Web Text Downloader v.2.0
User's guide

Osorio, Javier
javier.osoriozago@gmail.com

Reyes, Alejandro
alejandroe4@hotmail.com

May, 2014

Copyright

Terms of Use

How To Cite This Program

Please cite program as:

Javier Osorio and Alejandro Reyes. 2014. Web Text Downloader. Version 2.0.

BibTex citation:

```
@misc{Osorio2014,
address = {Puebla, Puebla},
author = {Osorio, Javier and Reyes, Alejandro},
title = {{Web Text Downloader v.1.0}},
url = {http://www.javierosorio.net/#!software/cqbi},
year = {2014}
}
```

Report Bugs

Please report any bugs to: `javier.osoriozago@gmail.com`

Updated copies of the Web Text Downloader program and manual can be found at `http://www.javierosorio.net/#!software/cqbi`

1 Introduction

Web Text Downloader is a program for automatically extracting and storing predetermined content from both open access and password-protected websites. The motivation to create this software emerged from the need to extract large numbers of news reports from an on-line news report repository that recurrently requested entering the username and password information. This software is an attempt to reduce the burden of such tedious effort. Web Text Downloader extracts web content that then can be processed with Web2Eventus to create the corpus of text of Eventus ID (Osorio and Reyes, 2014*b*,*a*).

For the purpose of this manual, Web Text Downloader is used to extract press releases issued by the Mexican Army (Secretaría de la Defensa Nacional, SEDENA). Alternatively, (Osorio, 2013) used this software to gather news reports contained in InfoLatina, a password-protected repository of newspapers. However, it is not possible to release such password. For this reason, the demonstration of the program exclusively focuses on SEDENA press releases.

Prior to using this program, the user must identify a series of links to be extracted. Notice that Web Text Downloader runs on Windows operating system.

2 Required Files

To extract web content, it is necessary to have two key files in the same directory:

1. The program file WebTextDownloader.exe
2. A plain text file in comma separated values (*.csv*) containing the links to be extracted and indicating the name used for saving the content of each link. For convenience, let's call this file ListOfLinks_DEMO.csv.

The ListOfLinks_DEMO.csv must contain the following information:

```
http://www.website.com/link_1  ,  FileName_1.txt
http://www.website.com/link_2  ,  FileName_2.txt
http://www.website.com/link_3  ,  FileName_3.txt
```

The content of ListOfLinks_DEMO.csv must have the following characteristics:

- Each line in the document must contain first the url of the link to be extracted and then the name of the file to be assigned to each link. These two elements must be separated by a comma (,).
- The link to be extracted must start with `http://`
- The file name must also indicate the type of file to be saved. In this case, the extension (*.txt*) refers to a plain text file.
- There should be no duplicate file names.

3 News Report File Naming Convention

This section indicates the elements necessary for adequately naming the news reports to be used in EVENTUS ID. Complying with this nomenclature is crucial for adequately formatting the corpus of text using the program WEB2EVENTUS.

3.1 Nomenclature Elements

The nomenclature consists of four main elements:

- Date
- Counter
- Source
- Extension

3.2 File Name

File names should be structured in the following way: `yyyymmddccc_SRC.ext`, where:

- `yyyy` is a four digit number representing the year (e.g. 2009).
- `mm` is a two digit number representing the month (e.g. 02 for February).
- `dd` is a two digit number representing the day (e.g. 17).
- `ccc` is a three digit number counting the number of reports issued by the same source in a given day. The three digits allow the counter to range from 001 up to 100.
- `SRC` is a short acronym indicating the name of the information source.
- `ext` is the extension indicating the format of the file (e.g. *.html* or *.txt*).

There should be no duplicate file names. The counter elements in the nomenclature (`ccc`) are an effective way of assigning unique file names even in contexts of a large number of news reports. For convenience, the counter considered in this example is set to three digits, but users can assign as many digits to the counter as they consider necessary.

3.3 File Name Example

For example, consider a set of on-line press release issued by SEDENA. There were two news reports on August 23, 2009 and another one on October 17, 2010. According to the nomenclature, the files should have the following names:

```
20090823001_SEDENA.html

20090823002_SEDENA.html

20101017001_SEDENA.html
```

4 Extraction Process

To extract web content using Web Text Downloader, users need to follow the following steps:

Step 1: Open Web Text Downloader

Open the program by double clicking on the executable file WebTextDownloader.exe.

Step 2: Access the main website

Enter the root address of website in the browser bar located in the upper left corner of the interface. In this example let's enter `http://www.sedena.gob.mx`, then click on "Ir" (Go). Users can specify any website they want. This website must be the root of all the urls contained in the list of links. Follow steps 1 and 2 in Figure 1.

Figure 1: Open a website within Web Text Downloader.

Users who want to extract content form a password-protected website should follow the same steps 1 and 2. Once the root website is loaded within Web Text Downloader's frame, the user must begin a new session by entering a valid username and password. Once the user opens a new working session, it is not necessary to enter the login information again. After finishing the extraction task, it is recommended that the user ends the session by logging out.

Many password-protected websites use cookies to verify the credentials of a valid user. Blocking cookies can prevent Web Text Downloader from performing as expected. This is a frequent problem with Windows Internet Explorer, but it can also occur with other browsers. Users must enable cookies in their browsers before using Web Text Downloader. Here is a link on how to allow cookies in Windows Internet Explorer:

`http://windows.microsoft.com/en-us/windows-vista/block-or-allow-cookies`

Step 3: Upload the list of links

To upload the list of links, click on "Abrir lista" (Open list). Then browse to the folder containing the file ListOfLinks_DEMO.csv. Select the file and click OK. This will upload the content and display the list of urls and their corresponding file names in the right window of the interface. Follow the arrow indicated in Figure 2. To open a single link, scroll down the list of links in the right window to identify the desired row, select the desired row by clicking on it and click F2 to open the url.

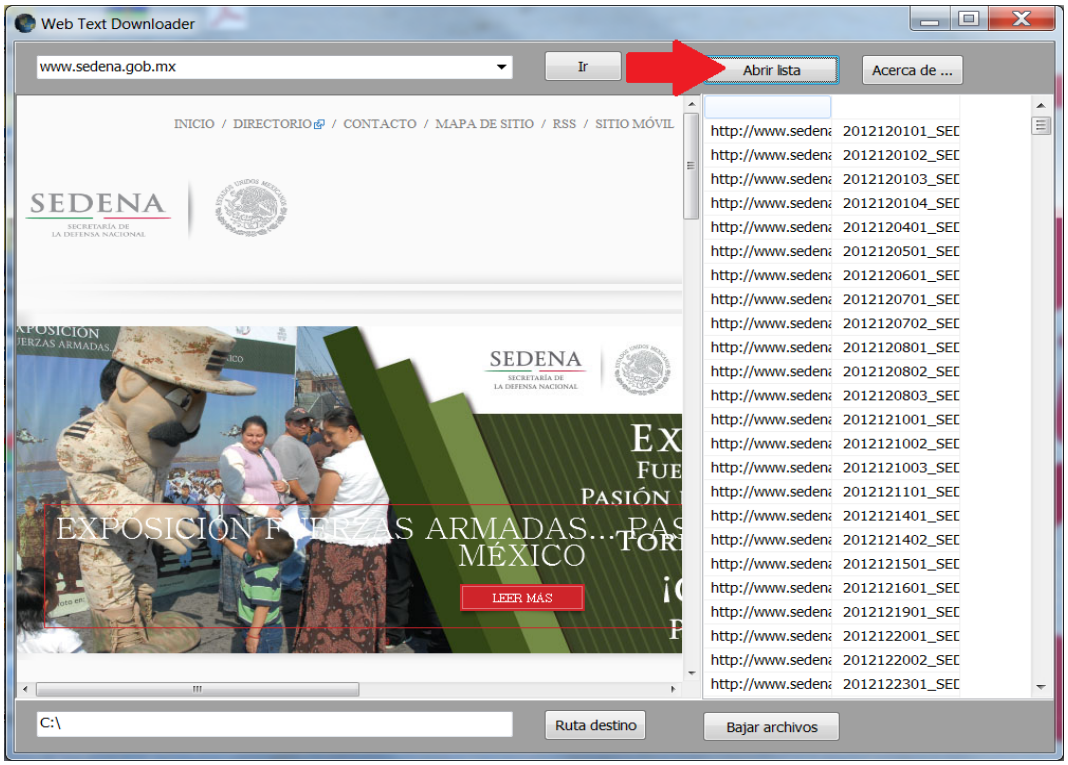

Figure 2: Upload the list of links to be extracted.

The demonstration file of Web Text Downloader contains a list of links of real press releases issued by the Mexican Army (Secretaría de la Defensa Nacional, SEDENA). The folder SEDENA_permission in the demonstration .zip file contains the official letter authorizing the use of these press releases as well as the required citations of the documents used in the list of links.

Step 4: Define a destination folder

The user must define a define a path for a destination folder to save the extracted files. To do so, click on "Ruta destino" (Destination route) located in the lower section of the interface. Then browse to the desired destination, select the folder to indicate the path and click OK. This step is illustrated in Figure 3.

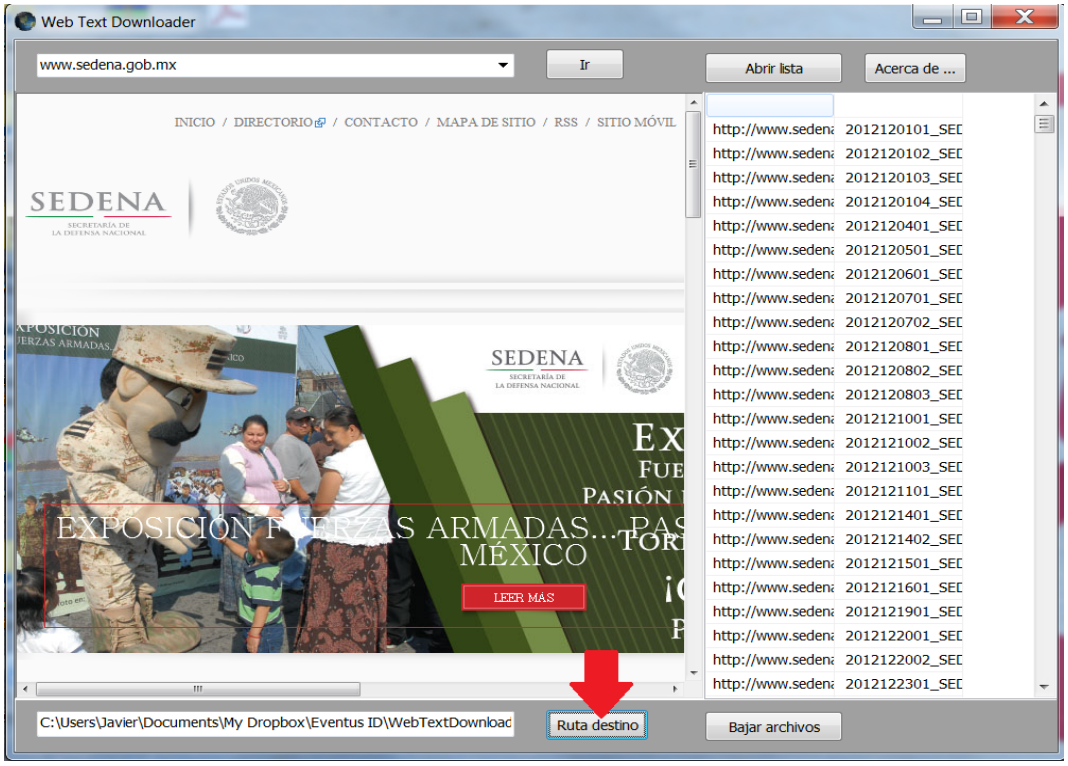

Figure 3: Select the destination folder.

Step 5: Extract the web content

To begin the automatic extraction of web content from the list of links click on "Bajar Archivos" (Download files). This button is located in the lower right section of the interface. See Figure 4. One url will be opened at the time in the internal browser of Web Text Downloader. The content of each url will be saved in an individual file according to the name specified in the list of links. Each file will be stored in the destination folder selected by the user in the previous step.

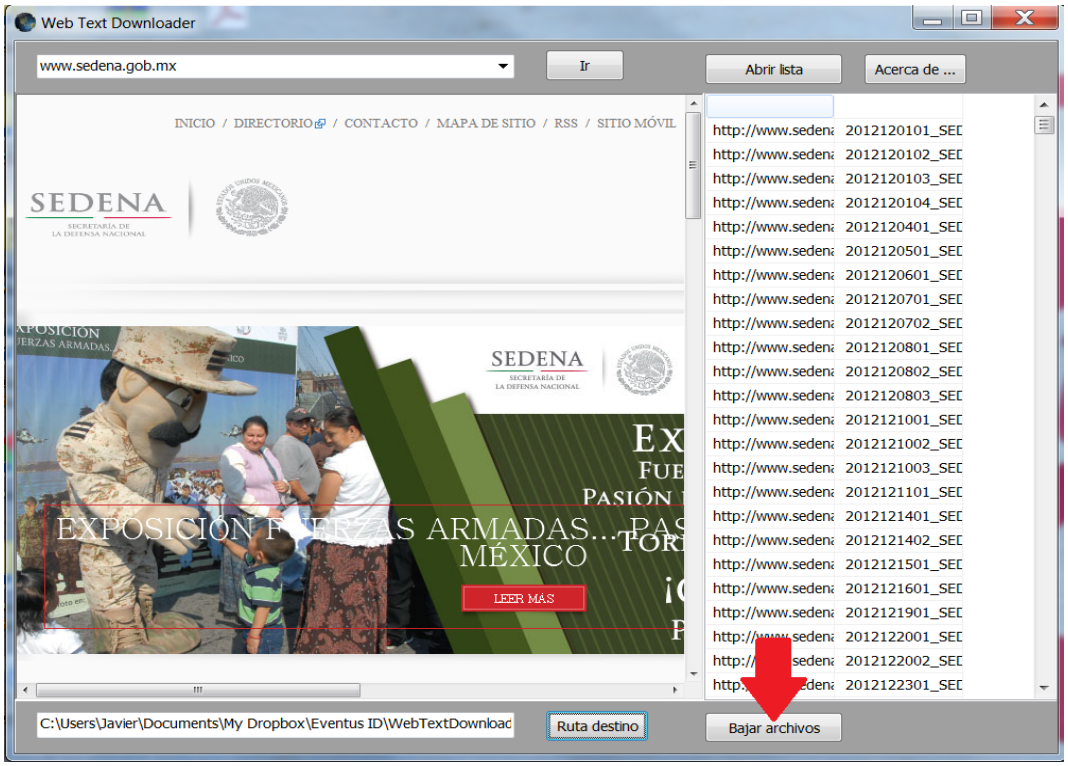

Figure 4: Select the destination folder.

References

Osorio, Javier. 2013. "Hobbes on Drugs: Understanding Drug Violence in Mexico.".

Osorio, Javier and Alejandro Reyes. 2014*a*. "Eventus ID. Supervised Event Coding From Text Written in Spanish.".
URL: *http://www.javierosorio.net/#!software/cqbi*

Osorio, Javier and Alejandro Reyes. 2014*b*. "Web 2 Eventus.".
URL: *http://www.javierosorio.net/#!software/cqbi*