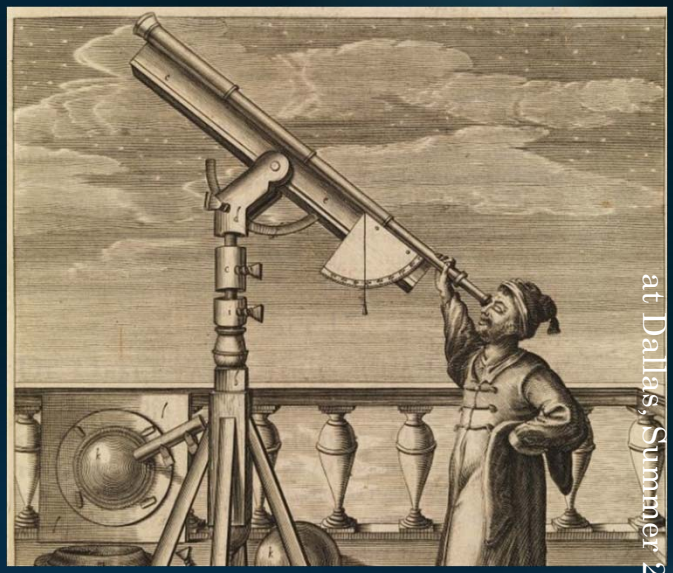
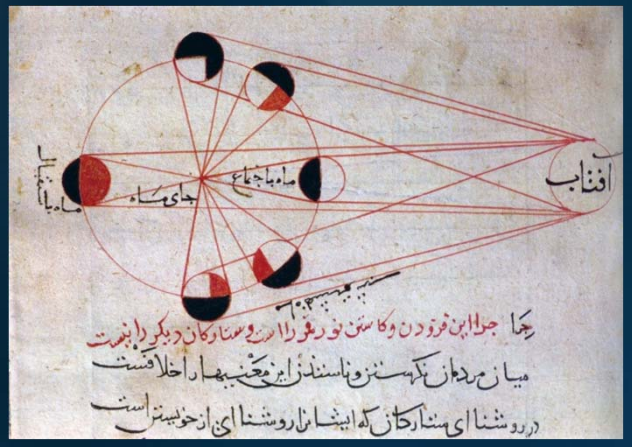
Data Collection Techniques

By: Meisam Hejazi Nia

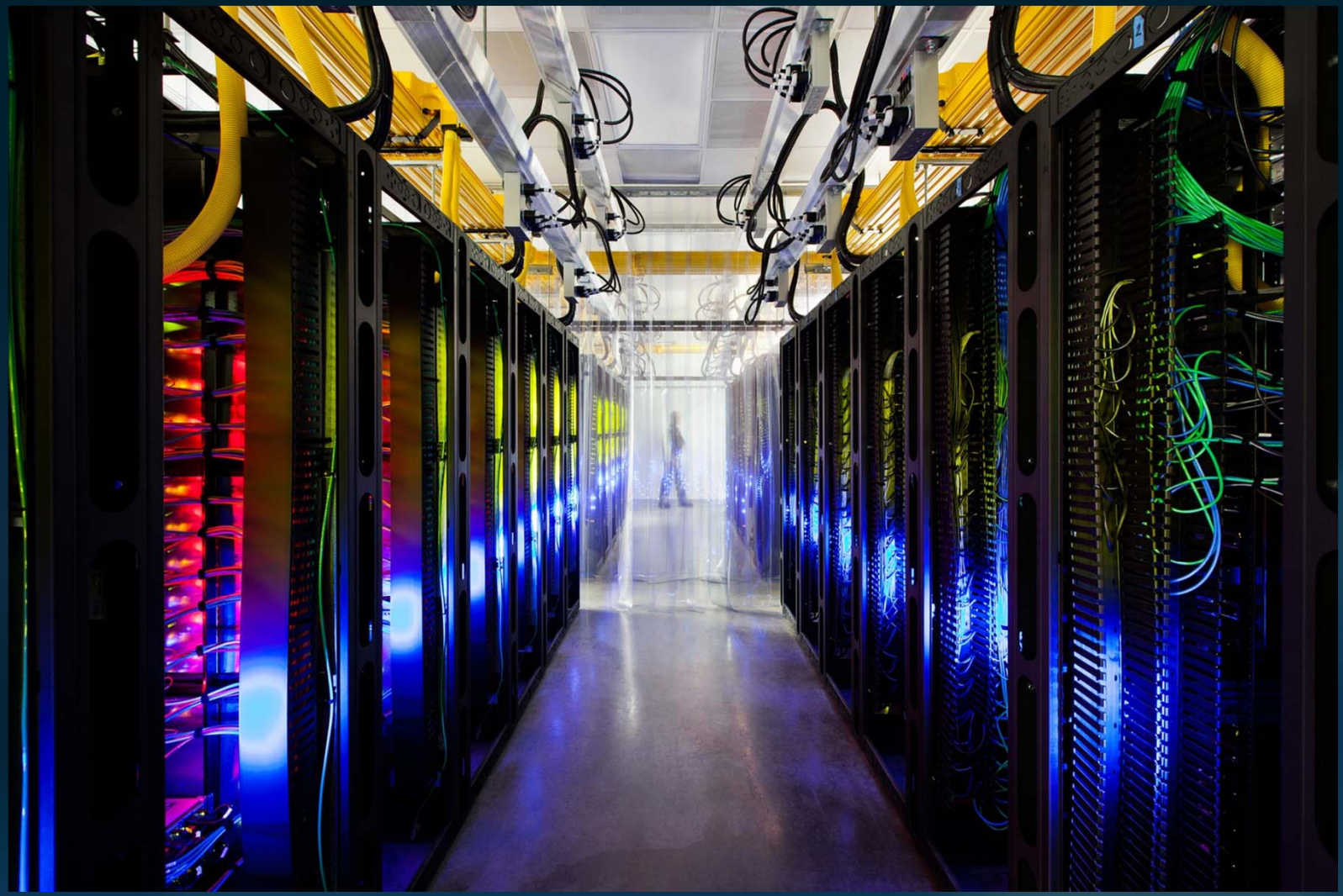
Data collection history



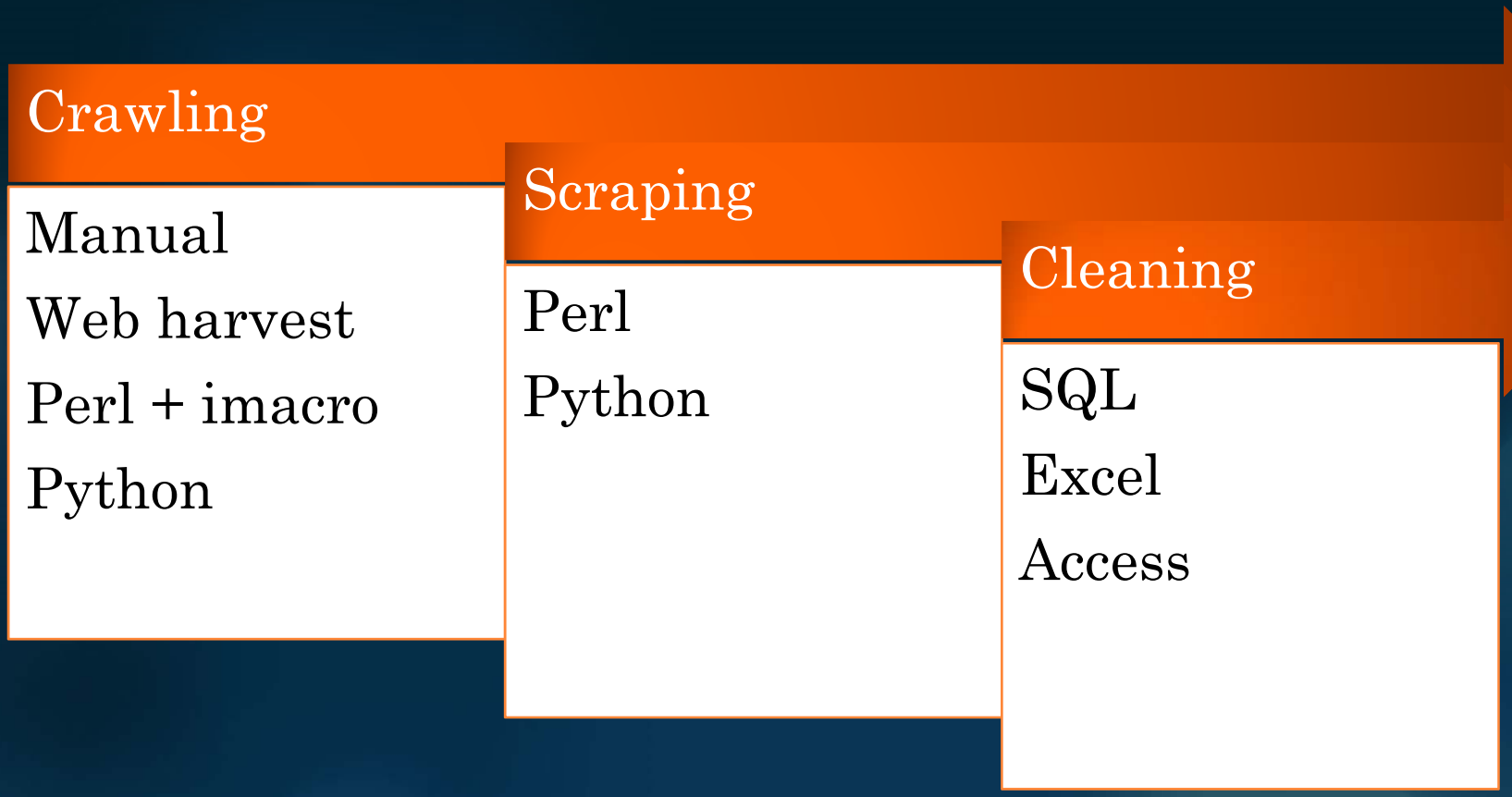


By Meisam Hejazi Nia University of Texas at Dallas, Summer 2014

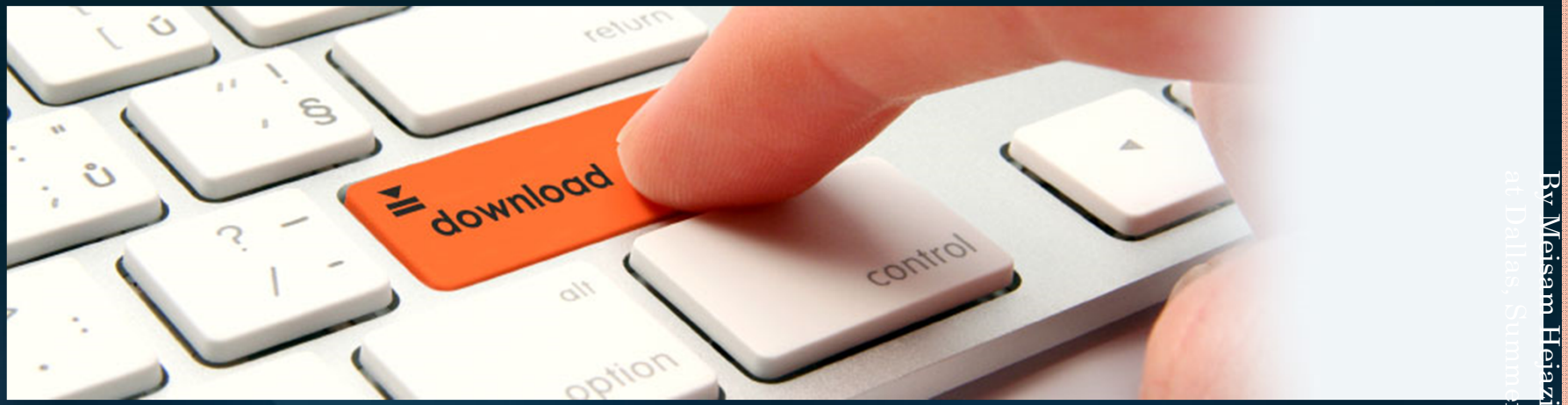
Data Collection Today



Steps of Data collection

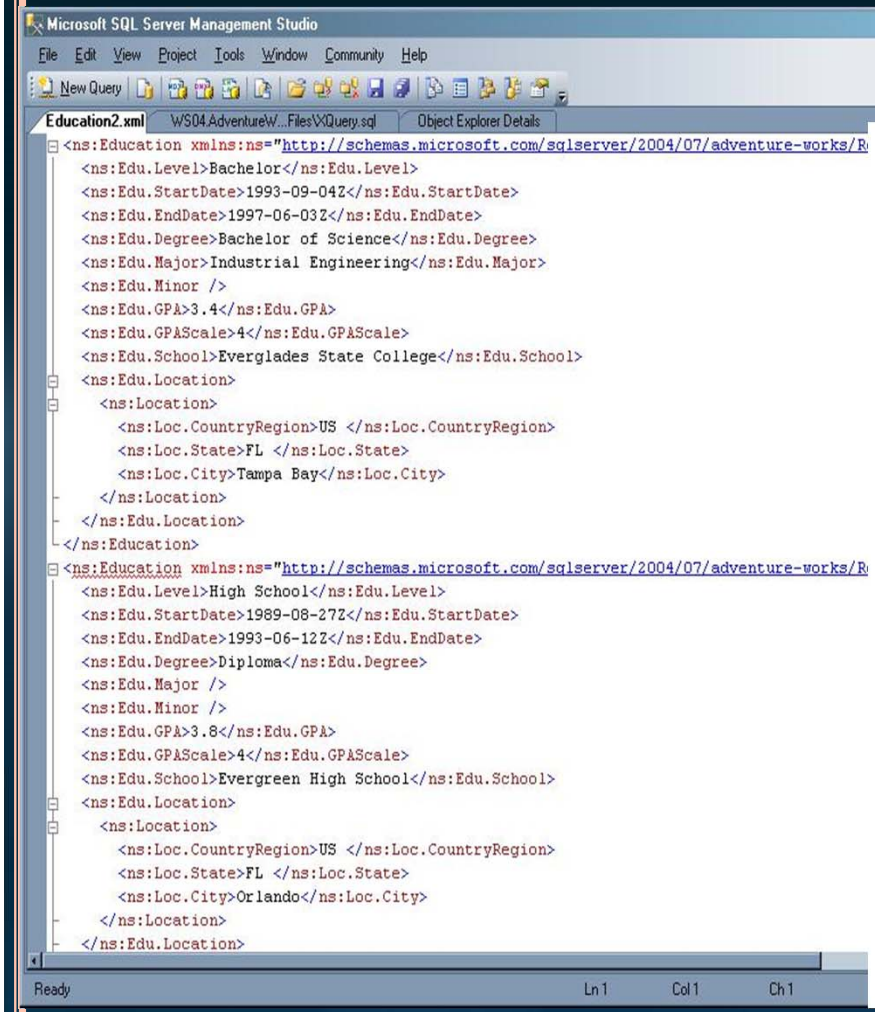


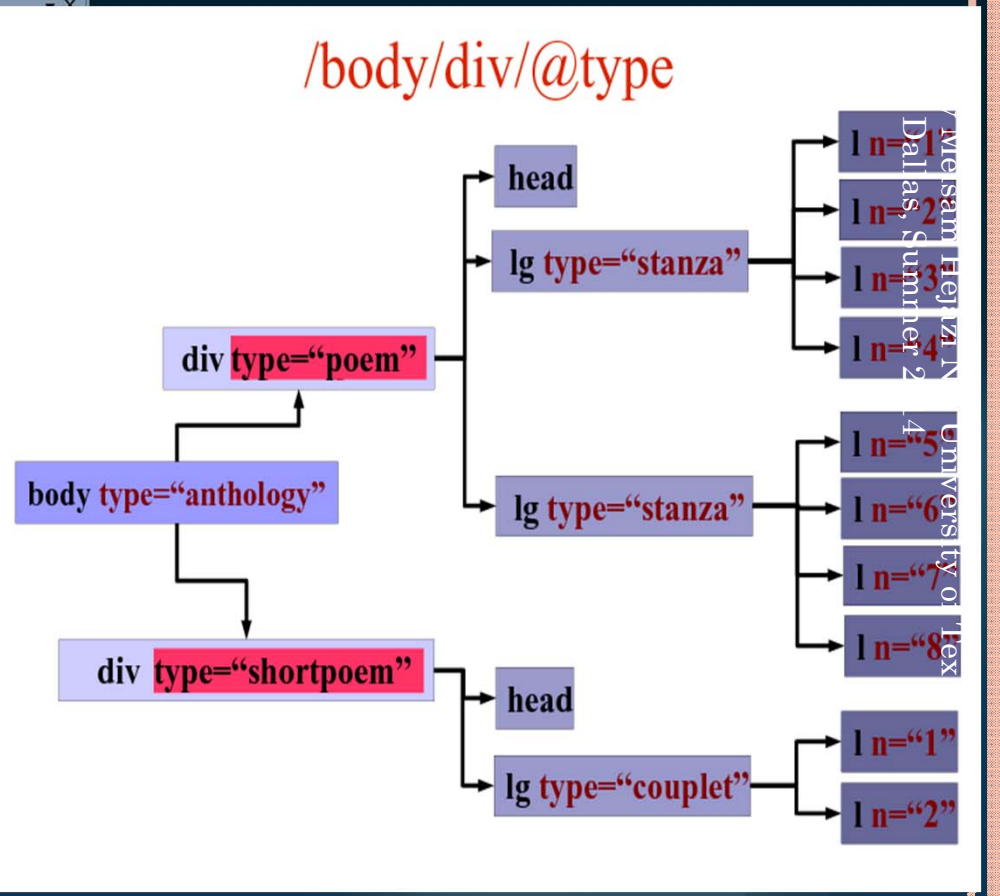
Manual



- 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





SELECTING NODE

Expression		Path Expression	Result
nodename	Selects all nodes with the name "*nodename*"	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: *In JavaScript:* *xml*.setProperty("SelectionLanguage","XPath");
/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.

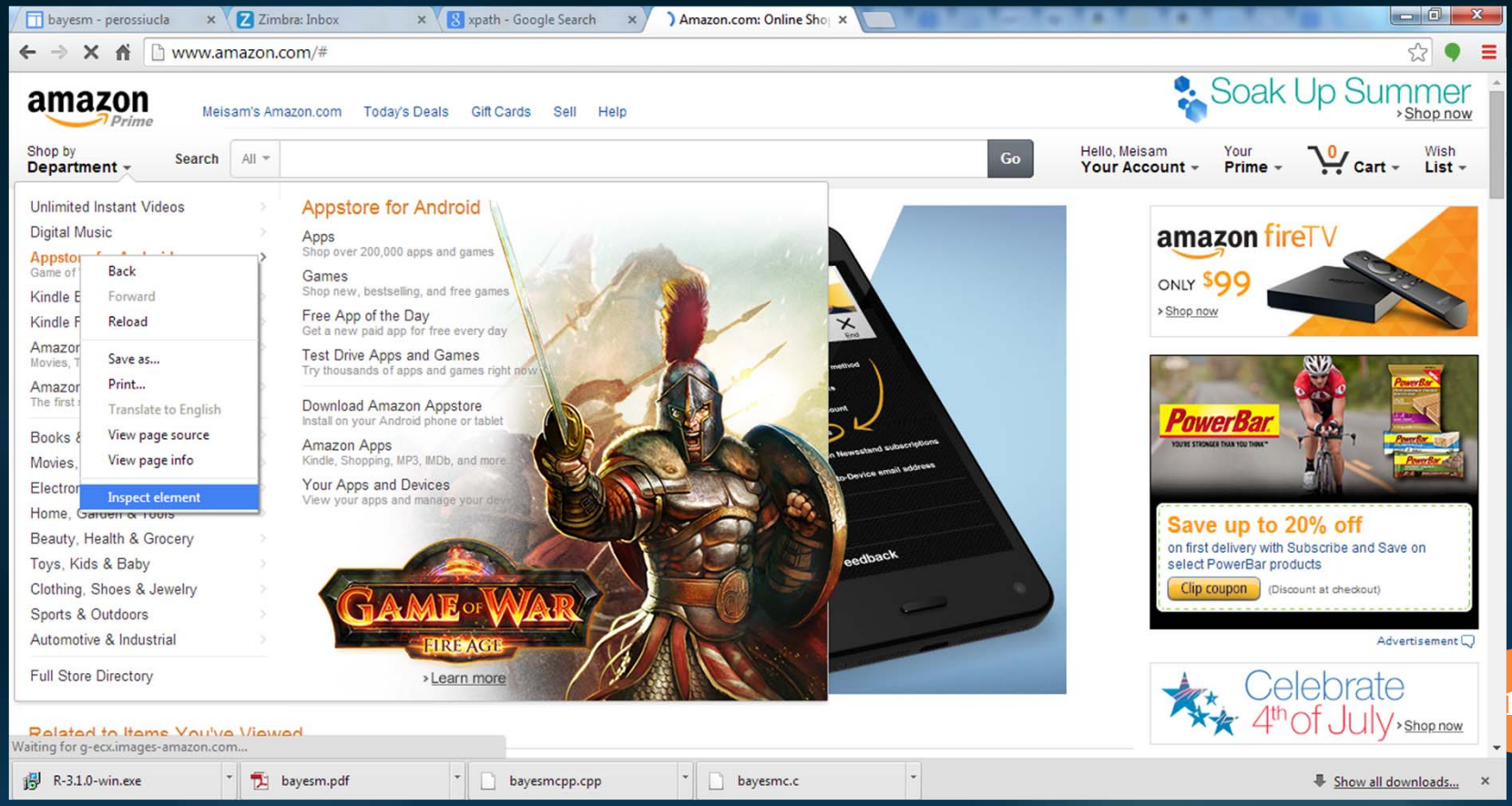
AxisName	Result	Example	Result
following-sibling	Selects all siblings after the current node	ancestor::book	Selects all book ancestors of the current node
namespace	Selects all namespace nodes of the current node	ancestor-or-self::book	Selects all book ancestors of the current node - and the current 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

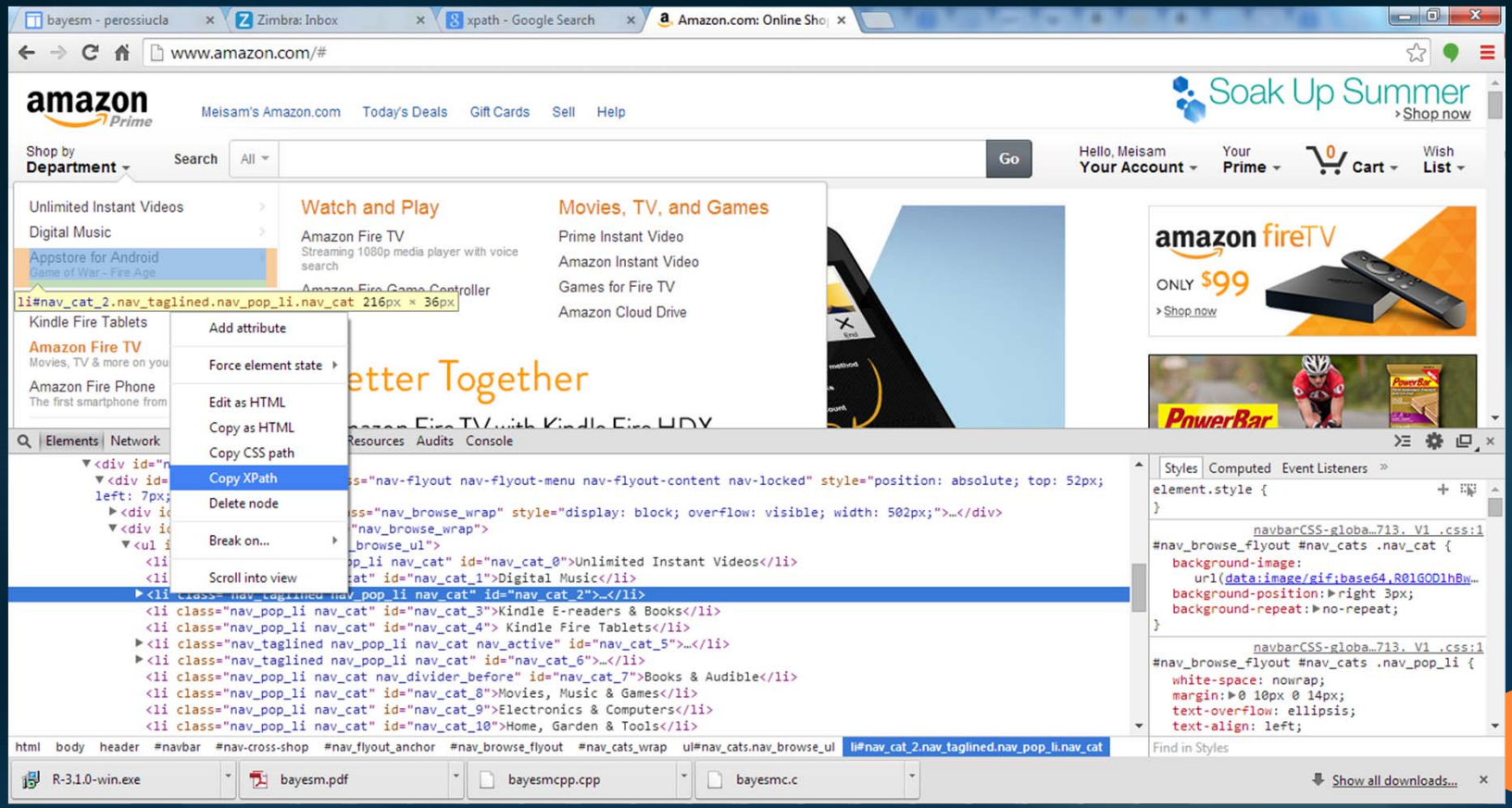
Easier way to get Xquery?

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



Easier way to get Xquery?

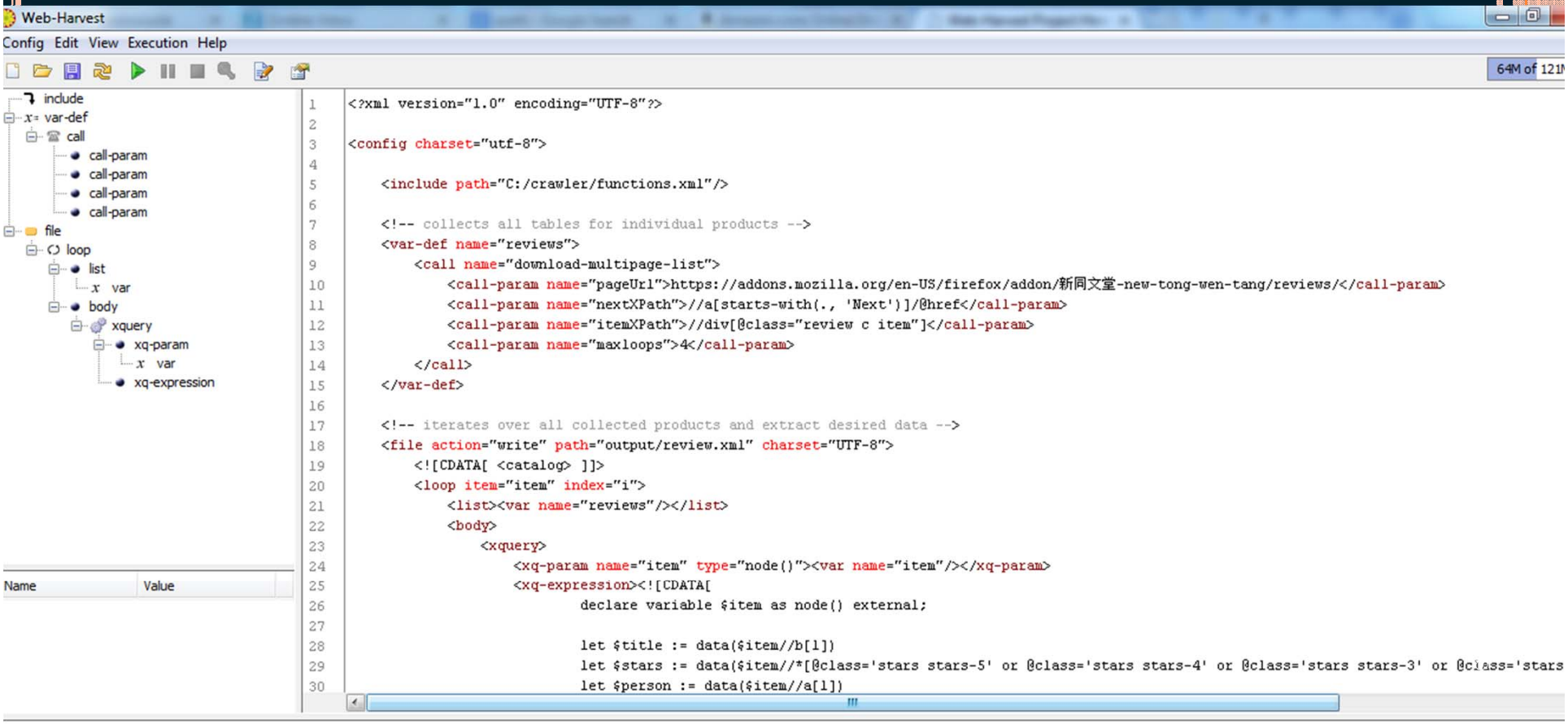
- Right click on the colored element and select copy XPATH



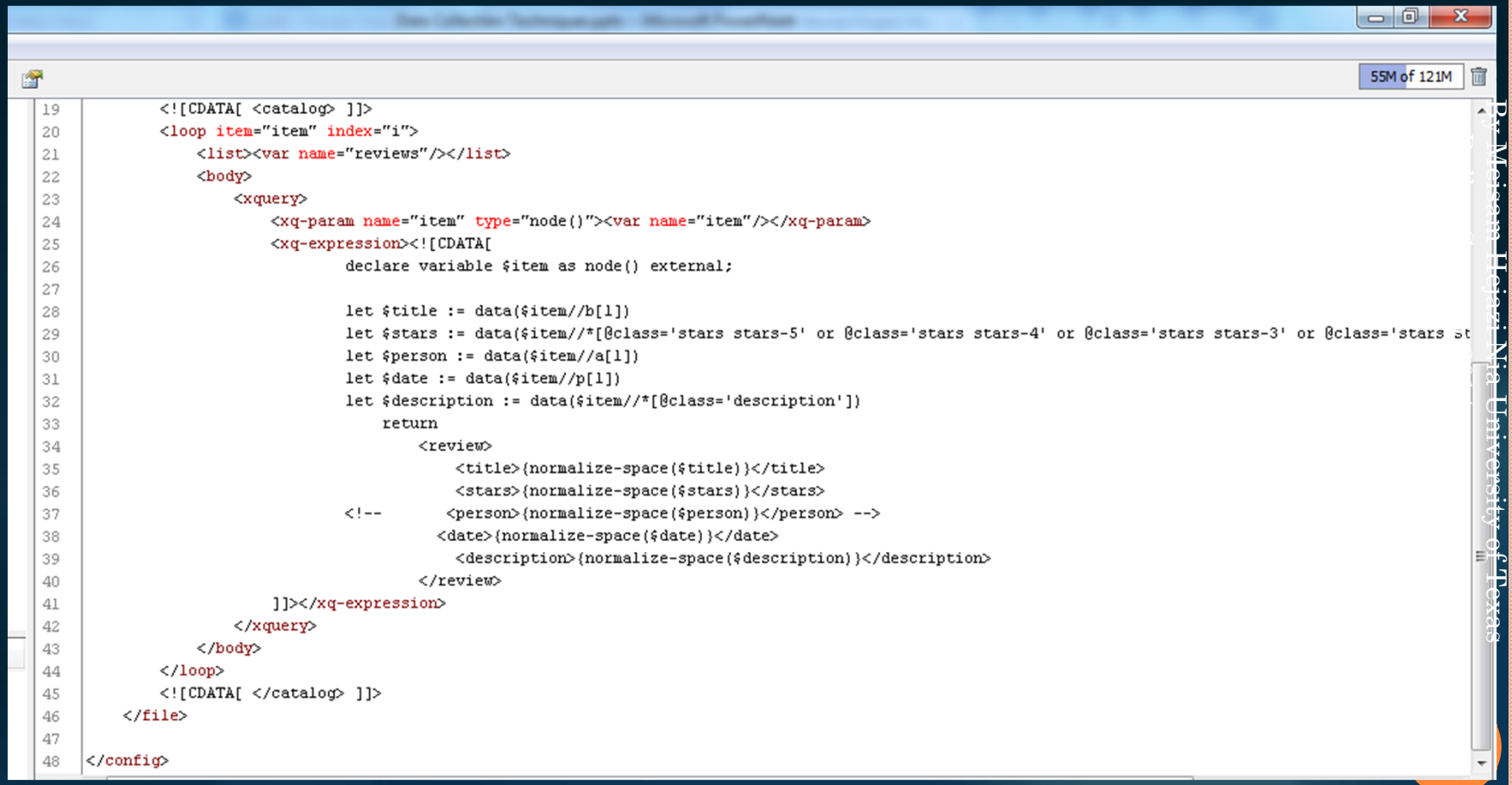
By Meisam Hejazi Nia University of Texas at Dallas, Summer 2014

Open Source Option: Web Harvest

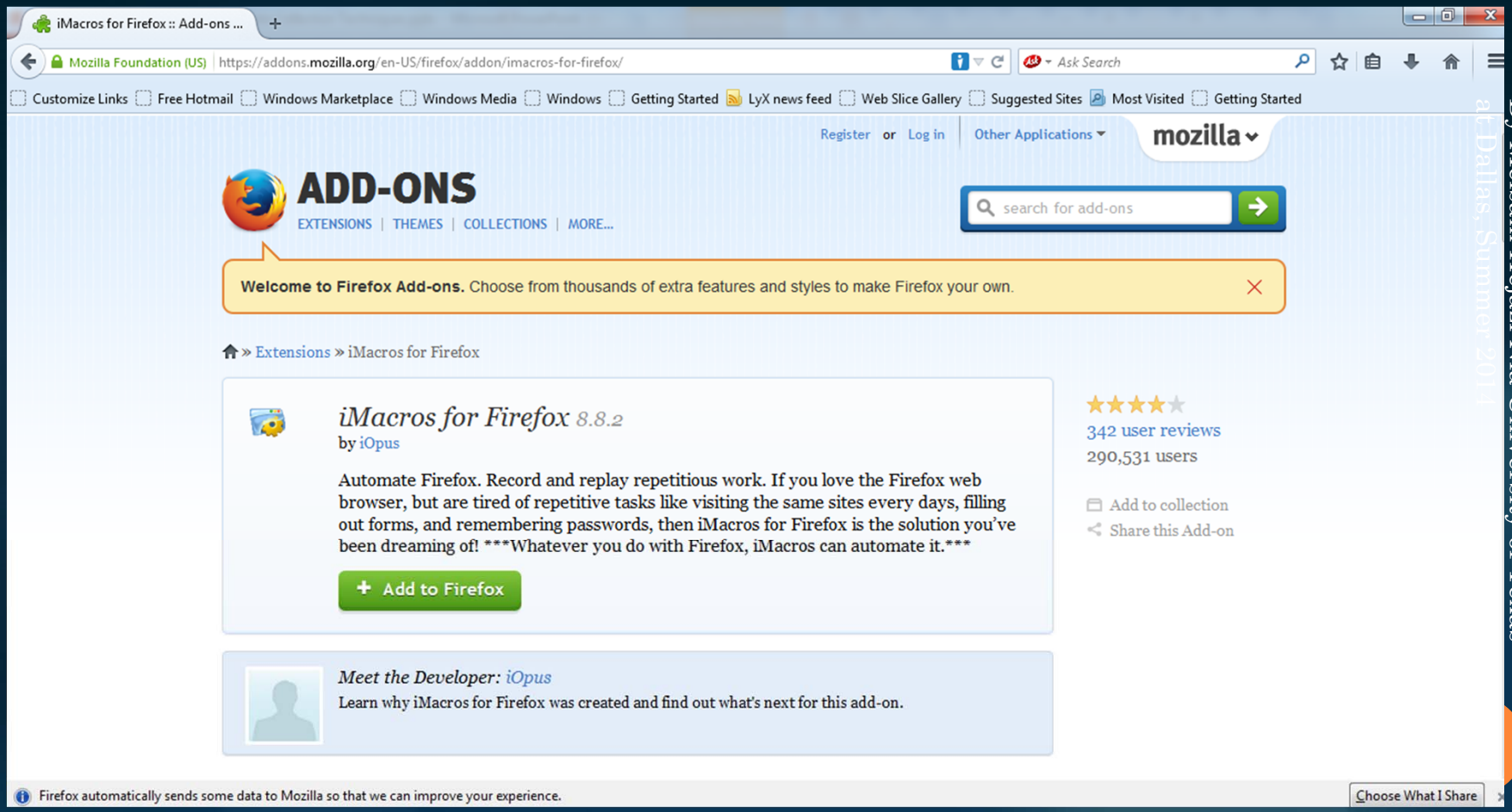
- Download from: http://web-harvest.sourceforge.net/



Open Source Option: Web Harvest

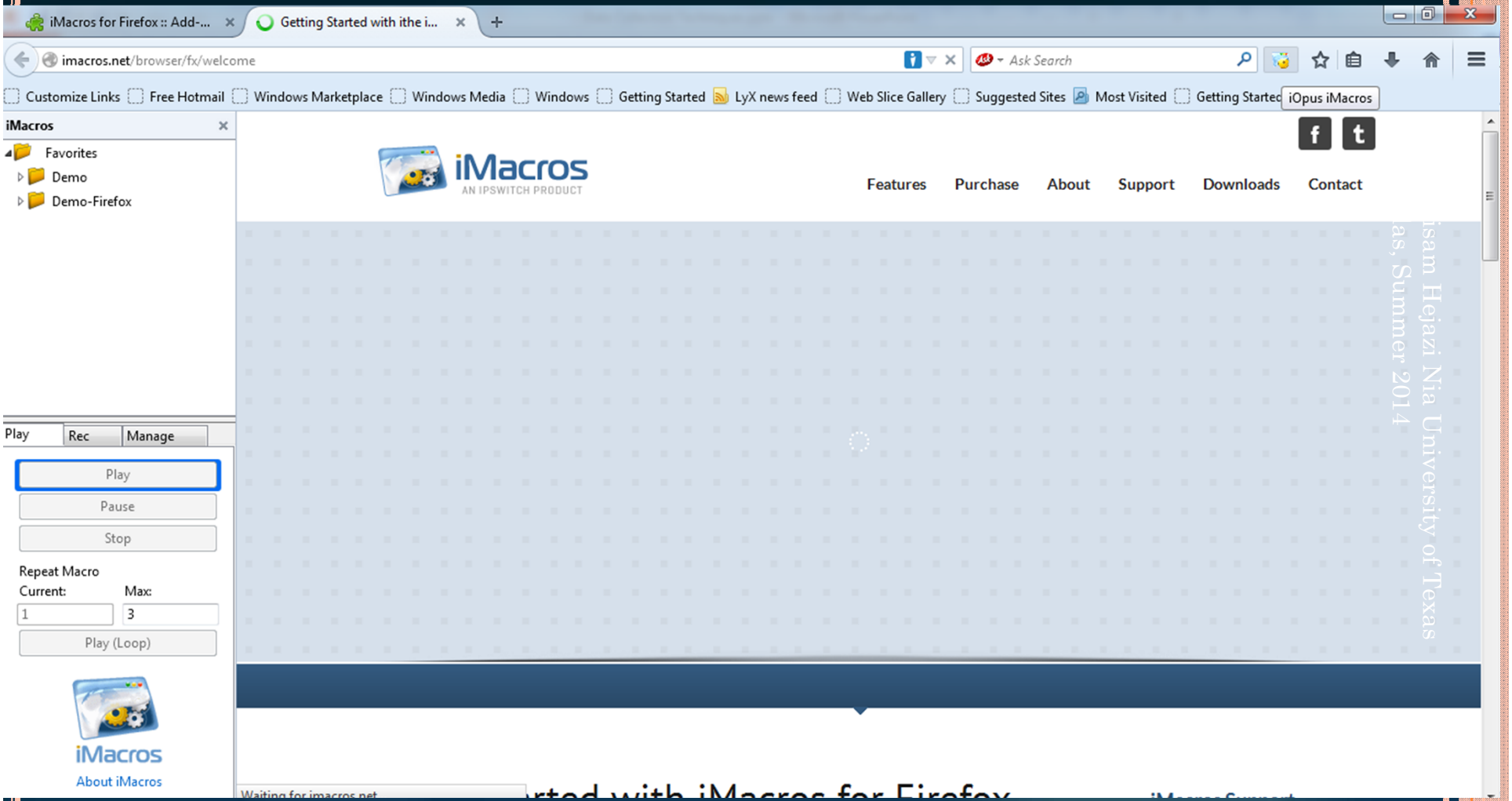


iMacro Recording

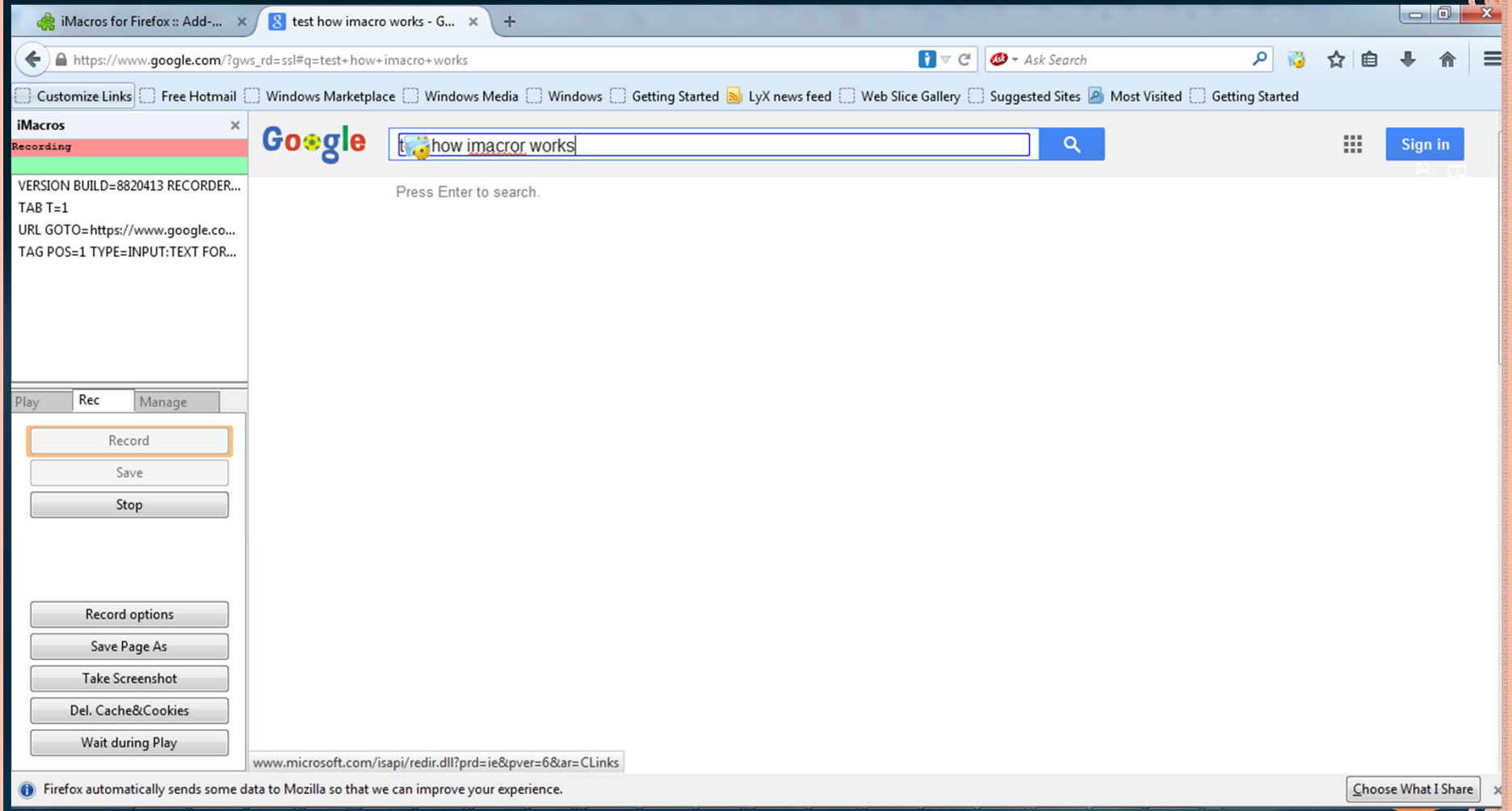


By Meisam Hejazi Nia University of Texas at Dallas, Summer 2014

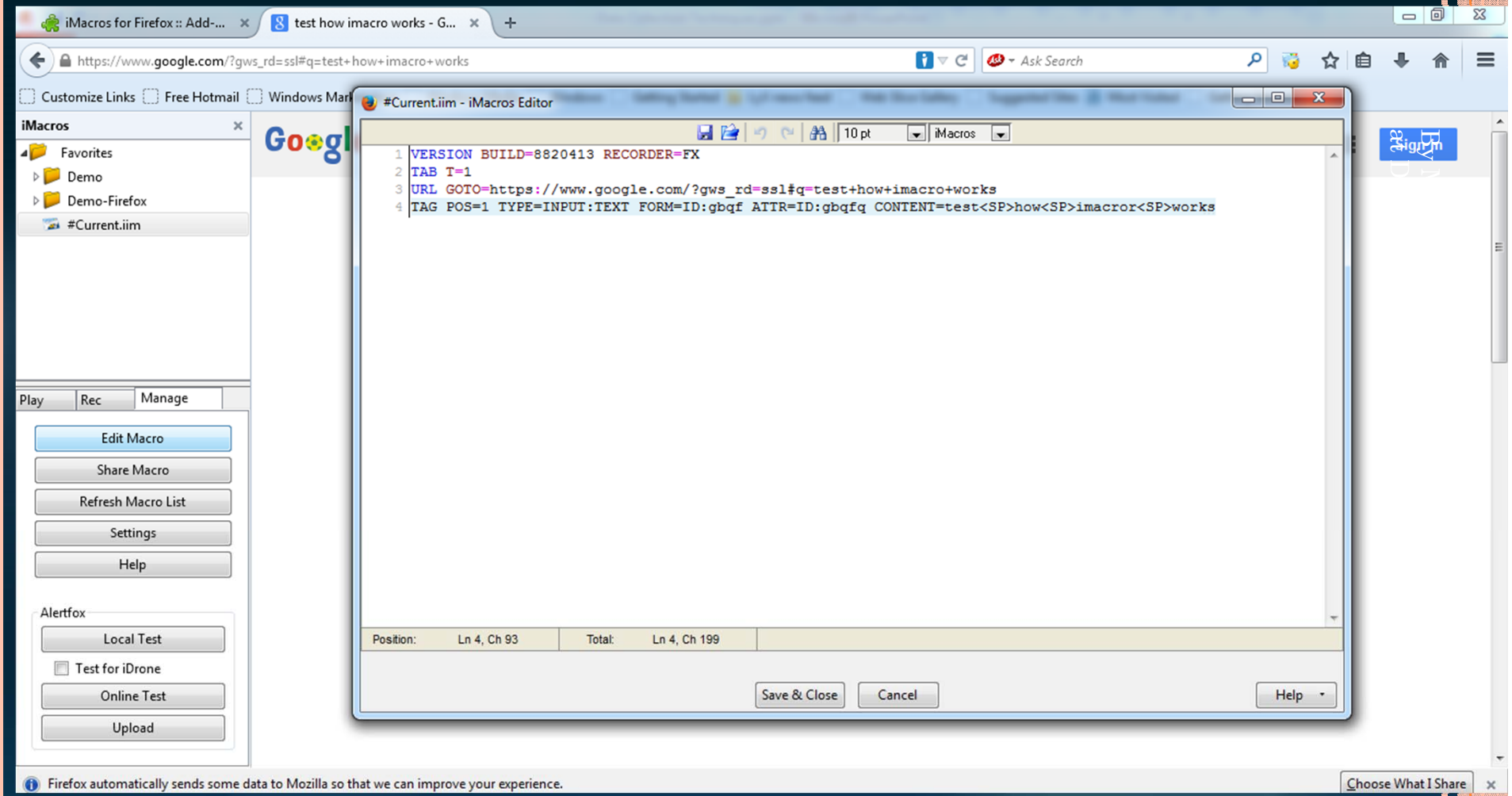
iMacro Recording



Recording


iMacros for Firefox :: Add-...
test how imacro works - G...
https://www.google.com/?gws_rd=ssl#q=test+how+imacro+works
Ask Search
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Getting Started
LyX news feed
Web Slice Gallery
Suggested Sites
Most Visited
Getting Started
iMacros
Recording
VERSION BUILD=8820413 RECORDER...
TAB T=1
URL GOTO=https://www.google.co...
TAG POS=1 TYPE=INPUT:TEXT FOR...
Play
Rec
Manage
Record
Save
Stop
Record options
Save Page As
Take Screenshot
Del. Cache&Cookies
Wait during Play
Google
t how imacror works
Sign in
Press Enter to search.
www.microsoft.com/isapi/redir.dll?prd=ie&pver=6&ar=CLinks
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share

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 releative 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
FIlTER 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:yymmdd_hhnnss}}.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
'==========================================
```

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%2Ck%3
A{{mangaName}}&page={{nextPageNum}}
WAIT SECONDS=1


SAVEAS TYPE=CPL FOLDER={{savepath}}
FILE={{mangaNumber}}.{{mangaName}}_SearchedPage_P{{nextPageNum}}_{{!NOW:ddmmyy_hhnnss}}.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/

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";
```

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

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();
```

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

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

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; *or*: shortest match
|	alternative
( )	grouping; "storing"
[ ]	set of characters
{ }	repetition modifier
\	quote or special

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*,}	at least *m* *a*'s
a{*m*,*n*}	at least *m* but at most *n* *a*'s
repetition?	same as *repetition* but the *shortest* match is taken
\t	tab
\n	newline
\r	return (CR)
\x*hh*	character with hex. code *hh*
\b	"word" boundary
\B	not a "word" boundary

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

Regular Expression – Strings

\w	**matches any *single* 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
characters]	matches any of the characters in the sequence
[*x*-*y*]	matches any of the characters from *x* to *y* (inclusively) in the ASCII code
[\-]	matches the hyphen character "-"
[\n]	matches the newline; other single character denotations with \ apply normally, too
[^*something*]	matches any character *except* those that [*something*] 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/Lattitudeoutput.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!

Data Cleaning in Excel-String methods

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

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

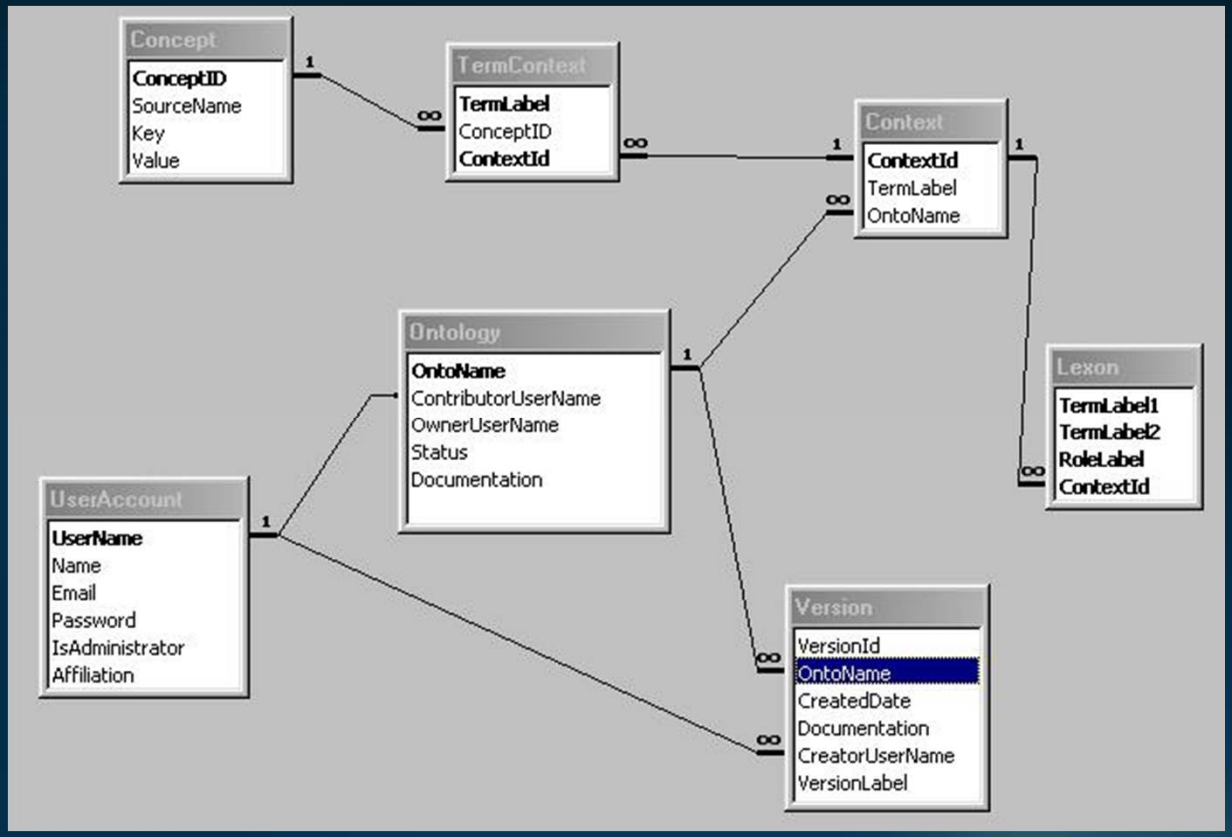
Data Cleaning in Excel-String methods

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

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


Concept
ConceptID
SourceName
Key
Value
TermContext
TermLabel
ConceptID
ContextId
Context
ContextId
TermLabel
OntoName
Ontology
OntoName
ContributorUserName
OwnerUserName
Status
Documentation
Lexon
TermLabel1
TermLabel2
RoleLabel
ContextId
UserAccount
UserName
Name
Email
Password
IsAdministrator
Affiliation
Version
VersionId
OntoName
CreatedDate
Documentation
CreatorUserName
VersionLabel

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

- **SELECT**
  - **FROM**
    - **WHERE**
      - **GROUP BY**
    - **HAVING**

SQL SAMPLE QUERIES

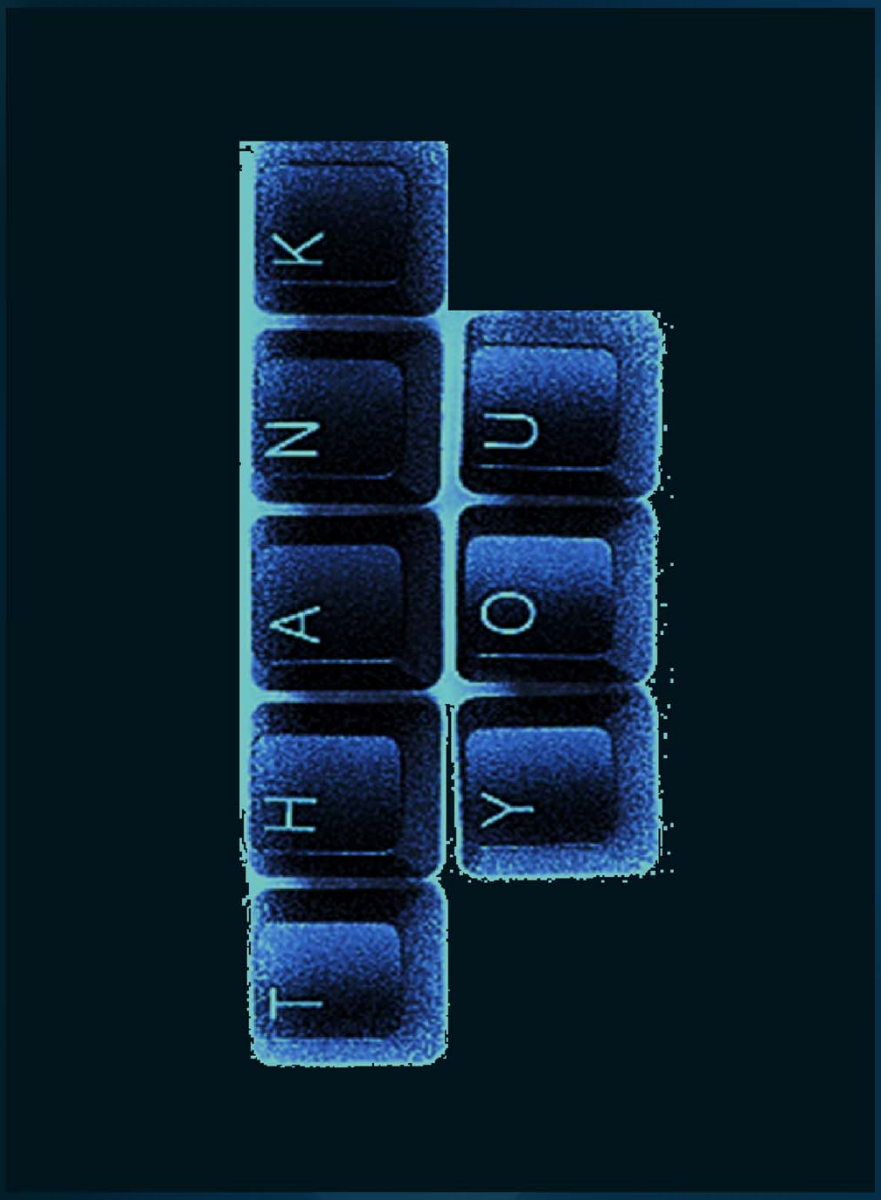
```
SELECT DailyPurchaseWithCategory.*,
UsersRegion.AreaID INTO
DailyPurchaseWithLocation
FROM DailyPurchaseWithCategory, UsersRegion
WHERE
(((UsersRegion.UserID)=[DailyPurchaseWithCateg
ory].[UserID]));
```

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 !


THANK YOU