УДК 004.4

ANALYZE OF THE MOST POPULAR WEB DATA EXTRACTION TOOLS

Эльфадил Хамза
Одеський Національний Політехнічний Університет, УКРАЇНА

ANNOTATION. The main motivation of this work is to show and analyze a categorization of thetools explaining the weak and strong points of data extraction tools and how to use them.

Introduction: Nowadays we live in a world where information is present everywhere in our dailylife.In particular we are going to concentrate on the information we find in Web pages. At the beginning, the Web was designed as a sourceof data for a human use. It was built to guarantee that the content and theinformation could easily be understood and read by humans but not prepared to beused as data able to be treated by other applications.

Purpose of the work: We are going to use a set of tools that have been specifically designed for thispurpose. We will explain the data extraction process and then we willcharacterize each of these tools.

Research result: The aim of this section is to give a general view to the reader of each of the toolsused in this document. Its main features are shown here.

- **Dapper**

Dapper is an online tool which allows the user to extract information from Websites.

To use it all what we need is an Internet browser and Internet connection as this service is only available online. Dapper is at the moment in beta phase but it is totally functional.The usage of Dapper is totally free and we only need to create a new account touse it. We can create our own wrappers (or Dapps as they are called) or usewrappers already created from other registered users.Dapper is one of the easiest tools to use as its interface is totally graphical. Apartfrom extracting data it allows you to create Flash widgets or alerts using theextracted information. Link a Dapp output to another Dapp input to create somenew Dapps is another useful functionality.With Dapper after following the standard steps to select the content of interest wecould receive all the information without problems. We grouped the informationdistinguishing the main title, the description and the players list.

- **Robomaker**

Robomaker is a Web 2.0 developer platform for creating mashups. The tool lets the user create RSS feeds, REST Web Services or Webclips in few steps.It is provided with powerful programming features including interactive visualprogramming, full debugging capabilities, an overview of the program state andeasy access to context-sensitive online help, this features make it really completeand dynamic. It can be used in both Windows and Linux platforms.Robomaker presented no problems when extracting simple data. We only had toselect the title and the description to extract these fields and introduce a loop toselect all the players.

- **Lixto**

The Lixto Visual Developer (VD) is a software tool that allows the user to define wrappers, which visually access data in a structured way, as well as configuring the necessary Web connectors.The program is originally from a research project of the Technical University ofVienna that becomes later in the Lixto Software. It provides businesses witheffective, user-friendly, and time critically viable wrapping, integration and deliveryof information all in the same product. First of all we have to know that Lixto VD only extracts our results in the XML format. First of all we have to create a Lixto Data Model to specify how the outputto our XML file will be.

- **WinTask**

WinTask is a Windows tool used to automate repetitive tasks or actions which should run at a certain moment. One of its features is data extraction of Web sites. WinTask can launch the URL to load, send a userid and an encrypted password if it is a secure site, conduct searches, and navigate to the different pages where some field contents have to be extracted. This tool is only available in the trial-version, if we want full functionality we have to buy it. It works by using its own scripts so at the beginning it can be a little hard to familiarize with the syntax.To extract data with WinTask we have to edit a script file that will extract all thefields of interest. First of all, we need two orders, one to open the Internet explorerand other one to load the Web page source.Then we only have to use the graphical interface to extract all the fields. Noproblems have been encountered with this tool and all the information has correctlybeen extracted.

- **Automation Anywhere**

Automation Anywhere is a Windows tool that lets the user record click and mouse movements and to create tasks in desktop that could interact with our programs. It can also record from the Web, this consists basically of creating a navigation sequence and extract data of our interest.We can also use templates to realize concrete tasks or use the task editor that letsthe user create a task using some predefined actions, conditions, scripts, mouseand keyboard activity... This tool is only available in the trial-version, if we want fullfunctionality we have to buy it. With Automation Anywhere we only have to create new variables to save the extracted values. Once created, we only have to select the specific content to extract and establish the relation to these variables. Then our results are extracted and can be outputted.

- **web Content Extractor**

Web Content Extractor is a Windows tool that allows the user to create a project for a particular site, extract data from it and store it in the current projects database. The extracted data can be exported to a variety of formats including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script or MySQL script.As it happens with the two tools analyzed before, we could only download the trialversionof Web Content Extractor.

- **Roadrunner**

Roadrunner is a project of the database departments of the *Università di Roma Tre* and the *Università della Basilicata*. This tool generates a wrapper for the analysis ofsimilarities and differences from several sample files of the same class.With this tool, a class is an amount of pages generated by the same script, sostructurally the same, but in some places both content are quantitatively different.This wrapper is a representation of the investigated sample files in the form of aregular expression or so-called union-free regular expression (UFRE).Web Content Extractor presented no problems when extracting these fields. Withthis tool we only have to select the Web page source and select the fields we wantto extract. We have to name each of the extracted fields to be referenced.

- **XWRAP**

XWRAP is a tool that was developed at the Georgia Institute of Technology. Its developers described it as an XML-enabled wrapper construction system for Web information sources.The toolkit includes three components: Object and Element extraction, filterinterface extraction and code generation. The wrappers are generated as Javaclasses. To use it we have to enter the URL of our desired Web site and the customization of the extraction process results is done via the Web by XWRAP.To use it we need a separate Web server (such as Apache Tomcat).Using this tool to configure the data extraction process we have to edit the*sample.php* file

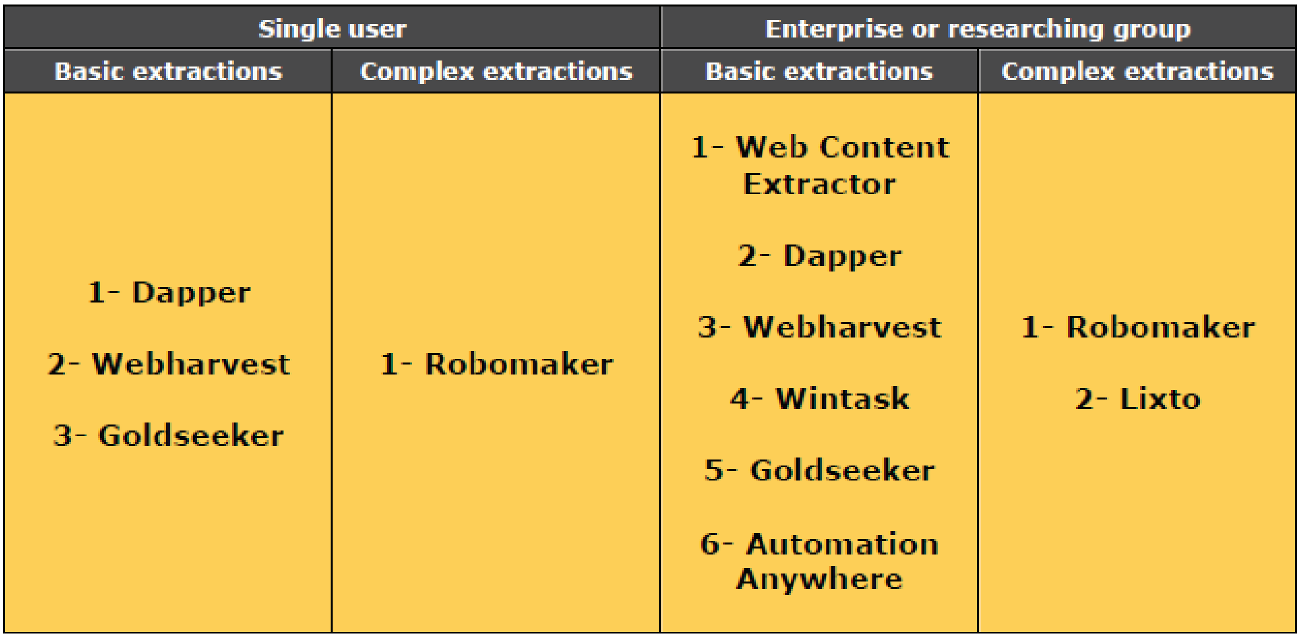
- **Webharvest**

Webharvest is an Open Source Web Data Extraction tool written in Java. It offers away to collect desired Web pages and extract useful data from them. In order to dothat, it leverages well established techniques and technologies for text/XMLmanipulation such as *XSLT*, *XQuery* and *Regular Expressions*. Web-Harvest mainlyfocuses on HTML/XML based Web sites which still make vast majority of the Webcontent.

- **Goldseeker**

Goldseeker is a data extraction tool, specifically a script under the GNU LGPL license. It was built to extract formatted data from HTML files, but it can be used with all kind of files. Its behavior is defined by a rule-based configuration file. It can process files on the local server or directly get Web pages via Internet. It is a development version, uncommented, undebugged and unfinished. Nevertheless, it can already be used for simple extractions.

Conclusion. On the other hand it is true that depending on the user profile and the dataextraction needs one tool will be more suitable than another. Considering that afinal user profile is a single user or an enterprise (or researching group) and takingcare of the complexity of the extractions and the price of the license, we are goingto construct a table. The presented order of the tools is used to give priority to thebest ones.



Single user		Enterprise or researching group	
Basic extractions	Complex extractions	Basic extractions	Complex extractions
1- Dapper 2- Webharvest 3- Goldseeker	1- Robomaker	1- Web Content Extractor 2- Dapper 3- Webharvest 4- Wintask 5- Goldseeker 6- Automation Anywhere	1- Robomaker 2- Lixto

REFERENCES

1. Baumgartner R.; Ceresna M.; Ledermüller G.: DeepWeb Navigation in Web DataExtraction. CIMCA/IAWTIC 2005

2. Chang C.; Kayed M.; Girgis M. R., Shaalan K. F.: A Survey of Web InformationExtraction Systems. IEEE Trans. Knowl. Data Eng. (TKDE) 18(10), 2006

3. 3,Crescenzi V.; Mecca G.; Merialdo P.: RoadRunner: Towards Automatic Data Extraction from Large Web Sites. VLDB 2001