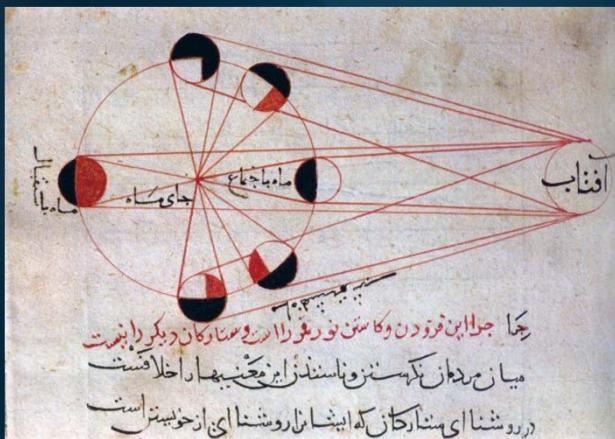


DATA COLLECTION TECHNIQUES

By: Meisam Hejazi Nia

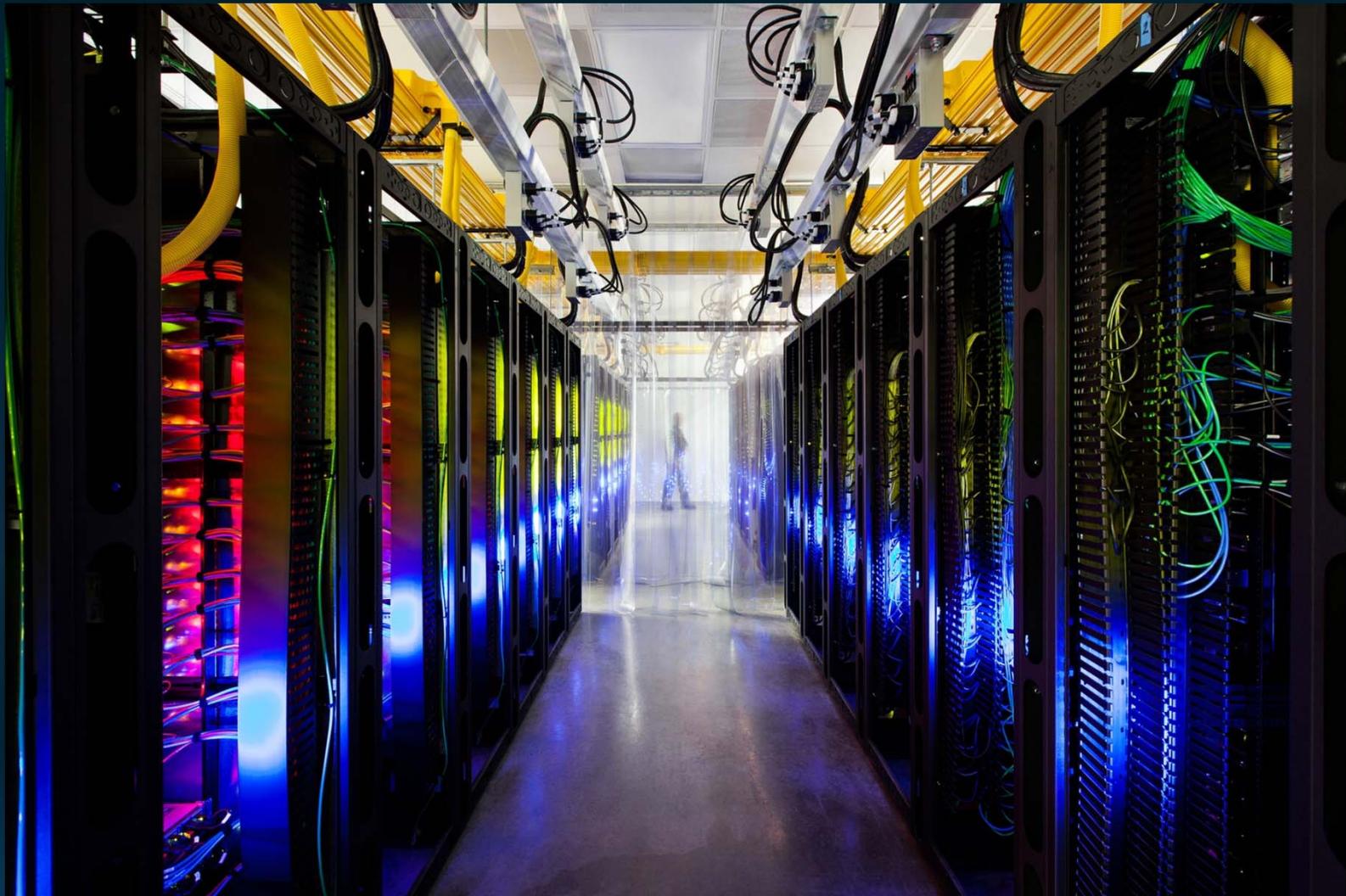
1

DATA COLLECTION HISTORY



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

DATA COLLECTION TODAY



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

STEPS OF DATA COLLECTION

Crawling

Manual
Web harvest
Perl + imacro
Python

Scraping

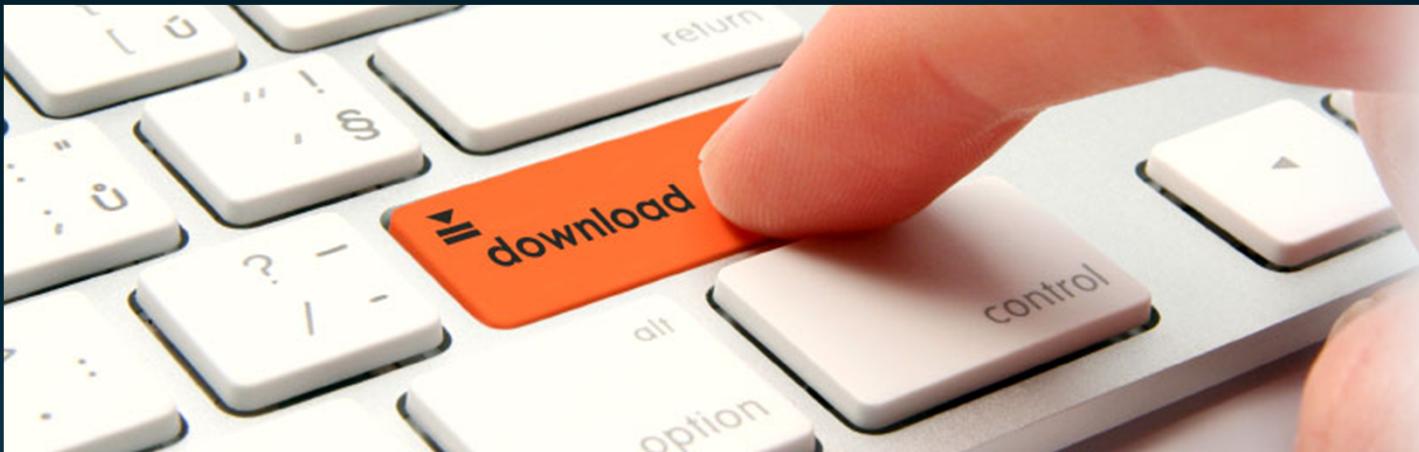
Perl
Python

Cleaning

SQL
Excel
Access

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

MANUAL



- Easier
- Doesn't need lots of coding
- Low Precision
- Lots of errors
- Sometimes very time consuming
- But:
 - Consider it whenever too much coding is required



XPATH: PATH IN MARKUP LANGUAGE

Microsoft SQL Server Management Studio

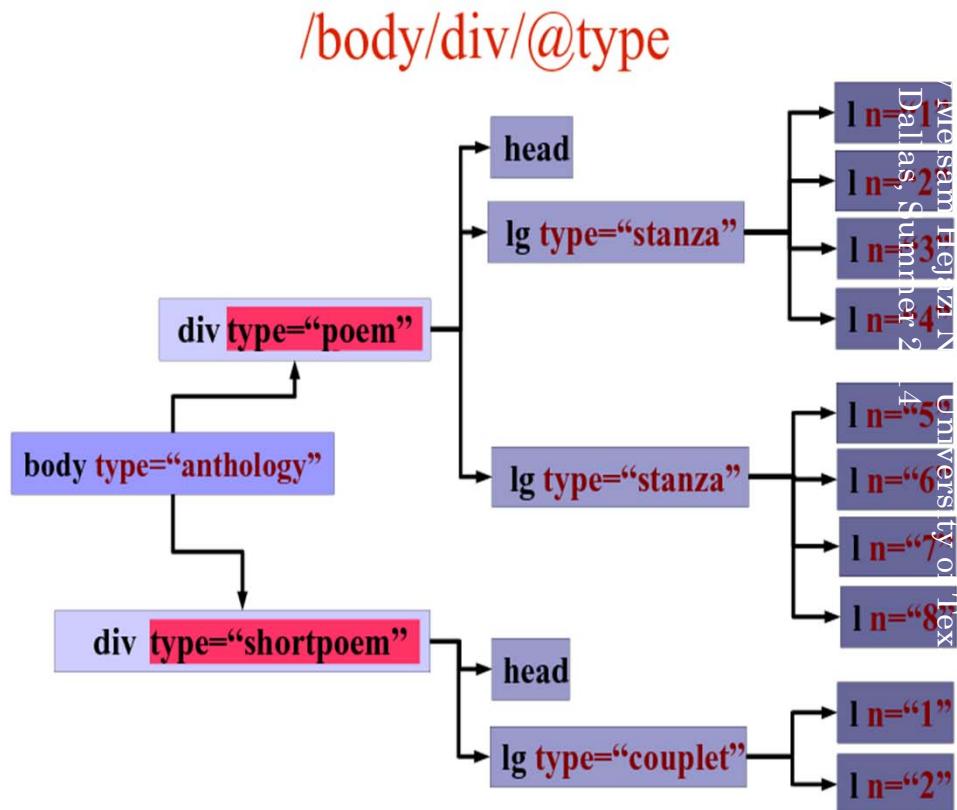
File Edit View Project Tools Window Community Help

New Query Object Explorer Details

Education2.xml WS04AdventureW...Files\Query.sql

```
<ns:Edu> Bachelor <ns:Edu> 1993-09-04Z <ns:Edu> 1997-06-03Z <ns:Edu> Bachelor of Science <ns:Edu> Industrial Engineering <ns:Edu> <ns:Edu> 3.4 <ns:Edu> 4 <ns:Edu> Everglades State College <ns:Edu> <ns:Edu> Location <ns:Location> US <ns:Loc> FL <ns:Loc> Tampa Bay <ns:Loc> </ns:Location> </ns:Edu> Location </ns:Edu> Education <ns:Edu> High School <ns:Edu> 1989-08-27Z <ns:Edu> 1993-06-12Z <ns:Edu> Diploma <ns:Edu> <ns:Edu> 3.8 <ns:Edu> 4 <ns:Edu> Evergreen High School <ns:Edu> <ns:Edu> Location <ns:Location> US <ns:Loc> FL <ns:Loc> Orlando <ns:Loc> </ns:Location> </ns:Edu> Location </ns:Edu> Education
```

Ready Ln1 Col1 Ch1



SELECTING NODE

Expression		Path Expression	Result
<i>nodename</i>	Selects all nodes with the name " <i>nodename</i> "	bookstore	Selects all nodes with the name "bookstore"
/	Selects from the root node	/bookstore	Selects the root element bookstore Note: If the path starts with a slash (/) it always represents an absolute path to an element!
//	Selects nodes in the document from the current node that match the selection no matter where they are	bookstore/book	Selects all book elements that are children of bookstore
.	Selects the current node	//book	Selects all book elements no matter where they are in the document
..	Selects the parent of the current node	bookstore//book	Selects all book elements that are descendant of the bookstore element, no matter where they are under the bookstore element
@	Selects attributes	//@lang	Selects all attributes that are named lang
	Select several path	//book/title //book/price	Select all the title and price elements of all book elements

PREDICATE TO FIND SPECIFIC NODE

Path Expression	Result
/bookstore/book[1]	Selects the first book element that is the child of the bookstore element. Note: In IE 5,6,7,8,9 first node is [0], but according to W3C, it is [1]. To solve this problem in IE, set the SelectionLanguage to XPath: <i>In JavaScript:</i> <code>xml.setProperty("SelectionLanguage", "XPath");</code>
/bookstore/book[last()]	Selects the last book element that is the child of the bookstore element
/bookstore/book[last()-1]	Selects the last but one book element that is the child of the bookstore element
/bookstore/book[position()<3]	Selects the first two book elements that are children of the bookstore element
//title[@lang]	Selects all the title elements that have an attribute named lang
//title[@lang='en']	Selects all the title elements that have an attribute named lang with a value of 'en'
/bookstore/book[price>35.00]	Selects all the book elements of the bookstore element that have a price element with a value greater than 35.00
/bookstore/book[price>35.00]/title	Selects all the title elements of the book elements of the bookstore element that have a price element with a value greater than 35.00

XPATH AXES

AxisName	Result	Example	Result
ancestor	Selects all ancestors (parent, grandparent, etc.) of the current node	child::book	Selects all book nodes that are children of the current node
ancestor-or-self	Selects all ancestors (parent, grandparent, etc.) of the current node and the current node itself	attribute::lang	Selects the lang attribute of the current node
attribute	Selects all attributes of the current node	child::*	Selects all element children of the current node
child	Selects all children of the current node	attribute::*	Selects all attributes of the current node
descendant	Selects all descendants (children, grandchildren, etc.) of the current node	child::text()	Selects all text node children of the current node
descendant-or-self	Selects all descendants (children, grandchildren, etc.) of the current node and the current node itself	child::node()	Selects all children of the current node
following	Selects everything in the document after the closing tag of the current node	descendant::book	Selects all book descendants of the current node

XPATH AXES CONT.

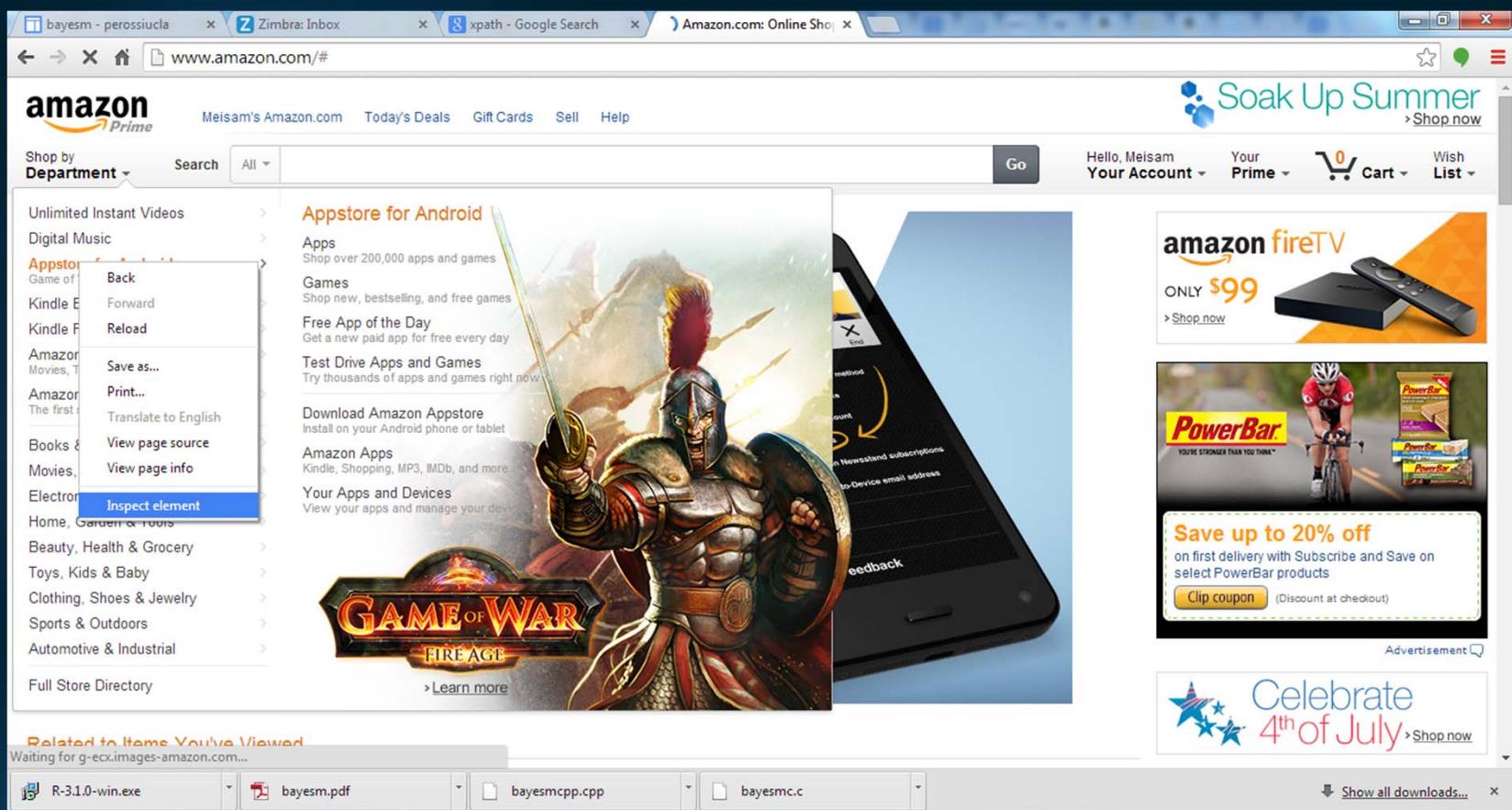
AxisName	Result	Example	Result
following-sibling	Selects all siblings after the current node	ancestor::book	Selects all book ancestors of the current node
namespace	Selects all namespace nodes of the current node	ancestor-or-self::book	Selects all book ancestors of the current node - and the current as well if it is a book node
parent	Selects the parent of the current node	child::* / child::price	Selects all price grandchildren of the current node
preceding	Selects all nodes that appear before the current node in the document, except ancestors, attribute nodes and namespace nodes	ancestor::book	Selects all book ancestors of the current node
preceding-sibling	Selects all siblings before the current node	ancestor-or-self::book	Selects all book ancestors of the current node - and the current as well if it is a book node
self	Selects the current node	child::* / child::price	Selects all price grandchildren of the current node

XPATH OPERATORS

Operator	Description	Example
	Computes two node-sets	//book //cd
+	Addition	6 + 4
-	Subtraction	6 - 4
*	Multiplication	6 * 4
div	Division	8 div 4
=	Equal	price=9.80
!=	Not equal	price!=9.80
<	Less than	price<9.80
<=	Less than or equal to	price<=9.80
>	Greater than	price>9.80
>=	Greater than or equal to	price>=9.80
or	or	price=9.80 or price=9.70
and	and	price>9.00 and price<9.90
mod	Modulus (division remainder)	5 mod 2

EASIER WAY TO GET XQUERY?

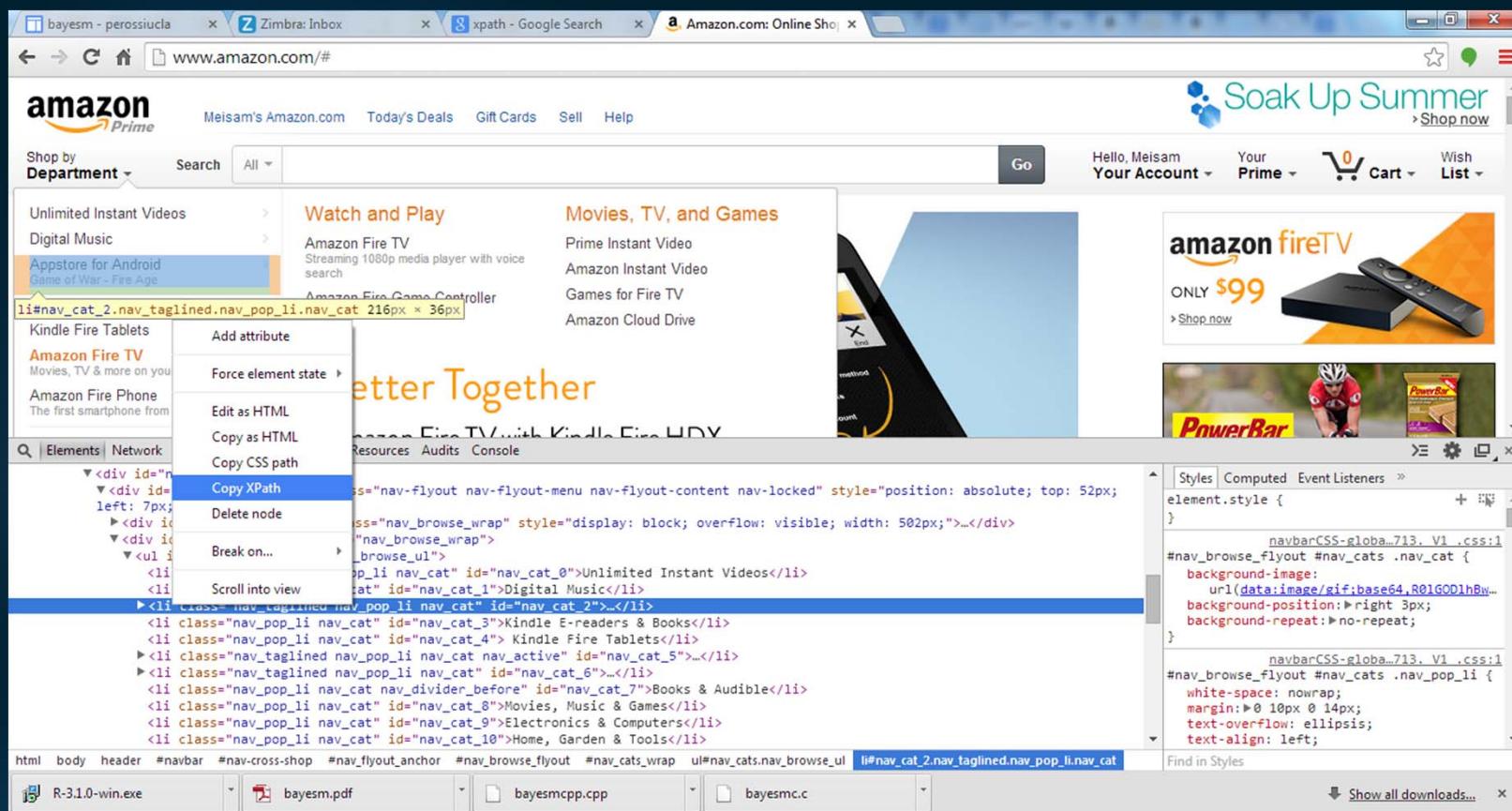
- Right click on part of the screen you want to know the xpath of in Google Chrome and select inspect the element



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

EASIER WAY TO GET XQUERY?

- Right click on the colored element and select copy XPATH



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

OPEN SOURCE OPTION: WEB HARVEST

- Download from: <http://web-harvest.sourceforge.net/>

The screenshot shows the Web-Harvest configuration editor window. The menu bar includes Config, Edit, View, Execution, Help, and a toolbar with various icons. The status bar at the bottom right shows "64M of 121M". On the left is a tree view of the configuration structure:

- include
- x= var-def
- call
 - call-param
 - call-param
 - call-param
 - call-param
- file
 - loop
 - list
 - x var
 - body
 - xquery
 - xq-param
 - x var
 - xq-expression

The main pane displays the XML configuration code:

```
<?xml version="1.0" encoding="UTF-8"?>
<config charset="utf-8">

    <include path="C:/crawler/functions.xml"/>

    <!-- collects all tables for individual products -->
    <var-def name="reviews">
        <call name="download-multipage-list">
            <call-param name="pageUrl">https://addons.mozilla.org/en-US/firefox/addon/新同文堂-new-tong-wen-tang/reviews/</call-param>
            <call-param name="nextXPath">//a[starts-with(., 'Next')]/@href</call-param>
            <call-param name="itemXPath">//div[@class="review c item"]</call-param>
            <call-param name="maxloops">4</call-param>
        </call>
    </var-def>

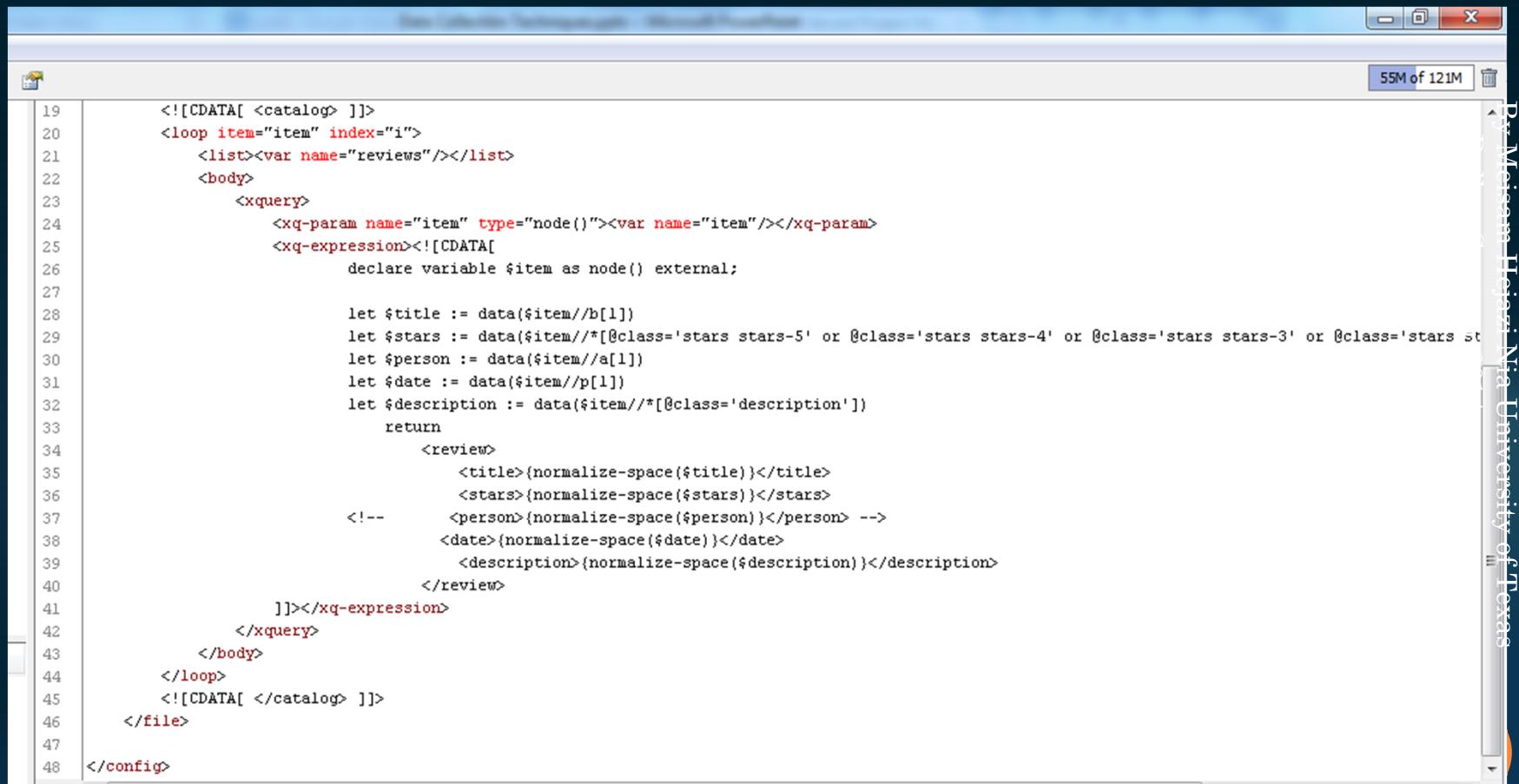
    <!-- iterates over all collected products and extract desired data -->
    <file action="write" path="output/review.xml" charset="UTF-8">
        <![CDATA[ <catalog> ]]>
        <loop item="item" index="i">
            <list><var name="reviews"/></list>
            <body>
                <xquery>
                    <xq-param name="item" type="node()"><var name="item"/></xq-param>
                    <xq-expression><![CDATA[
                        declare variable $item as node() external;

                        let $title := data($item//b[1])
                        let $stars := data($item//*[@class='stars stars-5' or @class='stars stars-4' or @class='stars stars-3' or @class='stars stars-2'])
                        let $person := data($item//a[1])
                    ]]>

At the bottom left, there is a table titled "Name" and "Value" with one row. The status bar at the bottom left shows "Welcome reviewCrawlerDownloadHelper.xml".


```

OPEN SOURCE OPTION: WEB HARVEST



The screenshot shows a Windows desktop environment with a code editor window open. The window title is 'Web Harvest Configuration File'. The status bar at the bottom right indicates '55M of 121M'. The code editor displays XQuery code for processing catalog items. The code includes declarations for variables like \$item, \$title, \$stars, \$person, \$date, and \$description. It uses XQuery functions like data() and normalize-space(). The code is structured with XML-like tags such as <catalog>, <loop>, <list>, <body>, <xquery>, <xq-param>, and <xq-expression>. The code editor has a dark theme with syntax highlighting for different tags and variables.

```
<![CDATA[ <catalog> ]]>
<loop item="item" index="i">
    <list><var name="reviews"/></list>
    <body>
        <xquery>
            <xq-param name="item" type="node()"><var name="item"/></xq-param>
            <xq-expression><![CDATA[
                declare variable $item as node() external;

                let $title := data($item//b[1])
                let $stars := data($item//*[@class='stars stars-5' or @class='stars stars-4' or @class='stars stars-3' or @class='stars stars-2' or @class='stars stars-1'])
                let $person := data($item//a[1])
                let $date := data($item//p[1])
                let $description := data($item//*[@class='description'])
                    return
                        <review>
                            <title>(normalize-space($title))</title>
                            <stars>(normalize-space($stars))</stars>
                            <!--
                                <person>(normalize-space($person))</person> -->
                            <date>(normalize-space($date))</date>
                            <description>(normalize-space($description))</description>
                        </review>
                ]]></xq-expression>
        </xquery>
    </body>
</loop>
<![CDATA[ </catalog> ]]>
</file>
</config>
```

IMACRO RECORDING

The screenshot shows a Mozilla Firefox browser window displaying the Mozilla Add-ons website. The URL in the address bar is <https://addons.mozilla.org/en-US/firefox/addon/imacros-for-firefox/>. The page title is "iMacros for Firefox :: Add-ons ...". The main content area shows the "ADD-ONS" section for Firefox. A large callout box highlights the "Welcome to Firefox Add-ons. Choose from thousands of extra features and styles to make Firefox your own." message. Below this, the "iMacros for Firefox" add-on is listed with a rating of 5 stars, 342 user reviews, and 290,531 users. A green "Add to Firefox" button is visible. At the bottom of the page, there is a note about Firefox sending data to Mozilla and a "Choose What I Share" link.

iMacros for Firefox :: Add-ons ...

Mozilla Foundation (US) | <https://addons.mozilla.org/en-US/firefox/addon/imacros-for-firefox/>

Customize Links | Free Hotmail | Windows Marketplace | Windows Media | Windows | Getting Started | LyX news feed | Web Slice Gallery | Suggested Sites | Most Visited | Getting Started

Register or Log in | Other Applications | mozilla ▾

ADD-ONS

EXTENSIONS | THEMES | COLLECTIONS | MORE...

search for add-ons

Welcome to Firefox Add-ons. Choose from thousands of extra features and styles to make Firefox your own.

[+ Add to Firefox](#)

iMacros for Firefox 8.8.2
by iOpus

Automate Firefox. Record and replay repetitious work. If you love the Firefox web browser, but are tired of repetitive tasks like visiting the same sites every days, filling out forms, and remembering passwords, then iMacros for Firefox is the solution you've been dreaming of! ***Whatever you do with Firefox, iMacros can automate it.***

★★★★★
342 user reviews
290,531 users

Add to collection | Share this Add-on

Meet the Developer: [iOpus](#)
Learn why iMacros for Firefox was created and find out what's next for this add-on.

Firefox automatically sends some data to Mozilla so that we can improve your experience.

Choose What I Share

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

IMACRO RECORDING

The screenshot shows the iMacros for Firefox extension interface. The browser title bar reads "iMacros for Firefox :: Add...". The main content area displays the iMacros.net website, which features a logo and navigation links for "Features", "Purchase", "About", "Support", "Downloads", and "Contact". On the left, a sidebar titled "iMacros" contains a tree view with "Favorites", "Demo", and "Demo-Firefox". Below this is a control panel with buttons for "Play", "Rec", and "Manage", where "Play" is highlighted. It also includes a "Repeat Macro" section with "Current: 1" and "Max: 3", and a "Play (Loop)" button. At the bottom, there's an "About iMacros" link and a status bar indicating "Waiting for imacros.net". A vertical watermark on the right side of the page reads "isam Hejazi Nia University of Texas at Dallas, Summer 2014".

RECORDING

The screenshot shows a Firefox browser window with the iMacros add-on active. The title bar indicates the user is testing how iMacro works on Google. The iMacros sidebar is open, showing the 'Recording' tab is selected. The sidebar displays the following information:

- VERSION BUILD=8820413 RECORDER...
- TAB T=1
- URL GOTO=https://www.google.co...
- TAG POS=1 TYPE=INPUT:TEXT FOR...

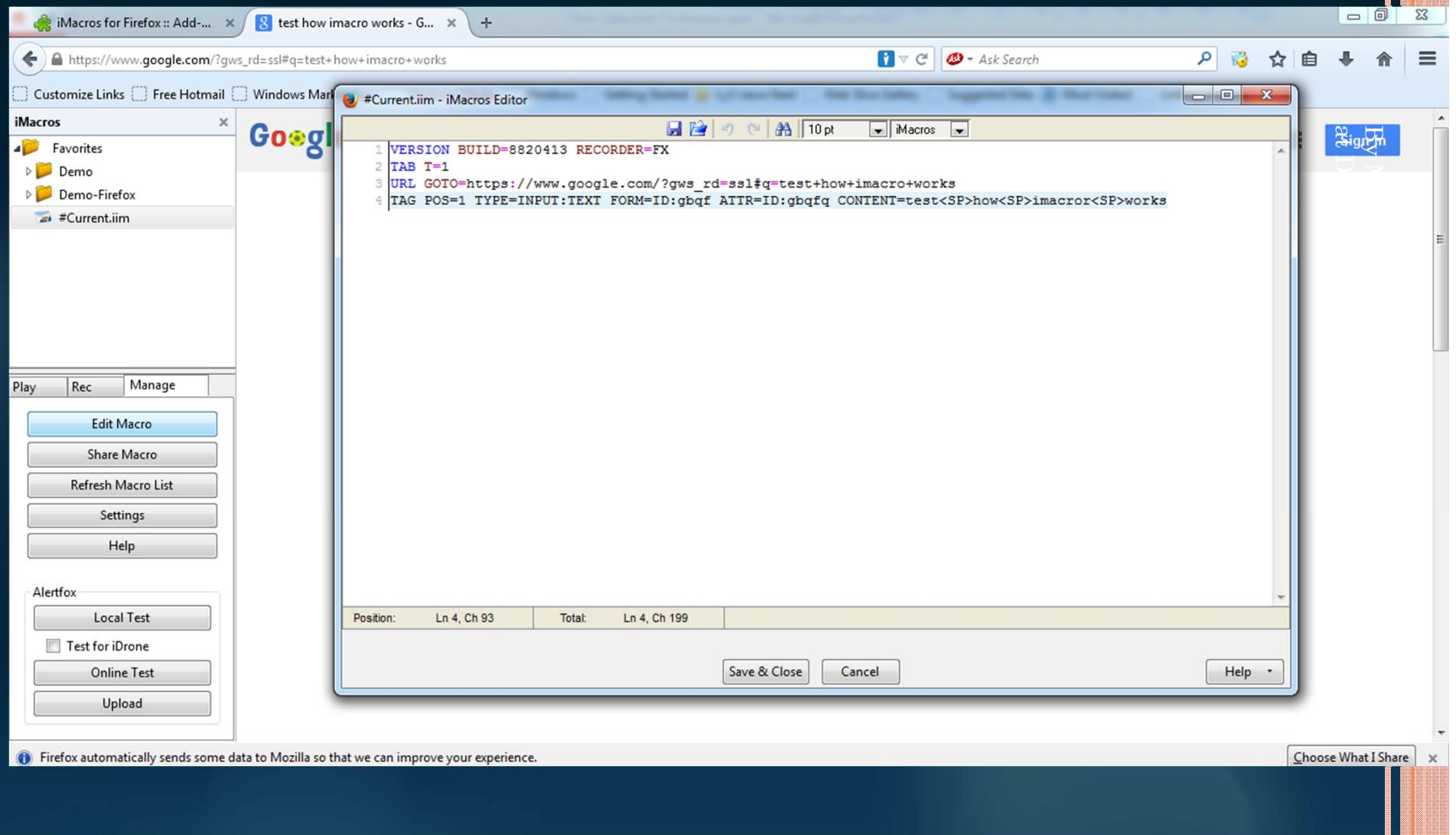
The main content area shows a Google search results page for "test how imacrom works". The search bar contains the query. Below the search bar, there is a message: "Press Enter to search.". On the left side of the main window, there is a toolbar with three buttons: Play, Rec (which is highlighted with an orange border), and Manage.

Below the toolbar, there is a list of options:

- Record
- Save
- Stop
- Record options
- Save Page As
- Take Screenshot
- Del. Cache&Cookies
- Wait during Play

At the bottom of the window, there is a status bar with the URL "www.microsoft.com/isapi/redir.dll?prd=ie&pver=6&ar=CLinks" and a note from Firefox: "Firefox automatically sends some data to Mozilla so that we can improve your experience." There is also a "Choose What I Share" button.

VIEW THE CODE OF RECORDED MACRO



IMACRO USEFUL COMMANDS

Command	Description
Wait seconds=x	Wait for x seconds
URL GOTO= javascript:window.scrollBy(x,y)	Scroll the page by x and y
SET !DATASOURCE address.csv	To save extracted item into address.csv
SET !DATASOURCE_LINE {{!Loop}}	Set the current line of the file to {{!Loop}}
{{!col1}}...{{!col5}}	Column of the file to write in
SET !EXTRACT NULL	To empty extract variable
SET !VAR2 {{!EXTRACT}}	To set the content of variable VAR2 to content that we have extracted
POS=R1	Define which relative position to use
“\n\t\\”	Next line, tab or \ variable
SET ERRORIGNORE YES	Ask to not through error
#EANF#	When imacro doesn't find extract item

IMACRO USEFUL COMMANDS

Command	Description
FIITER TYPE=IMAGE STATUS=OFF	To not load image when crawling
FRAME F=2	Select second frame in the html file
SAVEAS TYP:CPL Folder=* FILE=x MHT HTM TXT	Different format of file to save the page with name x
{!URLCURRENT{}}	The url of current page variable
SAVEAS TYP=EXTRACT FOLDER=* FILE=Mytable_{!Now:ymmd_hhnns}s}.csv	To save content that is extracted into file
\$b→iimSET("myvar",s);	To set variable myvar in script
\$b→iimGETLASTErrorMessage()	To get last error message
TAG XPATH="..."	To click on the tag with the xpath specified
\$b→iimGetLastExtract(i)	Get extracted item
\$b→iimplay("code:URL GOTO=...",timeout)	To run got to specific URL in a script with timeout

SAMPLE CUSTOMIZE MODULE OF IMACRO

```
'=====
' Meisam Hjazi Nia
' Manga Project Crawler Module to save rank page
'=====

SET !ERRORIGNORE YES
SET savepath {{currworkDirect}}\DataOfMangaPrj\{{!NOW:ddmmyy}}\{{realizationPath}}\
' no image to both reduce size and speed up
FILTER Type=IMAGES STATUS=ON
URL
GOTO=http://www.amazon.com/s/ref=sr_pg_2?rh=n%3A165793011%2Cn%3A165993011%2Cn%3A2514571011%2Ck%3Aanime+figures&
page=1&keywords=anime+figures&ie=UTF8&qid=1400719258&lo=toys-and-games
WAIT SECONDS=.5
DS CMD=MOVETO X=450 Y=410
DS CMD=MOVETO X=100 Y=410
DS CMD=MOVETO X=450 Y=100
DS CMD=MOVETO X=100 Y=100
WAIT SECONDS=3
'save the main page
SAVEAS TYPE=CPL FOLDER={{savepath}} FILE={{realizationPath}}MainRank_P{{pagenum}}_{{!NOW:ddmmyy_hhnnss}}.htm
WAIT SECONDS=1
TAG POS=1 TYPE=H2 ATTR=ID:s-result-count EXTRACT=TXT
'=====
' Termination is inside main page
'=====
```

By Meisam Hjazi Nia University of Texas
at Dallas, Summer 2014

SAMPLE CUSTOMIZE MODULE OF IMACRO

```
'=====
' Meisam Hjazi Nia
' Manga Project Crawler Module to search each manga name
'=====

SET !ERRORIGNORE YES
SET savepath {{currworkDirect}}\DataOfMangaPrj\{{!NOW:ddmmyy}}\{{realizationPath}}\
' no image to both reduce size and speed up
FILTER Type=IMAGES STATUS=ON

' search Manga's anime

URL
GOTO=http://www.amazon.com/s/ref=sr_pg_2?rh=n%3A165793011%2Cn%3A165993011%2Cn%3A2514571011%2Cn%3A{{mangaName}}&page={{nextPageNum}}

WAIT SECONDS=1

SAVEAS TYPE=CPL FOLDER={{savepath}}
FILE={{mangaNumber}}.{{mangaName}}_SearchedPage_P{{nextPageNum}}_{{!NOW:ddmmyy_hhnss}}.htm
WAIT SECONDS=1
SET !EXTRACT {{!URLCURRENT}}
'=====
' Termination is inside main page
'=====
```

PERL AND ITS COMPILER AND IDE

- Script language
- Interpreted and not compiled
- Quick to learn and write the code
- Install interpreter: <http://strawberryperl.com/>
- Install IDE: <http://padre.perlide.org/>



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

PERL-INPUT AND OUTPUT

```
$inputfilename="C:/Users/..../XXX.htm";
open( FILE2, $inputfilename );
open( OUT, '>C:/..../CleanedData/output.txt' );
while( $line = <FILE2> ) {
    # print $line;
    #$line=substr($line, 0, -1) ;
    $line =~ s/^[\s]+ | [\s]+$/g ;
    $fullData = join ' ', $fullData, ' ', $line;
    # print "Done...!\n";
}

print OUT "$firstvariable\t$myvariable\n";
print "XX ID is: $myvariable\n";
```

PERL CONTROL SEQUENCE

```
if ($fullData=~<DIV[^>*><A  
href=.http...subsite.website.com[^=]*=[^=]*=([0-9]*)[\s\S]*/>)  
    $VariableID = $1;  
    print "variable ID is: $ VariableID \n";  
}  
  
$fullData = <FILE2>;  
$fullData =~ s/^[\s]+ | [\s]+$/g ; # replace end of the line  
# for more commands check the file :  
SASPERLLATEXSummary
```

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

PERL USAGE OF IMACRO

```
$b = Win32::OLE->new('imacros') or die "iMacros Browser could not be started by  
Win32::OLE\n";
```

```
$b->{Visible} = 1;
```

```
$macro = "$cwd\\Modules\\AnimeScripts\\FirstSearchPagAnimeeSave.iim";
```

```
#Start the iMacros Browser - Use iimInit("-ie"/"-fx") to start iMacros for IE/Firefox  
instead.
```

```
$b->iimInit();
```

```
$animepageNum = 1;
```

```
$b->iimSet('currworkDirect', $cwd );
```

```
$b->iimSet('realizationPath', $realizationPath );
```

```
$b->iimSet('pagenum', $animepageNum);
```

```
$b->iimPlay($macro);
```

```
#$currentURL = $b ->iimGetLastExtract();
```

```
my $totalnumberOfCategoryPagesTxt = $b ->iimGetLastExtract();
```

```
my $totalnumberofitemsOfCategory=0;
```

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

PERL USAGE OF IMACRO CONT.

```
if ($totalnumberOfCategoryPagesTxt =~/([0-9,]+) results/){  
    my $tempstring = $1;  
    $tempstring =~ s/,//g;  
    #print "tempstring conteint is $tempstring\n";  
    $totalnumberofitemsOfCategory = $tempstring;  
    #print "from the text the extracted value is:  $numberpageresults";  
}  
  
my $totalNumCategPages = ceil($totalnumberofitemsOfCategory/60);  
print "Total number of Anime to be saved Extraction ($totalNumCategPages  
Pages).....Done\n";  
  
$b->iimExit();
```

HOW TO BE A GOOD CRAWLER?

Time

- Be careful about the time and put enough wait so that you would be like a human and not machine , otherwise they will block you

Various path and tricks

- There is no single way, think about it from many aspects, you have different tools so use them anytime necessary

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

REGULAR EXPRESSION – META CHARACTER

char	meaning
^	beginning of string
\$	end of string
.	any character except newline
*	match 0 or more times
+	match 1 or more times
?	match 0 or 1 times; <i>or</i> : shortest match
	alternative
()	grouping; “storing”
[]	set of characters
{ }	repetition modifier
\	quote or special

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

REGULAR EXPRESSION – REPETITION

a^*	zero or more a 's
a^+	one or more a 's
$a^?$	zero or one a 's (i.e., optional a)
$a\{m\}$	exactly m a 's
$a\{m,n\}$	at least m a 's
<i>repetition?</i>	same as <i>repetition</i> but the <i>shortest</i> match is taken
\t	tab
\n	newline
\r	return (CR)
\x <h></h>	character with hex. code hh
\b	“word” boundary
\B	not a “word” boundary

REGULAR EXPRESSION – STRINGS

\w	matches any <i>single</i> character classified as a “word” character (alphanumeric or “_”)
\W	matches any non-“word” character
\s	matches any whitespace character (space, tab, newline)
\S	matches any non-whitespace character
\d	matches any digit character, equiv. to [0-9]
\D	matches any non-digit character
<i>characters</i>	matches any of the characters in the sequence
[<i>x</i> - <i>y</i>]	matches any of the characters from <i>x</i> to <i>y</i> (inclusively) in the ASCII code
[\-]	matches the hyphen character “-”
[\n]	matches the newline; other <u>single character denotations with \</u> apply normally, too
[^ <i>something</i>]	matches any character <i>except</i> those that [<i>something</i>] denotes; that is, immediately after the leading “[”, the circumflex “^” means “not” applied to all of the rest
Examples	Check exampleOfRegular Expression file and SASPER....pdf file

REGULAR EXPRESSION IN PERL, HOSPITAL EXAMPLE

```
open( OUT, '>Output/Latitudeoutput.txt' );
for ($k=1;$k<=4;$k++){
    $inputfilename ="profile.php_00";
    $inputfilename=$inputfilename.$k;
    $inputfilename=$inputfilename.".html";
    open( FILE2, $inputfilename );
    open( FILE2, $inputfilename );
    $i=0;
    while( $line = <FILE2> ) {
        if ($line=~</td align="left">([0-9]*).....(E)...([0-9]*). ....(N)<\td>/){
            $i++;
            if (True){
                print "$3$4\n";
                print OUT "$3$4\n";
            }
        }
    }
}
```

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

HOW TO BE A GOOD SCRAPER?

Unique pattern

- Search for unique pattern that does not match in the file with anything except what you want

User General case

- Use . and .* as much as possible to match those items that does not make your pattern unique

Multiple ways

- There are always multiple ways, if one did not work give up and change your pattern, think!

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

DATA CLEANING IN EXCEL-STRING METHODS

Function	Description
<u>CLEAN</u>	Removes all nonprintable characters from text
<u>EXACT</u>	Checks to see if two text values are identical
<u>FIND, FINDB</u>	Finds one text value within another (case-sensitive)
<u>LEFT, LEFTB</u>	Returns the leftmost characters from a text value
<u>LEN, LENB</u>	Returns the number of characters in a text string
<u>LOWER</u>	Converts text to lowercase
<u>MID, MIDB</u>	Returns a specific number of characters from a text string starting at the position you specify
<u>REPLACE, REPLACEB</u>	Replaces characters within text
<u>REPT</u>	Repeats text a given number of times

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

DATA CLEANING IN EXCEL-STRING METHODS

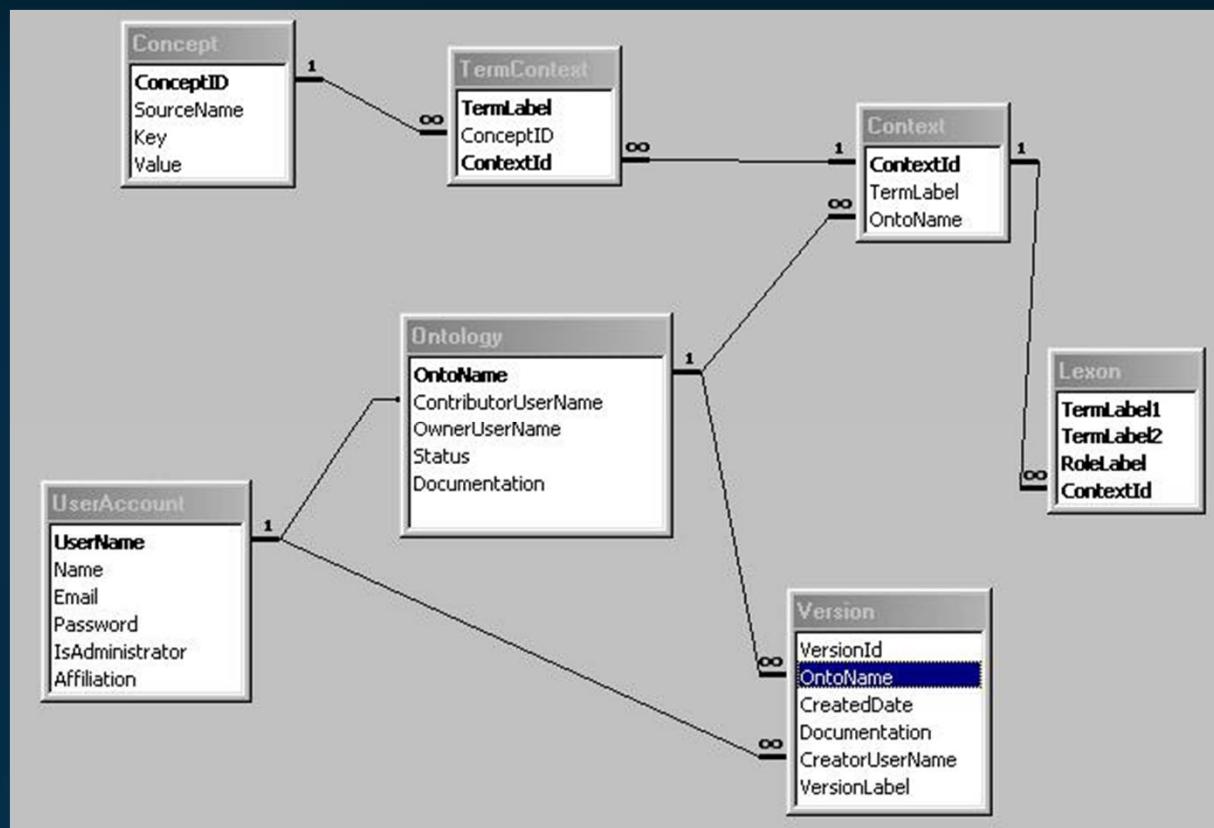
Function	Description
<u>RIGHT, RIGHTB</u>	Returns the rightmost characters from a text value
<u>SEARCH, SEARCHB</u>	Finds one text value within another (not case-sensitive)
<u>SUBSTITUTE</u>	Substitutes new text for old text in a text string
<u>TRIM</u>	Removes spaces from text
<u>UPPER</u>	Converts text to uppercase
<u>VALUE</u>	Converts a text argument to a number

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

RELATIONAL DATABASES

- Data is represented in the form of tables, and the model has 3 components
 - Data structure:
 - data are organised in the form of tables with rows and columns
 - Data manipulation:
 - powerful operations (using the SQL language) are used to manipulate data stored in the relations
 - Data integrity:
 - facilities are included to specify business rules that maintain the integrity of data when they are manipulated

EXAMPLE



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

SQL

- SQL: Structured Query Language
- Attempts to allow humans to ask questions of data sources using natural language
- retrieves and updates data in tables and views based on those tables
- The SAS System's SQL procedure enables you to
 - retrieve and manipulate data that are stored in tables or views.
 - create tables, views, and indexes on columns in tables.
 - create SAS macro variables that contain values from rows in a query's result.
 - add or modify the data values in a table's columns or insert and delete rows. You can also modify the table itself by adding, modifying, or dropping columns.
 - send DBMS-specific SQL statements to a database management system (DBMS) and to retrieve DBMS data.

FORMAT OF SQL COMMAND

- o **SELECT**

- **FROM**

- o **WHERE**

- o **GROUP BY**

- o **HAVING**

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

SQL SAMPLE QUERIES

```
SELECT DailyPurchaseWithCategory.*,
UsersRegion.AreaID INTO
DailyPurchaseWithLocation
FROM DailyPurchaseWithCategory, UsersRegion
WHERE
(((UsersRegion.UserID)=[DailyPurchaseWithCategory].[UserID]));
```

By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014

EXAMPLE: SIMPLE SELECTION, CREATE TABLE

```
SELECT
Category,count([3APPPU~1].DATETIME) as
DownloadCounts
FROM [3APPPU~1],[4APPCH~1]
WHERE
[3APPPU~1].PRODUCTID=[4APPCH~1].APPI
D and (Category ="Action/Strategy" )
GROUP BY DownloadDate,Category
ORDER BY DownloadDate
```

ACCESS AND SQL QUERIES

Import your excel file into access

Run different types of query

Sometimes it is good to first create another table and then run the query on the new table

Be patient and enjoy programming !



By Meisam Hejazi Nia University of Texas
at Dallas, Summer 2014