

Dmoz Directory Data in Easy to Edit CSV Files

Support: ezmozfiles@gmail.com

Website: <http://directorydata.codeisle.com>, <http://www.codeisle.com>

Disclaimer: the maker of the provided files in this distribution is not affiliated with AOL, Dmoz.org, or any other company associated with Dmoz.org.

About Dmoz

From Dmoz.org: “The **Open Directory Project** is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors.”

Dmoz dumps its directory databases on a weekly basis in the RDF file format. This data includes information about every website in its directory, such as: website URL, website category, description, etc.

About Dmoz Directory Data

Dmoz directory data is normally in RDF format. This poses several problems when trying to work with the data:

1. The RDF files cannot be opened using applications that do not support the format.
2. The RDF files are quite large to manage (over 1GB uncompressed).
3. The RDF files contain a lot of data that may be of no interest to many people.

To address these issues, the provided data files in this distribution have been formatted into easy to view and edit **comma separated values** files (CSV). These files can be opened in LibreOffice, Microsoft Excel, Microsoft Access, a text editor, or other programs supporting comma delimited data.

System Requirements

- An application capable of handling large CSV files (e.g. Microsoft Excel [2007 or greater recommended], LibreOffice, a text editor [not recommended]).
- RAM: Minimum of 3GB is recommended.
- An operating system supporting Unicode characters.

Folder Structure

Inside the root folder there are two types of files for each main category: files with a “_category” suffix, which contain data related to the categories in the Dmoz directory (category id, category path, and category name), and files with a “_listing” suffix, which contain information for each website in the Dmoz directory (category id, category path, website URL, website description, and website name). In this folder, you will also find two log files (with a “_generation_log” suffix). These log files contain useful statistics for the “_category” and “_listing” files.

Inside the main folder there is also a sub folder called “world.” This folder contains “_category” and “_listing” files for various non-English languages (in the “world” category).

How to Open

Note: Before opening the files, they must be unzipped! Refer to your zip program for information on how to do this.

Note about the unzip tool included with recent versions of Microsoft Windows: Microsoft’s unzip tool does not support certain non-English characters that some file names contain in the “world” folder. When attempting to unzip the main

folder, some files in the “world” folder will have parts of their names modified to incorrect characters. To avoid this, it is recommended to unzip all files (and the main folder itself) with an alternative zip program such as PeaZip (<http://peazip.sourceforge.net>).

Quick Open Method (if you don’t have a spreadsheet tool like Microsoft Excel):

1. Download and install LibreOffice.
2. (<http://www.libreoffice.org/download>).

During install, if you selected to associate Microsoft Office spreadsheets with LibreOffice, you can double-click to open the desired CSV file (the files suffixed with “_category” or “_listing”), then skip to step 8.

3. Open LibreOffice.
4. On the home screen, select “Spreadsheet.”
5. Select “Open” from the “File” menu. The “Open” dialog appears.
6. On the “Open” dialog, select “Spreadsheets” as the file filter option.
7. Open the desired CSV file (the files suffixed with “_category” or “_listing”). The “Text Import” dialog appears.
8. On the “Text Import” dialog, under the “Import” group, set “From row” to “1” (or “2” if you do not wish to include the header in the import).
9. Under the “Separator options” group, select “Separated by.”
10. In the same group, make sure only “Comma” is ticked.
11. In the same group, set “Text delimiter” to a double quote (“).
12. Under the “Other options” group, make sure only “Quoted field as text” is ticked.
13. Select “OK” to import the data into a new spreadsheet.

Other Open Methods:

- The files can be opened using Microsoft Access or Excel (refer to Microsoft Office’s documentation for more information).

- The files can be opened using a text editor (this is not recommended for some of the larger files).
- The files can be imported into your database (importing process varies by database; please refer to your database import tool for instructions).
- Remember: the files are comma delimited, so be sure to check the settings during import (refer to your application's instructions on how to open or import CSV files).

About the File Headers

All of the provided Dmoz directory data files contain headers that describe the data in the columns below them. If you wish to remove a header, you can open the desired file and delete it manually.

Support

Support and inquiries may be directed to the email address:

ezmozfiles@gmail.com.