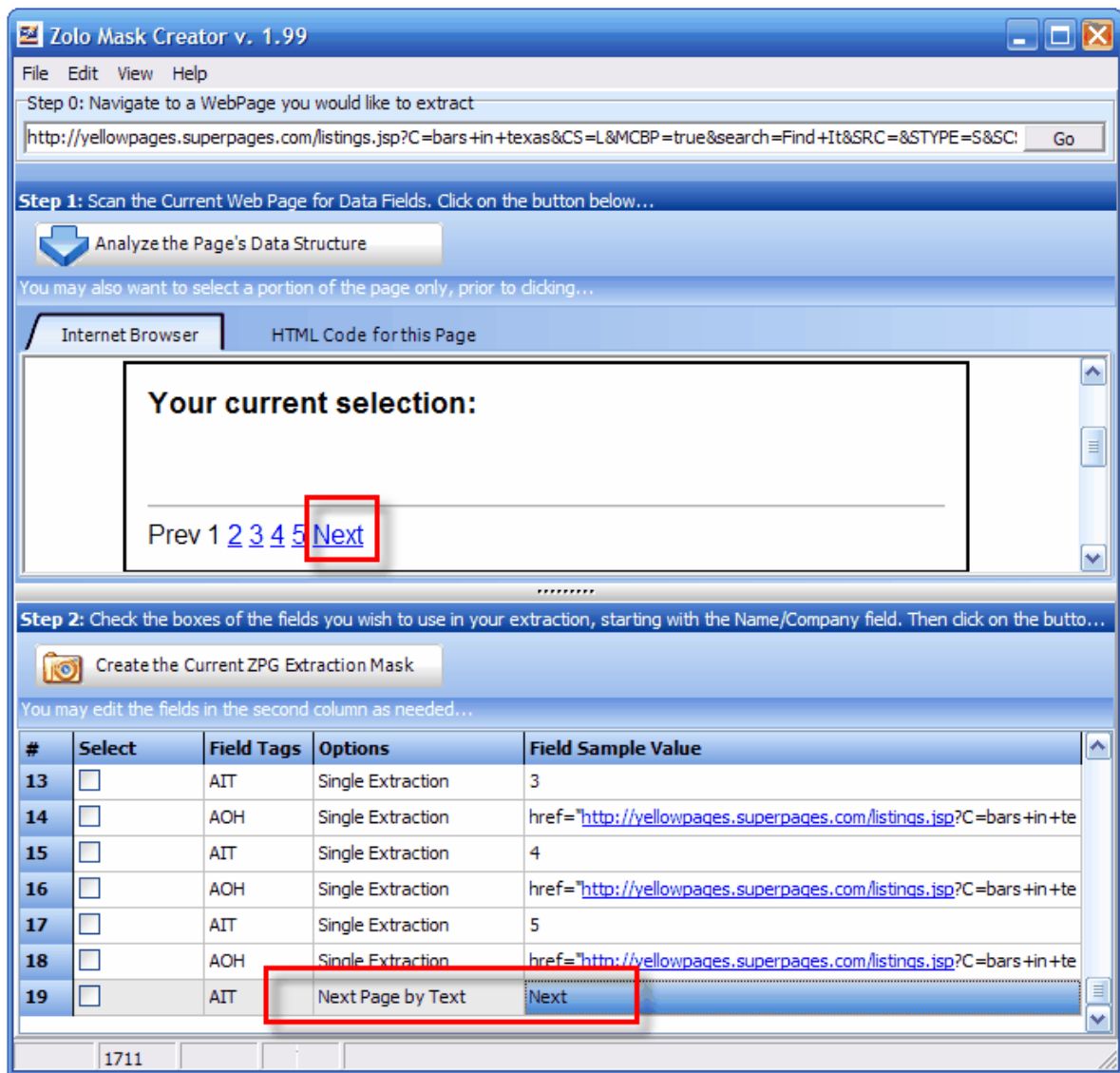
The bottom part of a yellow type of page often looks like this:



You have already scraped the initial page, and need to move on to the next. What do you do then?

Just select the portion of text above with your mouse, and then click on "**Analyze Page Data Structure**".

What you now see in the browser and in the grid is this:



Try to identify the link that, when clicked, will turn automatically to the next page. When you have found it, simply select **"Next Page by..."** in the Options column (see screenshot above).

You have clearly defined the next page as being invoked by clicking on a link whose **text** says **"Next"**;

Other choices at your disposal may also be located in the **ID tag** (as in, for instance, ` > `) or in the **class name** itself (such as ` 2 `).

In the ZPG file

The structure of the ZPG extraction template is modified thus: a section is added to the first items recorded (see [ZPG_Structure](#)^[22]), which looks like this.

```
[NextPage]
Type=NextPageByText
```

```
Pattern=Next
```

For the other possibilities mentioned above, the last two fields would be different, such as:

```
[NextPage]
Type=NextPageByID
Pattern=next-page-nav
```

```
[NextPage]
Type=NextPageByClass
Pattern=next_page
```

Four (4) possible values

The four values that you can use to turn a page are:

```
1.NextPageByText
2.NextPageByID
3.NextPageByClass
4.NextPageByTitle
```

Text, ID, Class and Title are all values that are found in a clickable link on a web page, such as:

```
<a href="http://www.zolopages.com/page2" Title="Next Page" id="next-page-intopic"
class="-next_page-class">Next Page in Topic (of course this would be the TEXT part)</a>
```

Not a dead shot!

Turning to the next page in ZoloPages Extractor is not an exact science. This cool tool is way too limited in itself to handle all possibilities used by the web source you're trying to mine data from.

For more professional web scraping solutions, please visit our other web sites (<http://www.pageraptor.com> and <http://www.catchapage.com>^[8]) and test our professional products PageRaptor and CatchaPage for free.

1.1.2.5 Options

The options you can get the extractor to use are the following, category by category:

1. Single or Multiple extraction

Tells the extractor to use a tag pattern once or several times. This is helpful when you have several telephone numbers to scrape, for instance.

[See Step 1](#)^[12]

2. Turn to Next Page options

There are four such option types, for which you also need to specify a pattern in the value column.

See also [Next Page topic](#)¹⁷.

Tells the extractor which method (type) to use and which pattern to apply when attempting to turn to the next page.

```
1 .NextPageByText
2 .NextPageByID
3 .NextPageByClass
4 .NextPageByTitle
```

Note: in the freeware version the page will turn automatically if such an option is used, but scraping won't start again until you click on the CAPTURE button again.

Our other tools, [pageraptor](#) and [catchapage](#) turn pages and scrape on doggedly!

2. Mask options

There are three such option masks:

1. **Email:** will find an email address in the line, such as "alfred@zolopages.com" in the line "please contact alfred@zolopages.com in case of a problem".
 2. **HTTP:** will extract a properly formed URL from a line, such as "<http://www.catchapage.com>" in line "Visit <http://www.catchapage.com> if you need a more powerful tool!".
 3. **Phone:** will extract any telephone number from a line, such as "(310).272.4141" in line "Call me at home, my number is (310).272.4141, if you need to chat".
-

3. Toggle telephone option

Sometimes a telephone number won't show on the page right away. You'll need to click on it before it is displayed and you can capture it.

This is the case for the Spanish yellow pages (paginasamarillas.es).

ZoloPages will do the clicking for you if you tell it where to click.

Select the string that needs to be matched (e.g. "teléfono" for Spain's Pages), and use this option. All phone numbers will show. Sheer magic! ;-)

4. Capitalize field

This is useful when some field data shows in lower case.

This is the case for Italy's yellow pages (paginegialle.it), where the company name will look like this: "[ristorante fellini](#)".

Using this option will process the name/field, and turn it into: "[Ristorante Fellini](#)".

This is it for now. Other options will be introduced as the need arises.

1.1.2.6 Wild Cards

A limited number of wild cards are available to help you work with variable elements in the HTML code.

1. ***** (star)

--> replaces any number of characters (but not line breaks).

Example:

H3nmwehtszdf*

matches

H3nmwehtszdf nmwehtclrdf

and

H3nmwehtszdf hbsgytrzesq

SPANzipVal*zipval

matches

SPANzipVal1zipval

SPANzipVal2zipval

SPANzipVal3zipval

SPANzipVal4zipval

SPANzipVal5zipval

2. **A** and **0** (any letter/any figure)

--> **Substitution for letters and numbers.**

Examples:

AA000

captures

AZ123, CA325, MN845, etc.

3. **{** and **}** (opening and closing brackets)

--> **begins and ends a capture block.**

Examples:

{AA}

captures

AZ, CA, MN, etc.

{00000}

captures

90210, 57050, 69006, etc.

Example

One for the road:

British Zip Code

```
{AA?00?A? 0A?A?}
```

captures

E17 3HX

SW4 7AA

SW18 4DW

EC1A 9LH

etc.

1.1.2.7 ZPG Structure

What you have managed to save to your hard drive is a simple text file (which you can edit with NotePad), structured this way:

```
[WebSource]
URL=http://www.superpages.com
```

```
[Fields]
FieldCount=4
0=H3nmwehtsz*
1=Paddress
2=zipval
3=SPANphonetxt
```

```
[Formulas]
1={*}..{*}, {AA} {00000}
```

```
[Multiple]
3=1
```

It is pretty self-explanatory.

So you see: you might actually create new ZPG templates without using ZoloMask at all!

Now, off you go: your data is waiting.

If you find this software useful, kindly drop me a line. And please don't forget to visit our "professional" web sites and start using our other nifty tools to "scrape the world"!

Test PageRaptor and CatchaPage for free: (<http://www.pageraptor.com> and <http://www.catchapage.com> ⁸). They are way more sophisticated than my very basic ZoloPages freeware app!

Best,

Alfred Zolo
alfred@zolopages.com

Alfred Zolo is an Independent Developer, member of ASP (<http://www.asp-software.org/> ⁸) - the World's #1 Trade Organization for Independent Software Developers and Vendors) and OISV (<http://www.oisv.com/> the Organization of Independent Software Vendors)

Endnotes 2... (after index)

Back Cover