

Data Collection on the World Wide Web using Excel¹

1. Introduction

The main purpose of this document is to take a better look at a few tools available within Microsoft Excel that facilitate the extraction of data directly from the Internet. For example, Excel allows users, at the simple click of a button, to automatically pull economic data and other type of information from your intranet server or from the Internet and transfer it directly into an Excel worksheet for tracking and analysis. This type of function could be particularly useful for OECD statisticians involved in the gathering and updating of data from various sources.

Much of the information presented to Internet users today is in HTML (Hyper Text Markup Language) tables. Indeed, tables are a useful way to organize information and display it effectively and attractively. It is possible to import HTML information, especially tables, directly into Excel worksheets.

This paper examines Excel's Web Queries. This feature goes beyond allowing the simple importation of HTML information and lets the user query a specific Web page or server, and receive the result directly in an Excel worksheet. A query can be automated, it can prompt for parameters, or it can use the contents of a worksheet as input. From Excel worksheets, it is possible to use Web Queries to pull "live" updated data from the Internet or an intranet, and then perform calculations and analysis with the updated data. Information can be refreshed automatically, and as often as needed, while maintaining worksheet layout and formulas unchanged, even if the amount of data returned changes.

Excel supports hyperlinks that allow users to click on a cell or object and connect to an Internet or intranet Web page, another Excel worksheet, or any Office document. Excel also supports several HTML extensions that allow tables to be displayed normally in a worksheet. An Active Server Page (ASP) web site can also be imported, using a Web Query, into an Excel worksheet. Finally, all of these tools can be fully automated using Excel Visual Basic for Applications for custom solutions.

In short, almost any static or dynamic Web page (HTML, ASP, Common Gateway Interface (CGI), etc.) that require parameters to be loaded and passed to a table page, can be imported into Excel using the Web Query functionality (see "In what condition are Web Queries possible?" in section 3 of this document).

This document² will cover the two following topics:

- Hyperlinks (including linked cells)
- Web Query

¹ This document has been prepared by Eric Déry, Principal Technical Assistant, Statistics Directorate (STD), Statistical Technology Section (STS), OECD

² A lot of the information in this document is based on information provided in "Microsoft Excel Web Connectivity Kit"

2. Hyperlinks

Many Statistical Institutions, Central Banks and other organizations disseminate their data on the web in Excel-type files, e.g. XLS, CSV, PRN, etc. Excel allows to easily link any cells or object to another location in the worksheet, another worksheet, or an Internet or intranet URL, or even another network address. Just as easily, it is possible to link a cell or an object to a specific file anywhere on the Intranet and Internet. A hyperlink can be created by clicking on INSERT >> HYPERLINK in the Excel menu or by typing the following formula into a cell:

=HYPERLINK("link or file location","friendly name")

This functionality can also be assigned to a Visual Basic command that will open any Excel-type file on the World Wide Web.

Example:

```
Workbooks.Open ("http://wdb.cnb.cz/cnbeng/docs/BALOFPAY/PB_EN.XLS")
```

This option can be extremely useful to extract many series coming from many different files stored in different pages on the Internet. These extractions can be completely automated using some simple Visual Basic procedures.

2.1 Linked cells

Just as it is possible to retrieve data from an Excel type file stored on a local Intranet or on the Internet by opening them using a hyperlink, Excel also allows cells to be directly linked to other cells on Excel files situated on the Web. By refreshing links in an Excel workbook, a user can actually extract live data coming from the Intranet without having to open any files or having to navigate on the Web.

3. Web Queries

3.1 What are Web Queries?

A Web Query is a feature in Excel that allows to retrieve data stored on an Intranet or the Internet. This feature creates a HTML page in an Excel worksheet by passing along the necessary parameters, required by the structure of the web page, to display the data in a workbook. A Web Query can use static parameters, dynamic parameters, or a combination of both. Queries with static parameters send a query without any input; queries with dynamic parameters prompt the user for input or can use a pre-determine range specified by the user. Regardless of the type of parameters in the query, the requested information is pulled from an Internet or an Intranet site, and the results are placed in a specified worksheet. The capacity to build queries with dynamic parameters enables Excel to be "linked" to a web site containing a database structure (ex.: NewCronos from EuroStat Web site).

3.2 In what condition are Web Queries possible?

If a user can display the data on the screen using a browser, then, there are good chances that he or she can use a Web Query to retrieve the data into an Excel worksheet.

On the other hand, there are some cases where Web Queries are not feasible. For example, if a web page stores “session variables” (i.e. parameters that are stored on the server and not displayed into the source of the web pages; their values are usually different every time one enters the web site) it could become impossible to create a Web Query unless the user knows the value of the session variables (e.g. predefined login and password details). Some web sites will assign an “order number” to a parameter which will change every time data is extracted from their database and it will only be valid for a short period of time. In such cases, it is impossible to know in advance the “order number” that will be generated by the web site, making Web Queries unfeasible.

3.3 How to create a WebQuery?

A Web Query is a text file with the file extension “.iqy”. It consists of three or four lines of text separated by carriage returns. Once a query is run in a worksheet and then the worksheet saved, the IQY file is no longer needed for that worksheet. The query information is saved with the worksheet and can be re-run anytime. IQY files are only used the first time a query is run in a given worksheet to establish the data location and parameters. However, an IQY file is not always necessary to create Web Queries; they can also be generated using Visual Basic. In such cases, all the information usually stored in the IQY file will be found inside the Visual Basic codes.

3.3.1 Creating Web Queries with a IQY file

An IQY file creating a Web Query is made with the following syntax:

Type of Query (optional)
Version of Query (optional)
URL (required)
POST Parameters (required for queries referencing POST forms/data)

A Web Query can be created using any text editor, such as Notepad. There are two basic types of queries: static and dynamic. A static query does not prompt the user for information, while a dynamic query prompts the user for values that it uses in the query. For information about static and dynamic queries, see “Static and Dynamic Parameters” later in this document.

➤ Type of Query:

The only valid entry for this field in a Web Query is WEB. If this value is omitted, Excel assumes WEB. This value is optional unless a version of query is specified. In other words, type of query and version of query must be used together or not at all.

➤ Version of Query:

The only valid entry for this field is 1. This value is optional unless a type of query is specified. If type of query is not specified, then this line should not be included.

➤ URL:

This is the file location where the query is sent and it is the only required field unless the Web page being queried is a POST type. It takes the form `http://server/file`; or in the case of a local file, `drive:\folder\file`; or on a network share, `\\server\share\folder\file`.

When building a query for an existing Web page, data can often be entered through the user's browser in the site's form. The resulting URL can then be copied from the browser's address field to this line in the query.

Copying the URL works for GET HTML forms, where the parameters are appended to the URL, such as `http://webservices.pcquote.com/cgi-bin/excelget.exe?TICKER=msft`. POST HTML forms require the parameters to be sent to the server as a separate line of text after the URL. For GET queries, however, the parameters are included immediately after the URL and a question mark (ex.: `http://server/file?parameters`)

With a HTML form, there are two methods for sending parameters to the server: GET and POST. Essentially, GET is used when sending small amounts of information, and POST is used for larger amounts of data. The GET method appends the data to the URL, whereas the POST method sends it as a separate line of text. GET is the default method, but because a URL of more than about 200 characters could cause problems, the POST method ensures that a large amount of data is processed correctly. The server application and the creator of the HTML form determine which method to use. POST parameters go in the fourth line of text after the URL in a Web Query.

➤ Static and Dynamic Parameters:

In Web Queries, static parameters send query data without prompting the user; while dynamic parameters prompt the user for one or more values, which are used in the query. Static queries include the parameter names and the values to be passed to the server. If there are multiple parameters, they are separated by an ampersand (&):

```
parameter1=value1&parameter2=value2
```

To create a dynamic parameter, the values must be replaced with two arguments in braces. The first argument is the argument name, and the second argument is the text, enclosed in quotation marks, to be displayed in the Excel-generated dialog box. Dynamic parameters are in the following format:

```
parameter1=["value1","Prompt for first value"]&parameter2=["value2","Prompt for second value"]
```

The braces in a dynamic query signal Excel to build a dialog box when the query is run that prompts the user with text in the second argument. The user's response becomes the value of the first argument.

3.3.2 Creating Web Queries through Visual Basic

The following is an example of a Visual Basic procedure generating a Web Query:

```
Sub Create_Web_Query()  
  With ActiveSheet.QueryTables.Add(Connection:="URL;URL_Address", Destination:=Range("A1"))  
    .Name = "query_name"  
    .PostText ("parameters")  
    .Refresh BackgroundQuery:=False  
  End With  
End Sub
```

3.4 How to run a Web Query using an IQY file

To run a Web Query, the user must (i) point to Get External Data on the Data menu; (ii) click Run Web Query; (iii) select from the available Web Queries; and then (iv) click Get Data. Once a query (IQY file) is selected, the user is prompted for the desired location of the results, as shown in Figure 1.

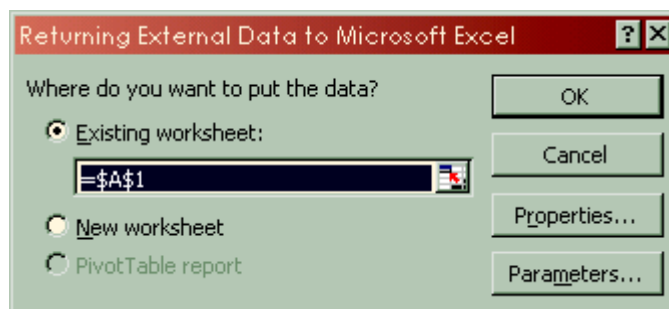


Figure 1: Excel asks the user where to place the results of the Web Query.

It is necessary to select the starting cell or range for the query results, or to select the New worksheet option. If the query is static (does not prompt for parameters), it runs, and displays the results in Excel. If the query is dynamic (prompts for parameters), Excel prompts the user with a series of dialog boxes like the one shown in Figure 2.

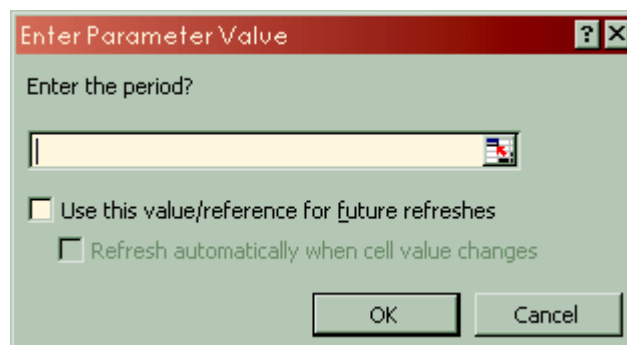


Figure 2: The Enter Parameter dialog box allows for dynamic parameters in a Web Query.

The user must enter the desired values and click OK. He or she can use the range selector button at the right of the entry field to select worksheet cells for the parameter values. If the user wants to specify multiple cells, Excel separates the range of cells from left to right and top to bottom to create the string of values to send to the server. The results of the query are displayed in the worksheet at the location specified by the user.

As mentioned above, once a query is run in a worksheet and the worksheet saved, the IQY file is no longer needed for that worksheet. The query information is saved with the worksheet and only needs to be refresh by using the refresh button in the Excel data menu as shown in figure 3.

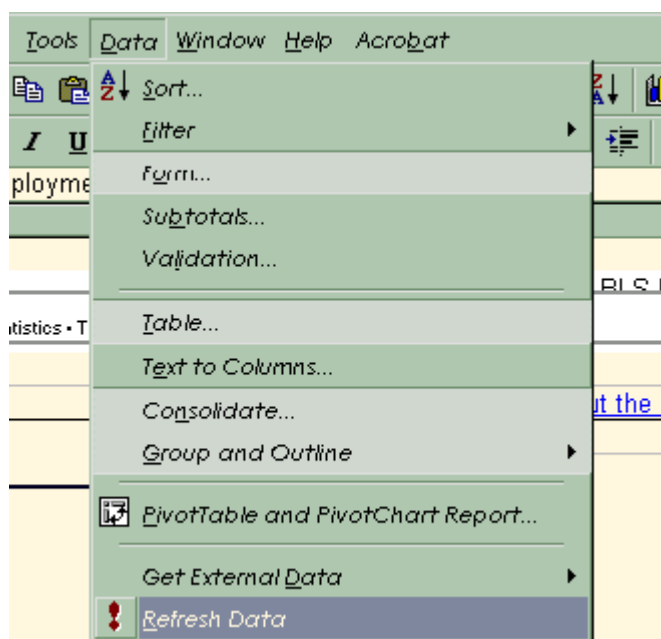


Figure 3: Refresh Data will rerun a Web Query.

IQY files are only used the first time a query is run to establish the data location and parameters. Once the Web Query has been created into a workbook using an IQY file, it cannot be modified within Excel. If some of the parameters in the Web site have been modified or if the URL address has changed, the IQY file has to be re-written, and the Web Query has to be re-imported into the workbook.

4. Conclusion

Web Queries are tools that can facilitate data collection from the World Wide Web. While the statistician will need some time up-front to set them up appropriately, Web Queries have the potential to significantly reduce the overall time spent on data collection.

APPENDIX A: The following is a list³ (so far identified) of member countries' Central Banks and National Statistical Institutions where Web Queries can be performed to retrieve data on a click of a button:

- ✓ National Bank of Belgium
- ✓ Statistics Denmark
- ✓ Banque de France
- ✓ Deutsche Bundesbank
- ✓ National Statistical of Greece
- ✓ Central Statistics Office Ireland
- ✓ Banca d'Italia
- ✓ De Nederlandsche Bank
- ✓ Statistics Norway
- ✓ National Bank of Slovakia
- ✓ Statistical Office of Slovak Republic
- ✓ Statistics Sweden
- ✓ Central Bank of the Republic of Turkey
- ✓ Bureau of Labor Statistics (United States)
- ✓ Statistical Office of the European Communities (EuroStat)

³ This list might not be complete since not all member countries' institutions were tested for possible Web Queries extractions. This list also include EuroStat as an International Organisation.