



Newsletter - July, 1999

In This Issue

- New Netscape Service
- Web Scraping in VB
- Ask Jeeves Flies
- Drucker Plugs Outside Info
- Seminars on Web Farming
- Pretexting & Web Surfing
- Search Engine Showdown
- KM Resource Center
- Invisible Web Catalog
- XML Support in Oracle
- More Domains for Infonautics
- U.S. Government Searching

R. Hackathorn
editor

[Comments?](#)

[Subscribe To This Newsletter](#)

[Previous Issues](#)

[PDF Version](#)

[Home Page for WebFarming.com](#)

[The Web Farming Book](#)

[An Introduction](#)

New Netscape Service

Netscape/AOL started a new discovery service with a new twist or two. First, they have partnered with **Google** (described in the last issue). Second, they have integrated the work of the **Open Directory Project** (ODP).



Through the efforts of 13,018 volunteer editors, 705,472 websites have been categorized into 105,733 categories with Yahoo-like descriptions (as of July 5, 1999). Impressive by any standards!

Netscape/AOL is freely licensing the ODP information to **Lycos, HotBot, DopPile, and others**. The ODP information is maintained with the **Resource Description Framework** of **W3C**. With all those volunteers, ODP's motto is "Humans do it better."

Seminars on Web Farming

Three-day seminars on **Web Farming** are offered by **DCI** on the following dates. For full details, see the [online brochure](#).
 - **September 15-17** in San Francisco
 - **November 10-12** in Dallas

Search Engine Showdown

The **Search Engine Showdown** is a new site monitoring global discovery services, much like the **Search Engine Watch**. The author,



Greg Notess, has done an excellent job in tracking the sizes, hits, overlaps, and dead links of the major sites. Note the *Feature Chart*.

Invisible Web Catalog

IntelliSeek has launched their **Invisible Web Catalog**, a by-product of **BullsEye** search tool. **Lycos** has licensed this catalog as a section called Reference > Searchable Databases. Must browse! 🐞

As quoted in a TechWeb [article](#), **Danny Sullivan** of the **Search Engine Watch** said that "there is lots of information that cannot be found via ordinary search engines because it is locked up in databases."

U.S. Government Searching

Web Scraping in VB

Like the screen scraping programs of the 3270-era, web scraping has become quite popular. Read about an example of scraping data from **Amazon.com** using Visual Basic. [more ▶](#)

Ask Jeeves Flies

The IPO of **Ask Jeeves** (NASDAQ: **ASKJ**) soared over 400% to \$72, giving it a market capitalization of almost two billion dollars! For the first quarter, revenues were \$1M with a loss of \$5M. Previously, this low-end search service has not attracted



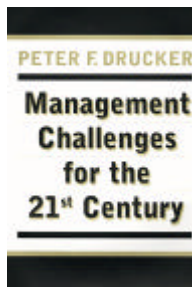
much attention. Its approach is to research recurring questions by its staff and respond to new questions with this research. The company is targeting corporate clients (like Dell and Compaq) for their growth by licensing their technology for in-house knowledge bases.

Pretexting & Web Surfing

Recent articles about pretexting caught my attention. They dealt with an information broker who calls a company to obtain confidential information about another person by giving a false pretext. In other words, pretexting is lying about who you are and why you are obtaining the information. Read more about this critical issue for web farmers. [more ▶](#)

Drucker Plugs Outside Info

In the latest [book](#) by **Peter Drucker** entitled *Management Challenges for the 21st Century*, the following statement makes the case for Web Farming: "And so is the one new area - and the most important one - in which we do not as yet have systematic and



organized methods for obtaining information: information on the **OUTSIDE** of the enterprise. These new methods are very different in their assumptions and their origins. Each was developed independently and by different people

An Introduction to Web Farming

One of the harder resources to search is all of the U.S. Government websites because they are so extensive and so fragmented. A new service by Northern Light alleviates this problem. **USGovsearch** charges fees based on a subscription - \$5 for a day pass, \$30 for a month, and \$250 for the year. This reminds one of the entrance fees to a national park! Remember that most content of the U.S. Government is free - if you can find it. . .



independently and by different people... They aim at providing information rather than data." p. 101

KM Resource Center

The **Knowledge Management Resource Center** has featured the Web Farming website. [more ▶](#)



More Domains for Infonautics

In the last issue, we noted that Infonautics, the provider of the successful Company Sleuth service, secure more domains for its future expansion. They recently added more. Guess what Infonautics will be sleuthing in the future!

- HOCKEYSLEUTH.COM
- BASEBALLSLEUTH.COM
- NASCARSLEUTH.COM
- FOOTBALLSLEUTH.COM
- SPORTSLEUTH.COM

XML Support in Oracle

Oracle announces XML support for its database server. Several XML parsers for C/C++ and Java were included, along with XML utilities and server-side applet to allow Oracle8i to process queries and reply with XML documents. However, Oracle's integration with XML is lacking without data cartridge support, as compared with IBM's XML Extender. [more ▶](#)



Newsletter - July, 1999

Web Scraping with Visual Basic

Back in the 1980's, a hot topic was "screen scraping" -- quick-and-dirty applications that pretended to behave like 3270 or vt100 terminals to legacy systems. These applications were not reliable, but they were cheap. They provided lots of bang-for-the-buck, as compared to doing it the proper way.

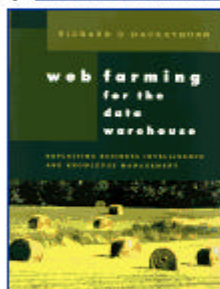
With web farming, we are in a similar situation with web scrapers, applications that process the HTML of a web page to extract meaningful data. Again, web scrapers are not reliable, but they are cheap and useful.

Let's examine a simple application that I have been operating for several months. It is written in Microsoft Visual Basic version 6. The application retrieves a web page from Amazon.com that contains a description of my Web Farming book. Within that description, it extracts the sales rank of the book and writes that value to a file.



Web Farming for the Data Warehouse (The Morgan

by [Richard D. Hackathorn](#)



List Price: \$44.95

Our Price: **\$35.96**

You Save: **\$8.99 (20%)**

Availability: Usually ships within 24 hours.

Paperback - 425 pages (November 1998)

Academic Press/Morgan Kaufmann; ISBN: 1558605037 ; Dimensions (in inches): 1.00

Amazon.com Sales Rank: 4,518

Avg. Customer Review: ★★★★★

Number of Reviews: 5

The calculation of the sales rank is based on Amazon.com sales and is updated regularly. The top 10,000 best sellers are updated each hour to reflect sales over the previous 24 hours. The next 100,000 are updated once a day. The rest of the list is updated monthly, based on various factors. The lower the number, the higher the sales for that particular title. When compared with more than two million books, this ranking is the closest indicator to a stock price for a book!


If you examine the HTML source for this page, you will find the following fragment amid lots of garbage:

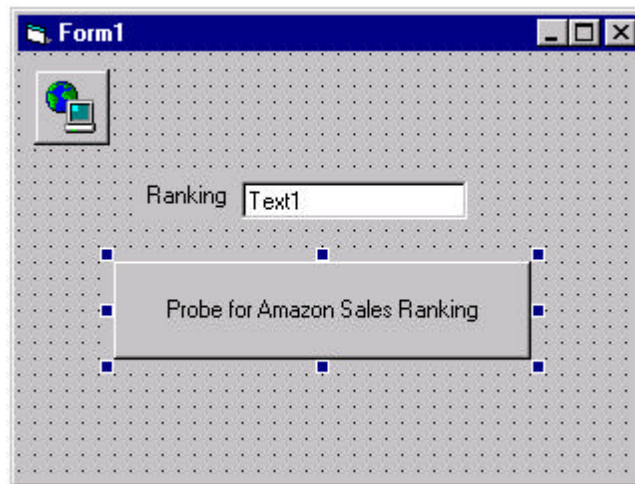
```
<font face=verdana,arial,Helvetica size=-1>
<b>Amazon.com Sales Rank: </b>
4,518
```

```
</font><br>
```

The task is to retrieve the HTML as a text string, search for an initial pre-pattern, search for a post-pattern, extract the text between the two patterns, convert the text to a numeric value, and write it to a file.

Sounds simple? Actually, it is! Here are the steps:

1. Open Visual Studio for VB and create a new standard EXE.
2. Under the Project menu, click on Components to open the Components dialog box. Check the box for Microsoft Internet Transfer Control 6 (which will load MSINET.OCX). Be sure to check the box, and then click OK. This is the secret ingredient!
3. You should see on your tool bar at the left a new icon  (a world with a terminal in front). Click on this icon. Then, drag a square on the form.
4. Also add a label with caption 'Ranking'. Add a textbox and a command button with caption 'Probe for Amazon Sales Ranking'. Your form should look like this. Not pretty, but simple.



5. Now for the fun stuff! Double click on the button to open the code window. For the routine Command1_Click, copy and paste the following. Watch for extra line breaks.

```
Private Sub Command1_Click()
    Dim strPage, strISBN, strURL As String
    On Error Resume Next

    ' set the proper URL to Amazon.Com asking for specific book
    strISBN = "1558605037" ' ISBN for the WFbook
    strURL = "http://www.amazon.com/exec/obidos/ASIN/" _
        & strISBN & "/"
    ' get the webpage content using Inet control
    strPage = Inet1.OpenURL(strURL, icString)
    ' put the ranking value into the textbox
    text1.Text = GetRank(strPage, "Sales Rank: </b>", "</font>")
End Sub
```

All the work is performed in the INET control with the OpenURL method. If there is an error, the string strPage remains empty because of the Resume Next.

6. One more piece of code is required. The GetRank routine must parse all of that messy HTML and extract the sales ranking. Note that the arguments are the HTML buffer, the pattern prior to the ranking, and the pattern after the ranking. Here is the code to copy and paste after the previous routine. Trust me; it works!

```
Private Function GetRank(strPage, strPrePat, strPostPat As String)
```

```
As String
  Dim iStart, iEnd As Integer
  Dim strIn, strOut As String

  GetRank = ""
  iStart = InStr(1, strPage, strPrePat) ' find first pattern
  If iStart <> 0 Then
    iStart = iStart + Len(strPrePat)
    iEnd = InStr(iStart, strPage, strPostPat) ' second
    If iEnd <> 0 Then
      strIn = Mid(strPage, iStart, iEnd - iStart)
      strOut = ""
      For iStart = 1 To Len(strIn) ' strip out control chars
        If Mid(strIn, iStart, 1) < " " Then
          strOut = strOut & " " ' add a blank instead
        Else
          strOut = strOut & Mid(strIn, iStart, 1)
        End If
      Next iStart
      GetRank = Trim(strOut) ' return extracted value
    End If
  End If
End Function
```

7. Now save and run. Be sure that your connection to the Internet is active. After clicking the button, there will be a pause and a number should appear in the text box. Hopefully, the number will be a low one for such an excellent book!

In the version that I use, I added a timer so that every hour the Amazon web page will be probed. Whenever the value changes, it is written to a comma-delimited text file for import into Excel for charting. So far, I have reliably recorded the sale ranking every hour for over four months.

I would like to hear about your experiences and enhancements.

- Richard Hackathorn
dick@webfarming.com



Newsletter - July, 1999

Pretexting and Web Surfing

Last week, articles in **Denver Post**, **New York Times**, and **ComputerWorld** about 'pretexting' caught my attention. They dealt with an information broker who obtained confidential information about other persons by giving a false pretext. In other words, pretexting (or pretext calling) is lying over the telephone about who



you are and why you are obtaining the information. Typical situations involve debt collection, fraud investigation, divorce proceedings, or tabloid articles about banking accounts, credit card histories, salaries, and medical histories from banks, insurance agencies, medical centers, and telephone companies, just to name a few.

It is interesting that pretexting is usually legal (except when impersonating a police officer or government official). No major corporations have been prosecuted for this deceptive practice. In fact, the practice is widely accepted in some industries.

The case in Denver involves the small firm of Touch Tone Information Acquisition. The State of Colorado is one of the few states having a law against impersonating someone to obtain confidential information for commercial gain. The precedence of this case may affect federal and state legislation across the nation.

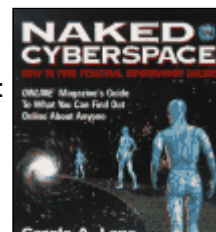
The article in the New York Times entitled "Law Confronts Seller of Private Data" gave additional details. [more](#) (requires a free registration with NYTimes.com)



The ComputerWorld [article](#) summarized the Colorado case and gave advice to companies on how to safeguard their information systems. Some suggestions for 'stopping the leaks' were: to show dialog boxes

to remind agents to verify the caller's identity, to track frequency of confidential inquiries against an account alerting if certain limits are exceeded, and to monitor outsourcing services for any sensitive customer support functions. Clearly state the policy for releasing confidential information to all employees (and contractors). And then, apply lots of common sense.

The issues and techniques for obtaining confidential information about persons are well explained in Carole Lane's book *Naked in Cyberspace*. The Web Farming [book](#) carried the following comment about Lane's book: "A must reference for the Web farmer... A sobering and balanced description of the privacy and need-to-know issues." However, nowhere in this book are suggestions for deceptive practices, like pretext calling. In contrast, Lane clearly calls for a high level of professionalism by information brokers. In particular, it is ethical practice by brokers to



information brokers. In particular, it is ethical practice by brokers to identify honestly the caller and the reasons for the call.



How does pretext calling relate to web farming? The efficient exchange of information for web farming will depend upon honest authentication of both producers and consumers of information. Deception by the parties in this exchange will add tremendous burdens.

It is widely accepted that it is wrong to falsify identity through misusing credit cards for an e-commerce purchase via the Web. However, is it wrong to falsify identity to obtain a white paper from your competitor's website? Is it wrong to not identify your web crawler when you spider your competitor's website? And so on. . .

The point is that web farmers must act in highly ethical manner if we are to create a viable profession from this discipline. See the suggested [Code of Ethics](#) for web farming. Note specifically the point about disclosure.

I would like to hear about your comments on this critical issue.

- *Richard Hackathorn*
dick@webfarming.com