

# Extreme DMOZ Extractor Users Guide

Extreme DMOZ Extractor Users Guide written by Bruce J Peresky © 2008 All Rights Reserved.

No portion of this document may be duplicated without written consent of the author.

Extreme DMOZ Extractor © 2007 NiceCoder, All Rights Reserved

Additional guides for NiceCoder products can be found at [IndexUSupport.com](http://IndexUSupport.com)

[Nicecoder Forums](#) [Nicecoder Support](#)

[Part 1 - Installing the Extreme DMOZ Extractor](#)

[Part 1.5 - Installation issues with the Extreme DMOZ Extractor](#)

[Part 2 - Download DMOZ data](#)

[Part 3 - A Quick Glance at the Extreme DMOZ Extractor](#)

[Part 4 - Extracting Categories](#)

[Part 5 - Extracting Links](#)

[Part 6 - Import into IndexU](#)

[Part 7 - Completing your import](#)

[Part 8 - Special Note](#)


## Part 1 - Installing the Extreme DMOZ Extractor

The first thing you need to do is to get your license key and copy of the DMOZ Extractor from the Nicecoder Client Area. Log into the client area and you will see something similar to the screen below.


The screenshot shows the NiceCoder client area interface. At the top, there is a navigation bar with links for 'Clients Login', 'Helpdesk', and 'Contact Us'. Below this is the NiceCoder logo with the tagline 'MAKE YOUR WEBSITE WORK' and a McAfee 'TESTED' badge. A main menu is visible on the left, and a 'Your Orders' table is displayed on the right. The table lists several orders, with the 'Extreme DMOZ Extractor' order (Order ID #2584, Cart ID #86cb6437) highlighted in blue.

Order ID	Cart ID	Ordered	Create
#5566	#6338e1ab	IndexU Deluxe Unlimited Domain License (from IndexU 5)	05-07-2008
#5464	#92b89ee5	IndexU Unlimited Domain License (Extras Only)	04-26-2008
#5371	#76d5c932	iDesk 10 Domain License	04-07-2008
#4214	#4414e5bd	IndexU Free Version	09-17-2007
#2586	#bd2eba5f	PRLoader Single Domain License	11-29-2007
#2585	#67f0f70f	iDesk Single Domain License	11-29-2007
<b>#2584</b>	<b>#86cb6437</b>	<b>Extreme DMOZ Extractor</b>	<b>11-29-2007</b>
#2583	#8af6b01c	Indexu Unlimited Domain License	11-29-2007

Click on the View button beside Extreme DMOZ Extractor.



[Clients Login](#) | [Helpdesk](#) | [C](#)



Home
Products
Community
Testimonials
Purchase

**Main Menu**

- ➔ [Home](#)
- ➔ [Logout Now](#)
- ➔ [Place a New Order](#)
- ➔ [News & Information](#)
- ➔ [Edit Your Profile](#)
- ➔ [View Your Orders](#)
- ➔ [View Your Invoices](#)
- ➔ [View Your Licenses](#)
- ➔ [iDev Affiliate](#)

**Your Orders > Order #2584**

Cart ID:	#86cb6437
Last Invoice ID:	#2584 [ <a href="#">view invoice</a> ] [created on 11-
Product Ordered:	Extreme DMOZ Extractor
Previously Purchased Extras:	None purchased.
Order Status:	Active
Base Product Cost:	\$65 USD One Time

**Your Licenses:**

Extreme DMOZ Extractor

▶ Reissued key: XXXXXXXXXX [Vi](#)

**Upgrade Packages Ordered:**

One year free update

▶ *Upgrade Package Term: 1 Year*

---

One year free update

▶ *Upgrade Package Term: 1 Year*


**Support Packages Ordered:**

No support packages found for this order.

**Addons Ordered:**

No addons found for this order.

Now click on View & Download and you will see the page below.



**NICECODER**  
MAKE YOUR WEBSITE WORK

[Clients Login](#) | [Helpdesk](#) | [Contact Us](#)

---

Home
Products
Community
Testimonials
Purchase

Main Menu

- ➔ [Home](#)
- ➔ [Logout Now](#)
- ➔ [Place a New Order](#)
- ➔ [News & Information](#)
- ➔ [Edit Your Profile](#)
- ➔ [View Your Orders](#)
- ➔ [View Your Invoices](#)
- ➔ [View Your Licenses](#)
- ➔ [iDev Affiliate](#)

**Your License > Order #2584 > Viewing License Details**

Notice:

Your new registered details will be locked in on your first access

License Created on:	11-29-2006 04:46:00
License Activated on:	11-29-2006 04:46:00
License Expires on:	Never Expires
Current License Status:	Reissued
Your License Key:	[REDACTED]

➔ Agreements and Documents

#	Agreement Name
1.	<b>[IMPORTANT!] Extreme Dmoz Extractor License Key</b>
2.	<b>End User License Agreement</b>

➔ Product Downloads Available

#	File Name
1.	Extreme DMOZ Extractor Guide
2.	extreme_dmoz_extractor.exe

➔ Update Packages Purchased

#	Package	Term:
1.	[Expired] <span style="color: green;">[Renew]</span> One year free update	1 Year
2.	[Expired] <span style="color: green;">[Renew]</span> One year free update	1 Year

**Keep this page open when installing the DMOZ Extractor.**

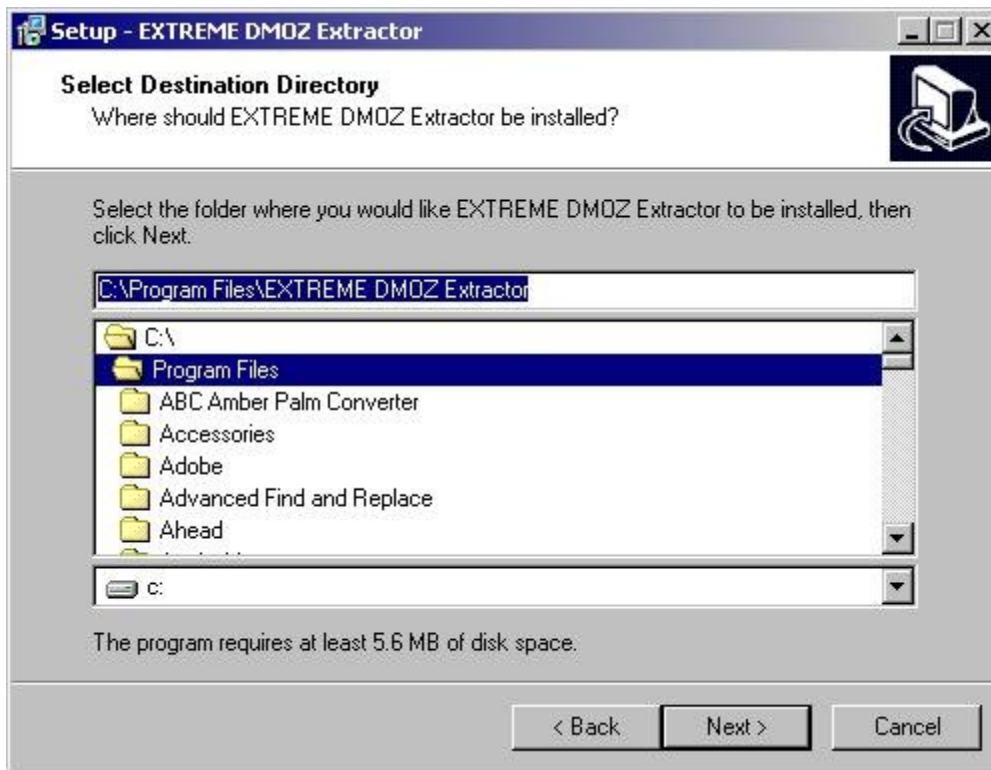
Click on Download beside extreme\_dmoz\_extractor.exe and the file will start to download. Save it to your desktop. Once the download has completed now go to your desktop and double click on extreme\_dmoz\_extractor.exe and you will see the screen below.



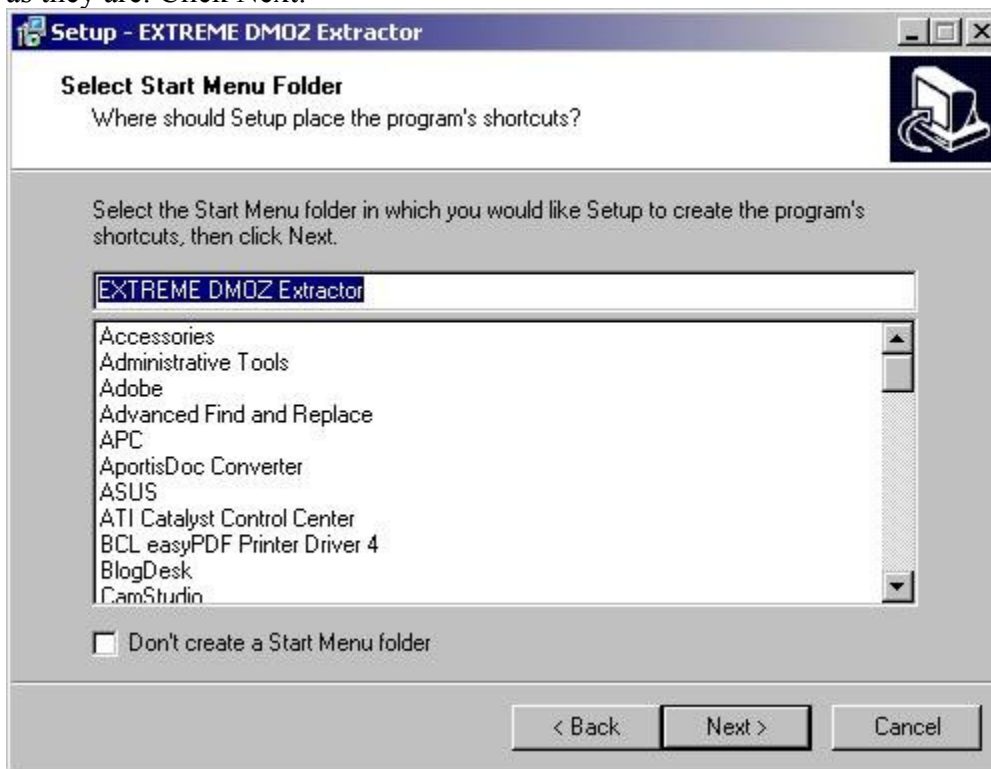
Click Next and you will see the license agreement.



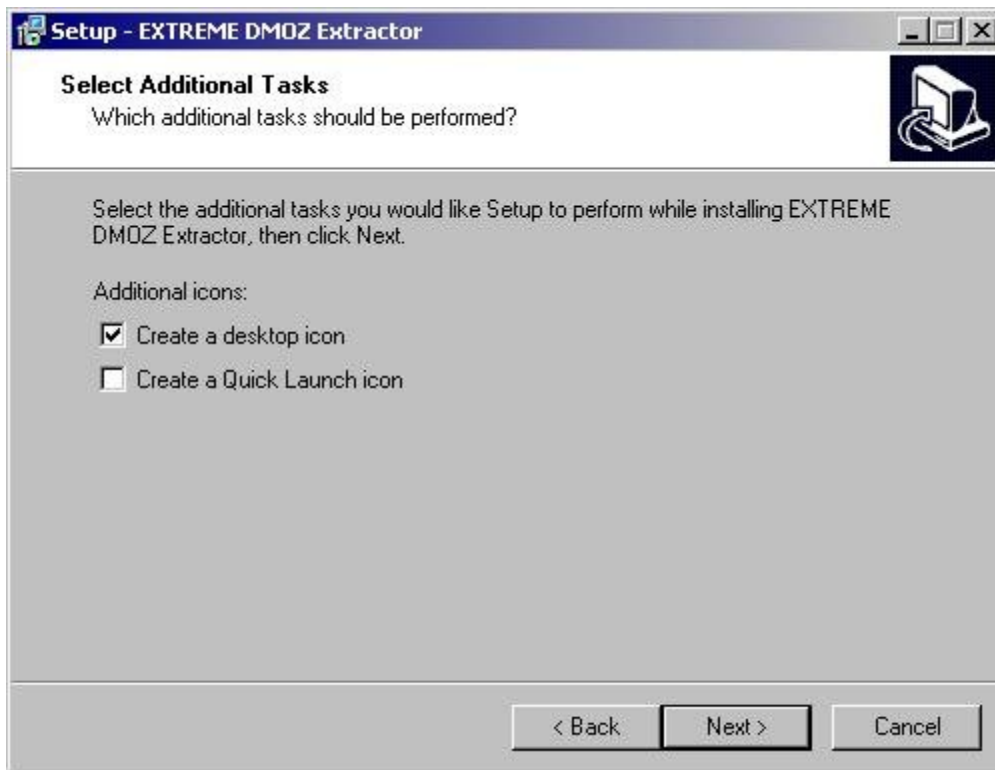
Read the license agreement and if you agree then click on "I accept this agreement" and click on Next and you will see the screen below.



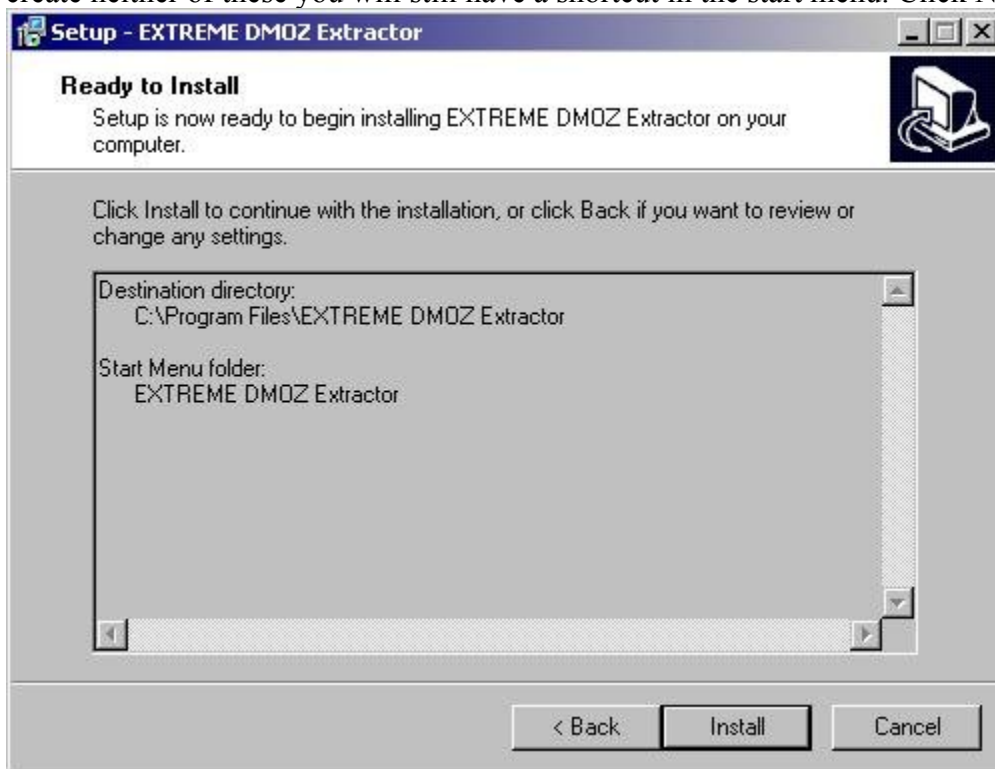
Choose the location to install the DMOZ Extractor to as shown above. It is fine to just leave the settings as they are. Click Next.



Choose the Start Menu folder where you want a shortcut for the DMOZ Extractor to appear. It is fine to just leave the settings as they are. Click Next.



Now you need to choose whether to create desktop and quick launch icons as above. If you choose to create neither of these you will still have a shortcut in the start menu. Click Next.

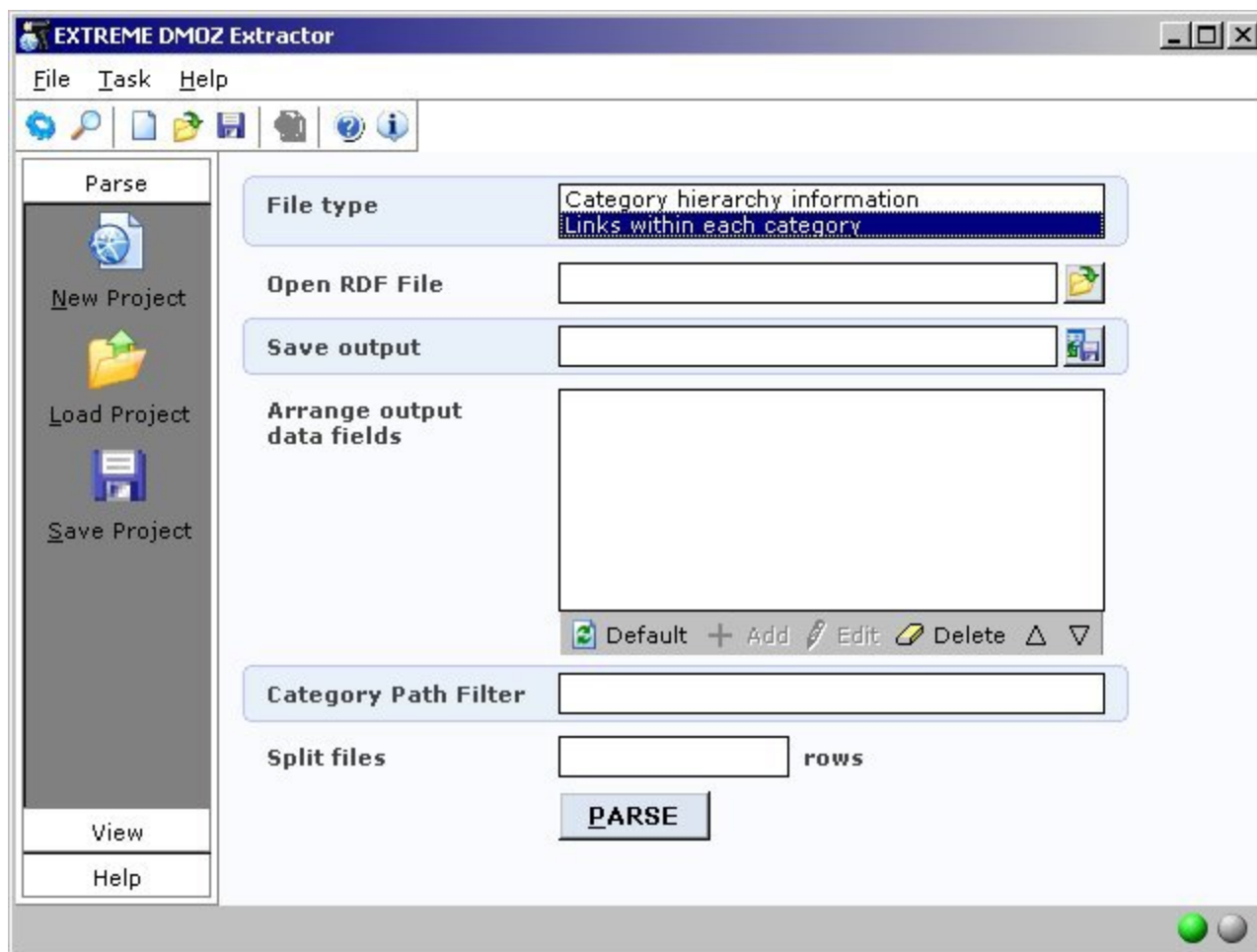


Confirm your settings as per the screenshot above. Click Install and installation will begin.

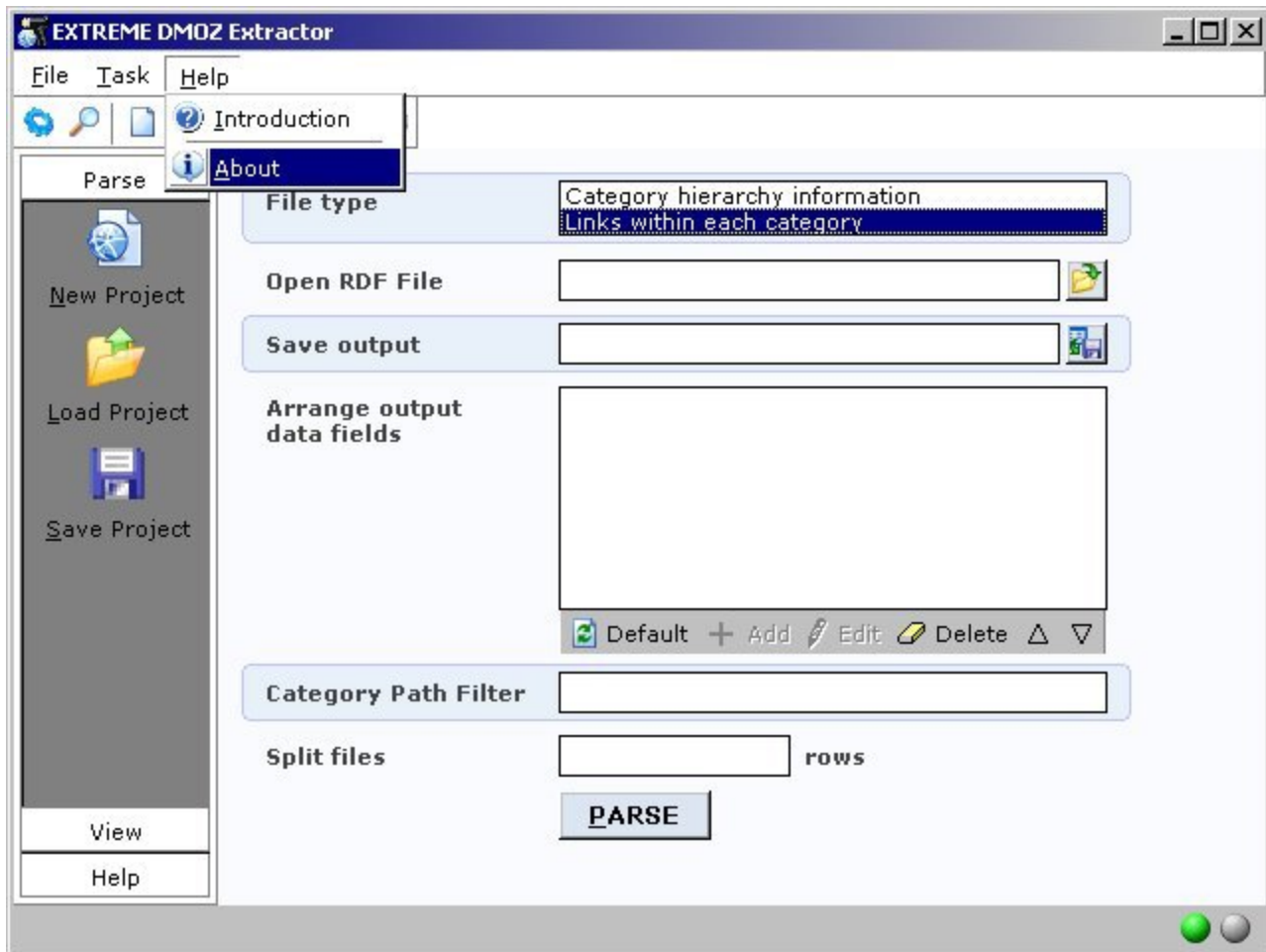


Click finish and the installation is complete

Now you need to register your copy of the DMOZ Extractor  
Start up the DMOZ Extractor



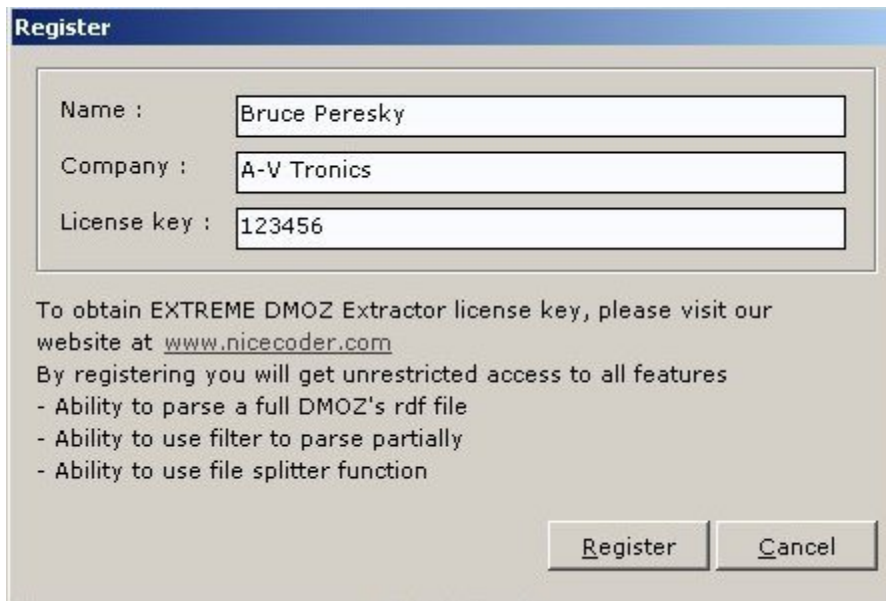
Click on Help then click on About



Now you will see the screen below



Now you can click on the Register button. Enter your name, company name and license number and click on Register



**Register**

Name :

Company :

License key :

To obtain EXTREME DMOZ Extractor license key, please visit our website at [www.nicecoder.com](http://www.nicecoder.com)

By registering you will get unrestricted access to all features

- Ability to parse a full DMOZ's rdf file
- Ability to use filter to parse partially
- Ability to use file splitter function

If all the data is correct you will see the screen below



Click OK, then click Close. You are now ready to use the Extreme DMOZ Extractor

## Part 1 - Installation issues with the Extreme DMOZ Extractor

If you receive the following error during installation

RICHTX32.OCX The existing file is newer than the one Setup is trying to install

Simply click on Yes and installation will continue normally

If you receive the following error during installation

COMDLG32.OCX The existing file is newer than the one Setup is trying to install

Simply click on Yes and installation will continue normally

If you receive the following error during installation

msxml4.dll The existing file is newer than the one Setup is trying to install

Simply click on Yes and installation will continue normally

If you receive the following error

Run time error '429' ActiveX component can't create object

You must download and install the following fix from Microsoft.

[msxmlcab.exe](#)

Choose RUN when you click on the link above and follow the directions to install the fix. Once you have installed the MS XML fix above you can then proceed to install the Extreme DMOZ Extractor normally.

[Return to the top](#)

## Part 2 - Download DMOZ data

Visit <http://rdf.dmoz.org/> and download the following files

structure.rdf.u8.gz (~65MB Uncompressed)

content.rdf.u8.gz (~300MB Uncompressed)

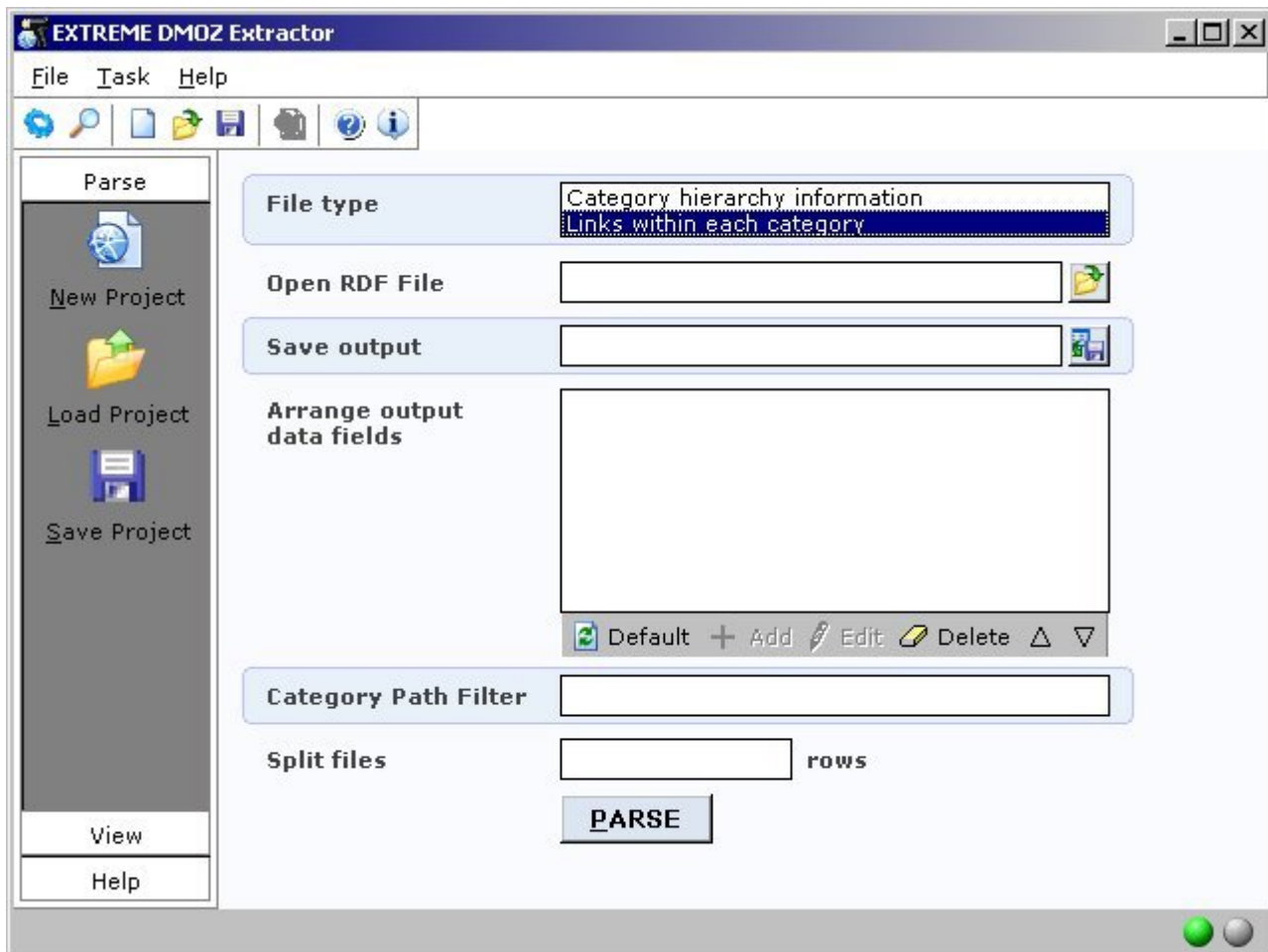
structure.rdf.u8 will be approx 650MB when uncompressed and contains the categories

content.rdf.u8 will be approx 1.9GB when uncompressed and contains the links

**NOTE** - Failure to extract the files will create a large number of errors.

[Return to the top](#)

## Part 3 - A Quick Glance at the Extreme DMOZ Extractor



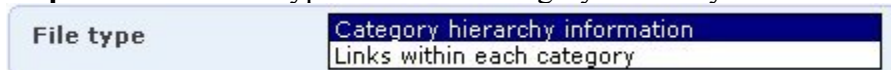
[Return to the top](#)

## Part 4 - Extracting Categories

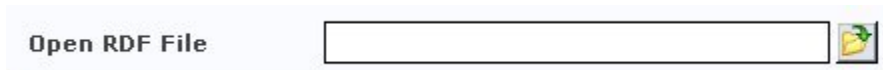
### Extract Categories

**Step 1** - Start the DMOZ Extractor, it can be found at Start - Programs - Extreme DMOZ Extractor

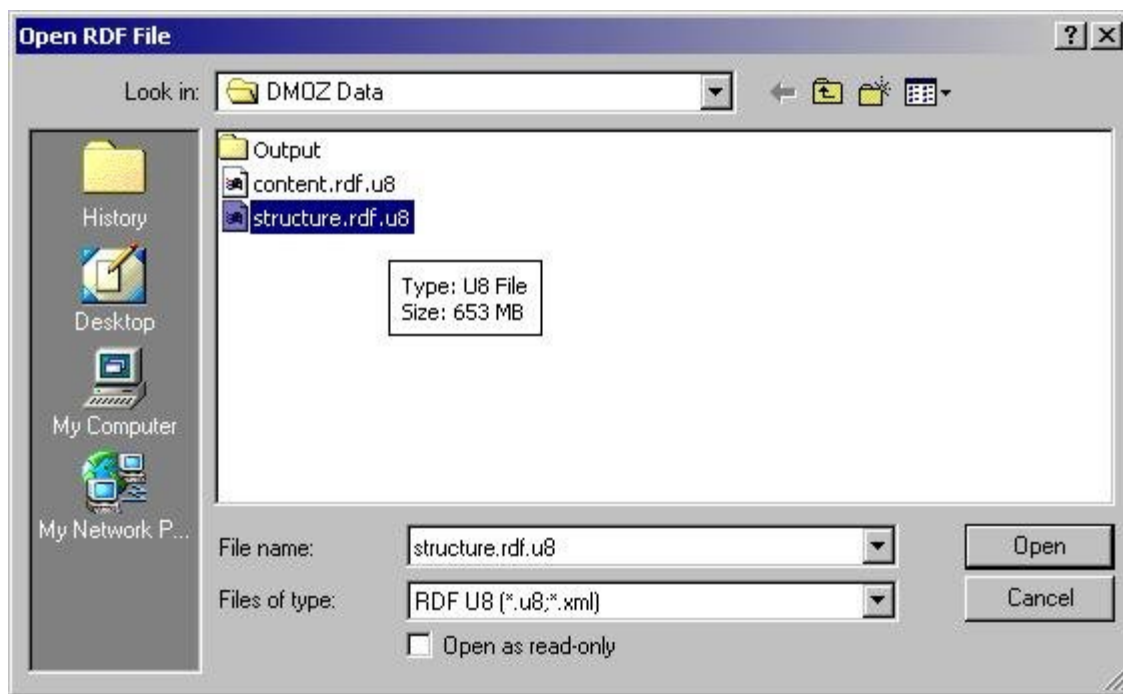
**Step 2** - Beside File Type - click on "category heirarchy information" as shown below



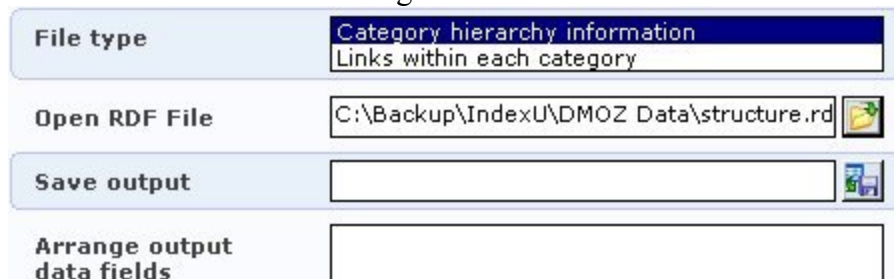
**Step 3** - Beside Open RDF File you see an entry box and then a folder with an arrow on it as shown below.



Click on the folder and browse to the location where you extracted the structure.rdf.u8 and content.rdf.u8. Highlight the file structure.rdf.u8 and click on Open as shown below.



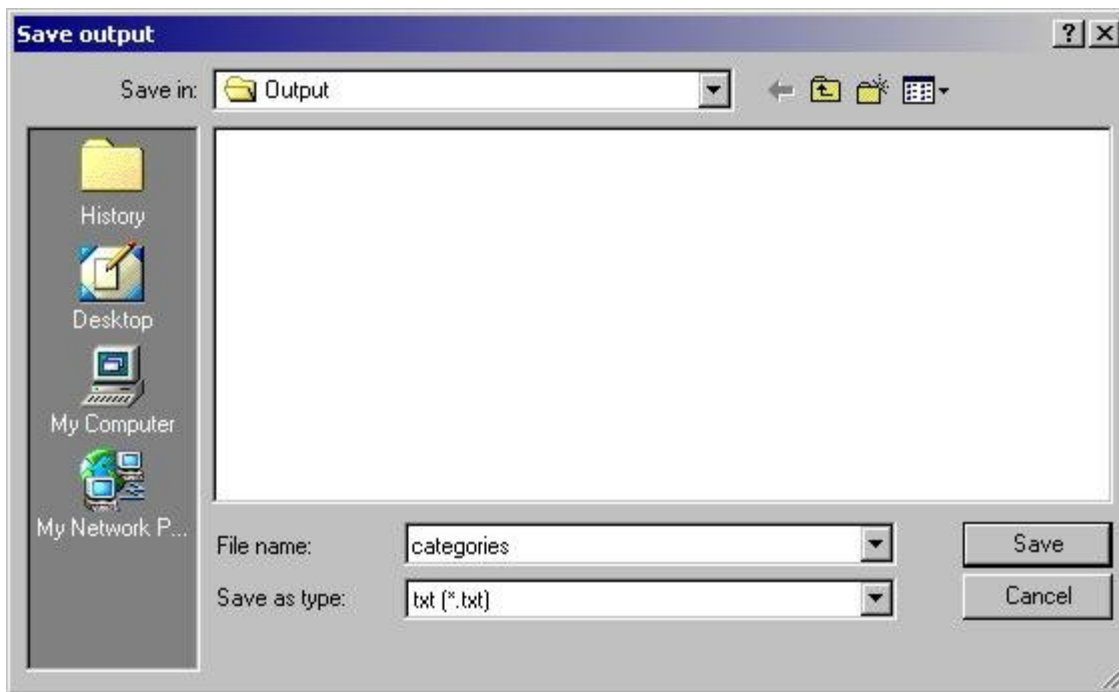
You will now see the following



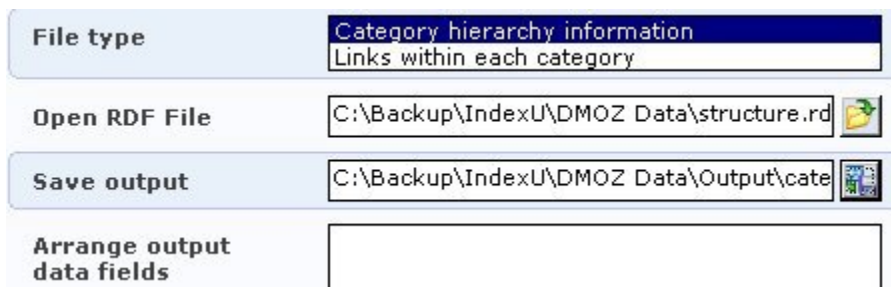
**Step 4** - Beside Save output - you see an entry box and then a small box with a disk on it as shown below.






Click on the button and browse to the location where you want to save the output that the DMOZ Extractor makes. Once you have done that you will need to enter in a file name. The screen below shows you what you will see.



I suggest you make the filename descriptive and also use the word category or cat so you know this is the category output. Now click on Save and you will see the following.



**Step 5** - Beside Arrange Output Data Fields you will see a box with some buttons at the bottom. Click on the Default button and you will see the screen below

<b>File type</b>	Category hierarchy information Links within each category						
<b>Open RDF File</b>	C:\Backup\IndexU\DMOZ Data\structure.rd 						
<b>Save output</b>	C:\Backup\IndexU\DMOZ Data\Output\cate 						
<b>Arrange output data fields</b>	<table border="1"> <tr><td>category_id</td><td></td></tr> <tr><td>name</td><td></td></tr> <tr><td>category_path</td><td></td></tr> </table> <p>Default + Add Edit Delete </p>	category_id		name		category_path	
category_id							
name							
category_path							
<b>Set a root</b>	<input type="text"/>						
<b>Split files</b>	<input type="text"/> rows						
<b>PARSE</b>							

**Step 6** - Beside Set A Root is a text entry box as shown below.

<b>Set a root</b>	<input type="text"/>
-------------------	----------------------

You must enter your "Root" here. The Root is the same as the URL for DMOZ except that you replace <http://www.dmoz.org> with Top/

Examples

<http://www.dmoz.org/Business/> becomes Top/Business/

<http://www.dmoz.org/Recreation/Humor/> becomes Top/Recreation/Humor/

[http://www.dmoz.org/Kids\\_and\\_Teens/](http://www.dmoz.org/Kids_and_Teens/) becomes Top/Kids\_and\_Teens

[http://www.dmoz.org/Games/Video\\_Games/](http://www.dmoz.org/Games/Video_Games/) becomes Top/Games/Video\_Games/

**NOTE** - No spaces are allowed in any circumstances! The DMOZ categories always have an underscore instead of a space and it works the same way in the DMOZ Extractor. Even if you do not see the space on the DMOZ site, it really is there!

**NOTE** - The Root MUST start with Top/ and must end with /

**NOTE** - You must capitolize the categories and the word Top as shown above. Failure to do so may result in no results being returned.

Lets assume we want to extract categories from <http://www.dmoz.org/Regional/Asia/Japan/> we would change the URL to Top/Regional/Asia/Japan/ and enter this in the "Set A Root" box as shown below

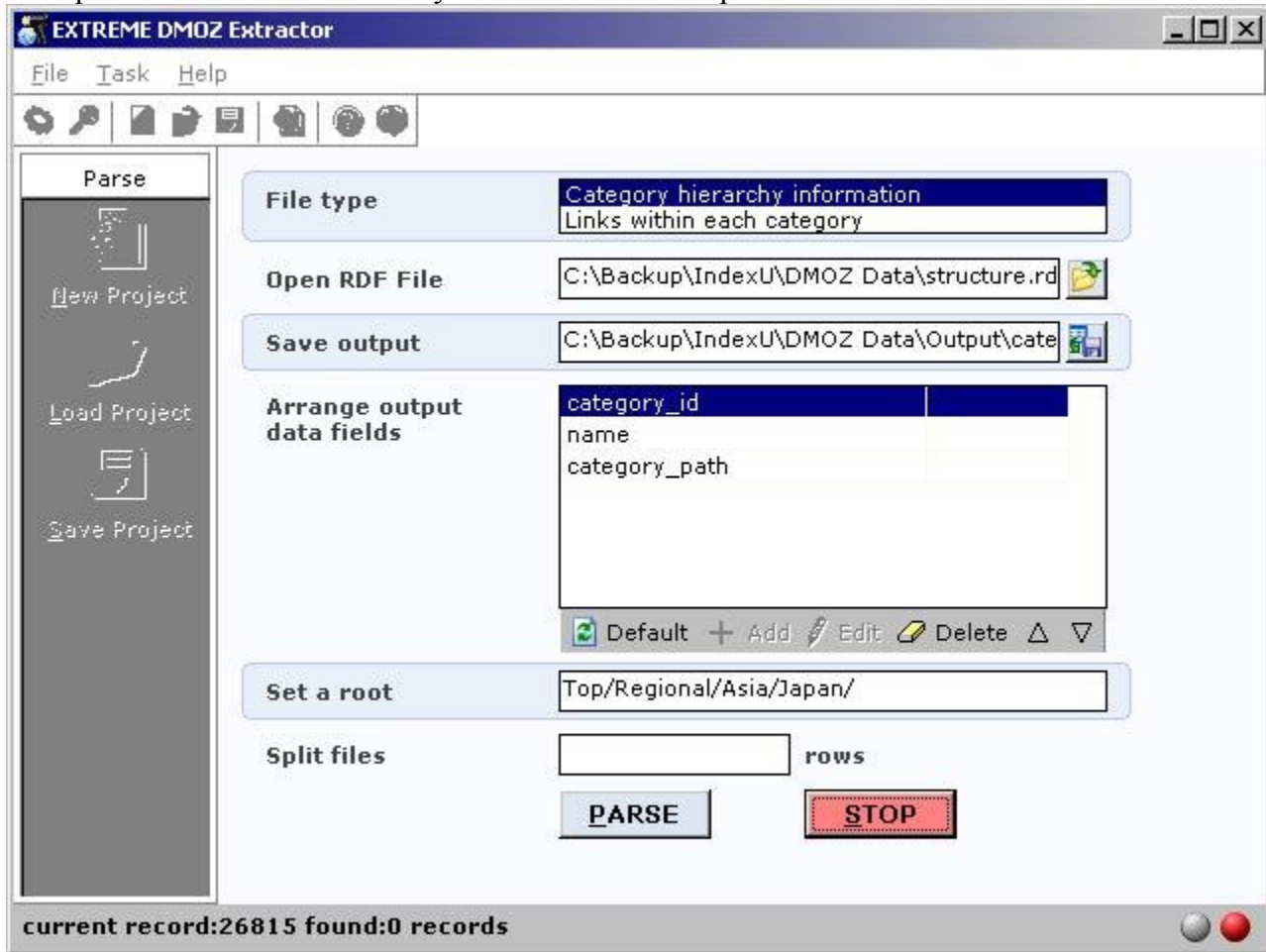
<b>Set a root</b>	Top/Regional/Asia/Japan/
-------------------	--------------------------

**Step 7** - Beside Split Files is another text entry box as shown below.

<b>Split files</b>	<input type="text"/> rows
--------------------	---------------------------

If you do not know the size of files you are allowed to upload to your webserver can be then enter the value **10000** here. Otherwise adjust accordingly or leave this entry blank for users who have set their php.ini to allow any file size to be uploaded.

**Step 8** - Click on Parse and the Extreme DMOZ Extractor starts getting to work. You can click on Stop to stop the DMOZ Extractor at any time before it is complete.



Extracting Categories is now complete. A screenshot of what is displayed is below.

**Summary**

## **SUMMARY**

<b>File Type</b>	Category hierarchy information
<b>RDF File</b>	C:\Backup\IndexU\DMOZ Data\structure.rdf.u8
<b>Output</b>	C:\Backup\IndexU\DMOZ Data\Output\categories.txt
<b>Filter</b>	Top/Regional/Asia/Japan/
<b>Split File</b>	No
<b>Created File</b>	1
<b>Total Record</b>	741993
<b>Record Found</b>	816
<b>Total Time</b>	00 hour 02 minute 41 seconds
<b>Status</b>	<b>Done</b>

You can see that the Extreme DMOZ Extractor extracted 816 records. Since we are extracting categories that means we now have 816 categories. Click on Close to complete this part of the process.

[Return to the top](#)

## Part 5 - Extracting Links


**Step 1** - Beside File Type - click on "Links Within Each Category" as shown below.

**File type**

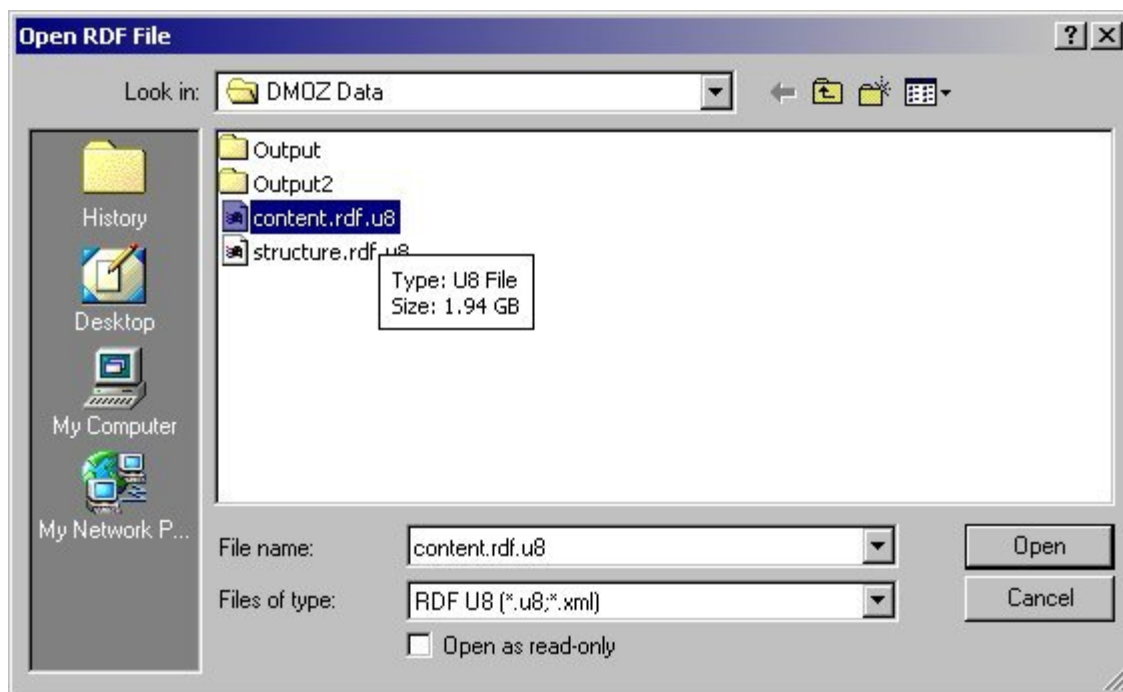
Category hierarchy information  
Links within each category

**Step 2** - Beside Open RDF File you see an entry box and then a folder with an arrow on it as shown below.

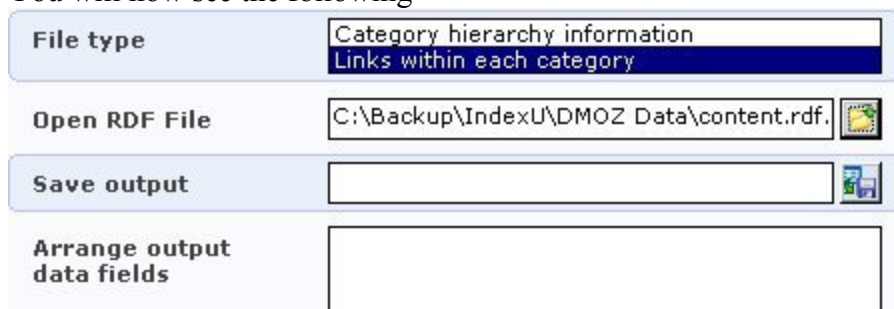
**Open RDF File**



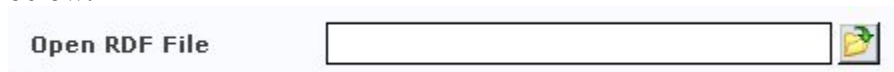
Click on the folder and browse to the location where you extracted the structure.rdf.u8 and content.rdf.u8. Highlight the file content.rdf.u8 and click on Open as shown below.



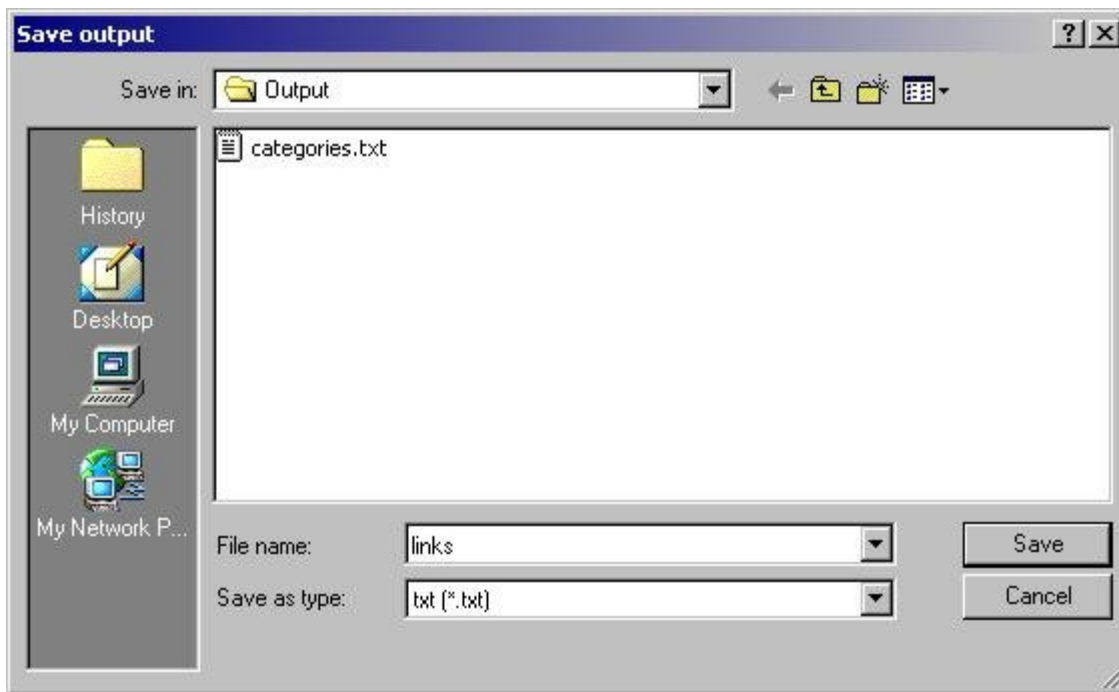
You will now see the following



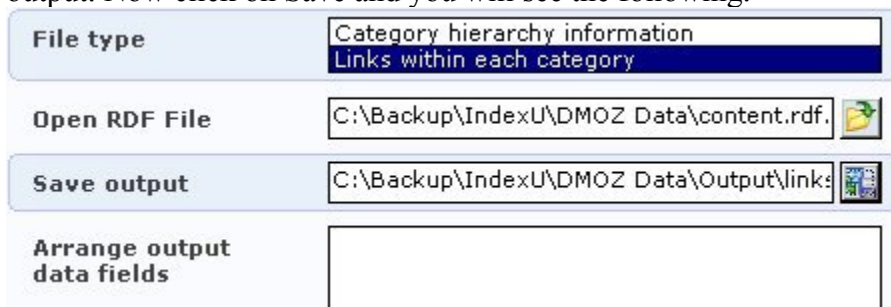
**Step 3** - Beside Save output - you see an entry box and then a small box with a disk on it as shown below.



Click on the button and browse to the location where you want to save the output that the DMOZ Extractor makes. Once you have done that you will need to enter in a file name. The screen below shows you what you will see.



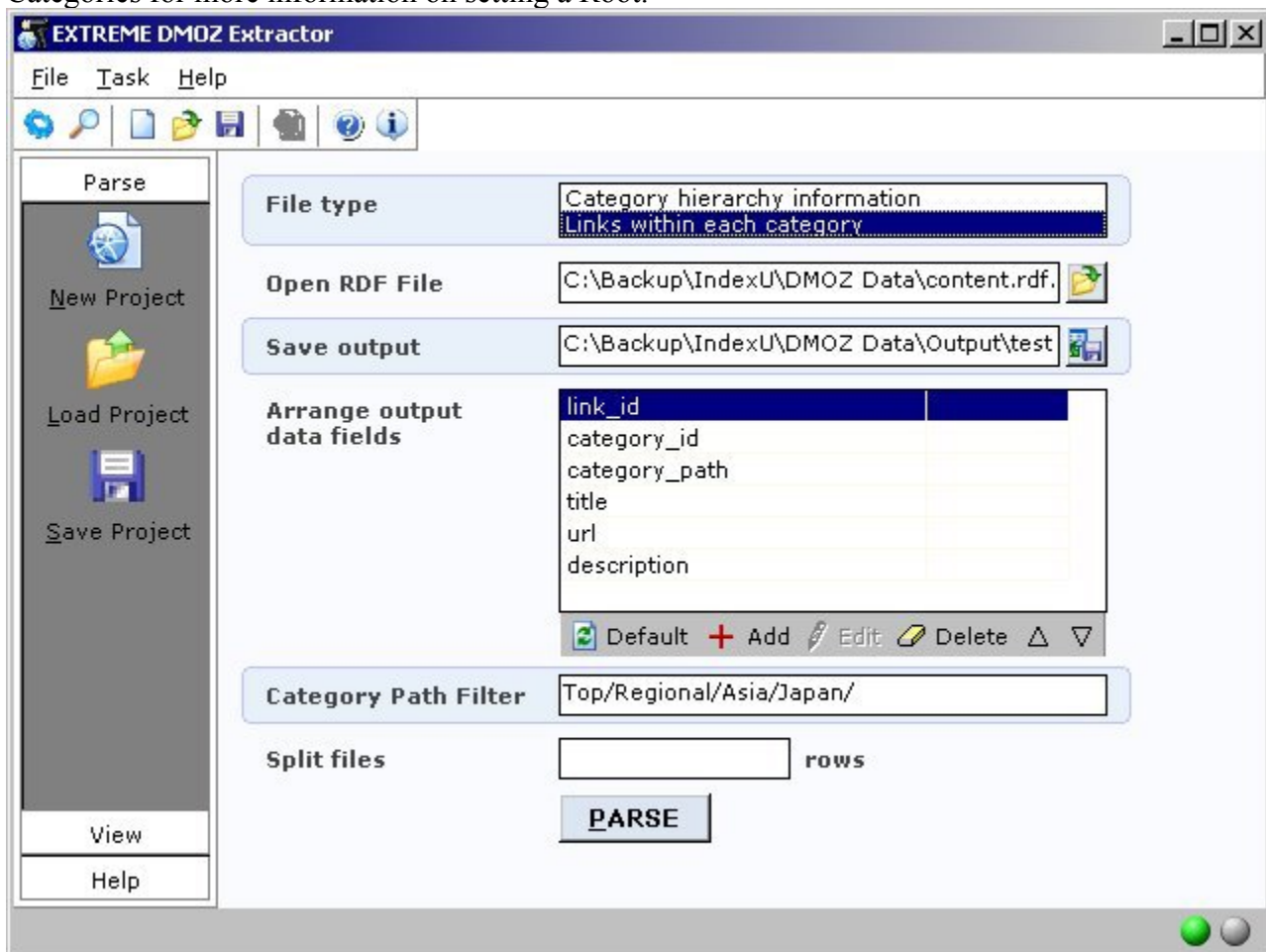
I suggest you make the filename descriptive and also use the word links so you know this is the link output. Now click on Save and you will see the following.



**Step 4** - Beside Arrange Output Data Fields you will see a box with some buttons at the bottom. Since you already did the categories the Extreme DMOZ Extractor will still show the data fields for categories. Clicking on Default will change this information to the data needed for links. If there is no data here click on Default anyways.

<b>File type</b>	Category hierarchy information Links within each category												
<b>Open RDF File</b>	C:\Backup\IndexU\DMOZ Data\content.rdf.												
<b>Save output</b>	C:\Backup\IndexU\DMOZ Data\Output\test												
<b>Arrange output data fields</b>	<table border="1"> <tr><td>link_id</td><td></td></tr> <tr><td>category_id</td><td></td></tr> <tr><td>category_path</td><td></td></tr> <tr><td>title</td><td></td></tr> <tr><td>url</td><td></td></tr> <tr><td>description</td><td></td></tr> </table> <p>Default + Add Edit Delete</p>	link_id		category_id		category_path		title		url		description	
link_id													
category_id													
category_path													
title													
url													
description													
<b>Category Path Filter</b>	Top/Regional/Asia/Japan/												
<b>Split files</b>	<input type="text"/> rows												

**Step 5** - Set a root - Since you just extracted categories your root will already be set. See Extracting Categories for more information on setting a Root.

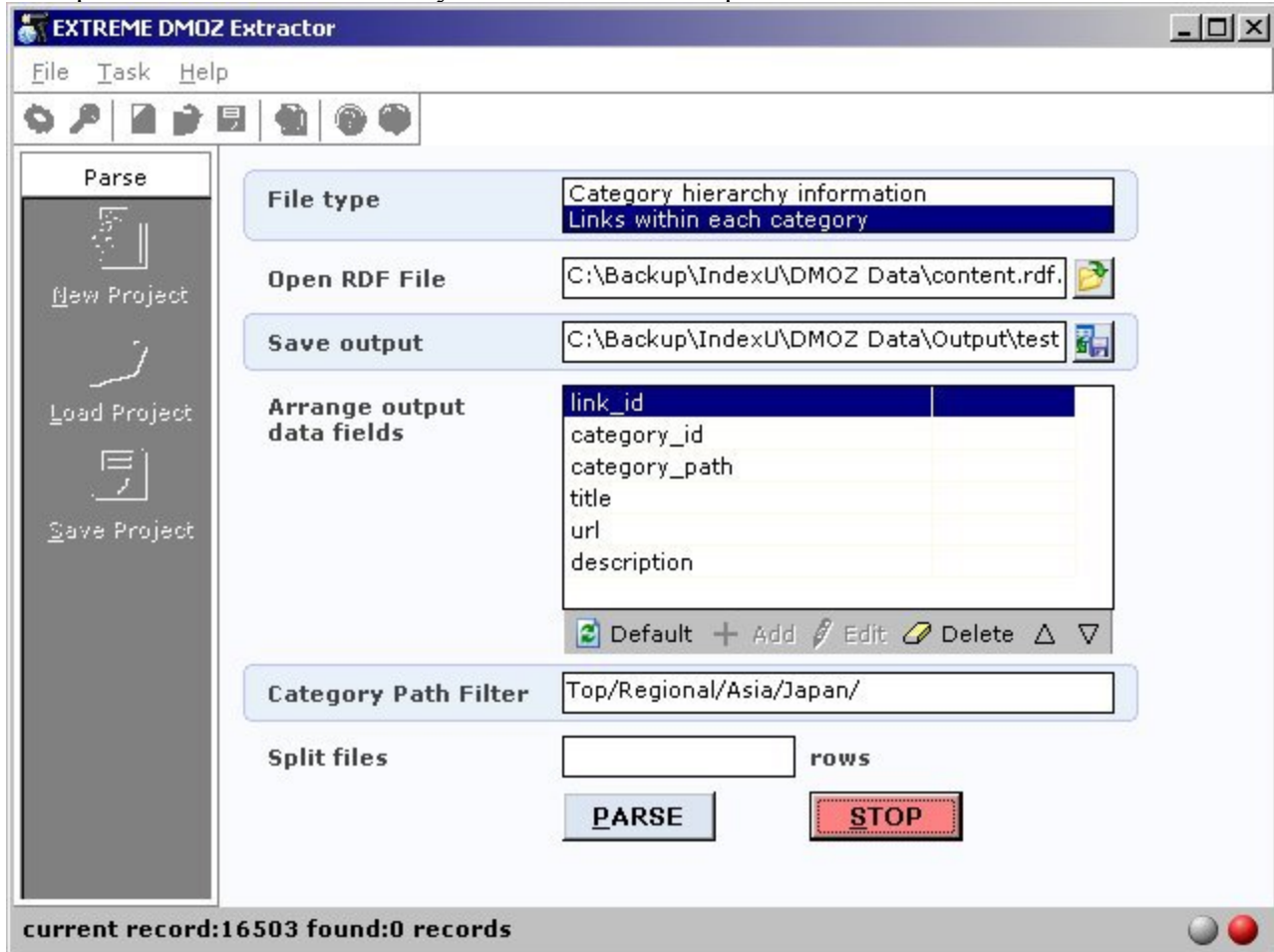


**Step 7** - Beside Split Files is another text entry box as shown below.

<b>Split files</b>	<input type="text"/>	rows
--------------------	----------------------	------

If you do not know the size of files you are allowed to upload to your webserver can be then enter the value **2000** here. Otherwise adjust accordingly or leave this entry blank for users who have set their php.ini to allow any file size to be uploaded.

**Step 7** - Click on Parse and the Extreme DMOZ Extractor starts getting to work. You can click on Stop to stop the DMOZ Extractor at any time before it is complete.



Extracting Categories is now complete. A screenshot of what is displayed is below. Click on Close to complete your extraction.

### Summary

## **SUMMARY**

<b>File Type</b>	Category hierarchy information
<b>RDF File</b>	C:\Backup\IndexU\DMOZ Data\structure.rdf.u8
<b>Output</b>	C:\Backup\IndexU\DMOZ Data\Output\categories.txt
<b>Filter</b>	Top/Regional/Asia/Japan/
<b>Split File</b>	No
<b>Created File</b>	1
<b>Total Record</b>	741993
<b>Record Found</b>	816
<b>Total Time</b>	00 hour 02 minute 41 seconds
<b>Status</b>	<b>Done</b>

You can see that the Extreme DMOZ Extractor extracted 6292 records. Since we are extracting links that means we now have 6292 links. Click on Close to complete this part of the process.

Done, now import your data into your admin panel of IndexU

[Return to the top](#)

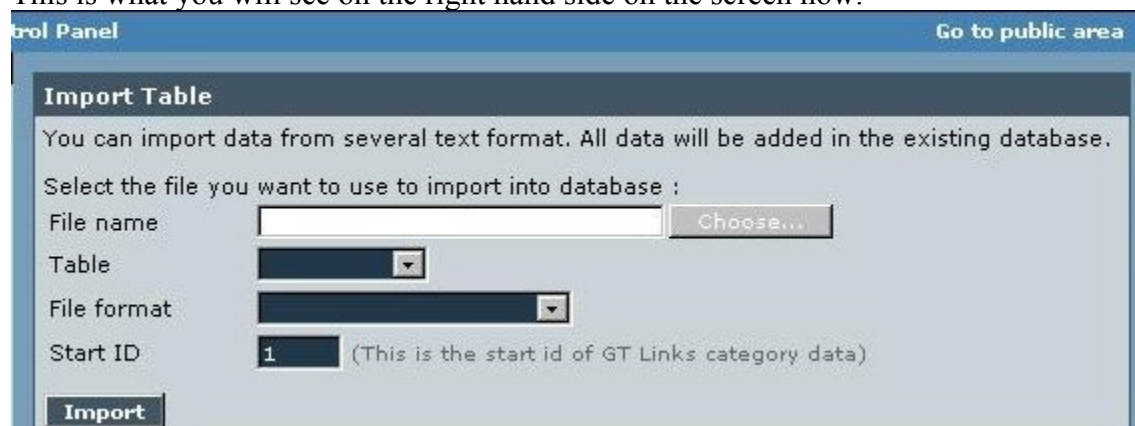
## Part 6 - Import into IndexU

To import your new DMOZ links and cateogires go the the Admin panel of your IndexU installation and at the bottom of the left hand menu go to Database and then to Import. A screenshot is provided below

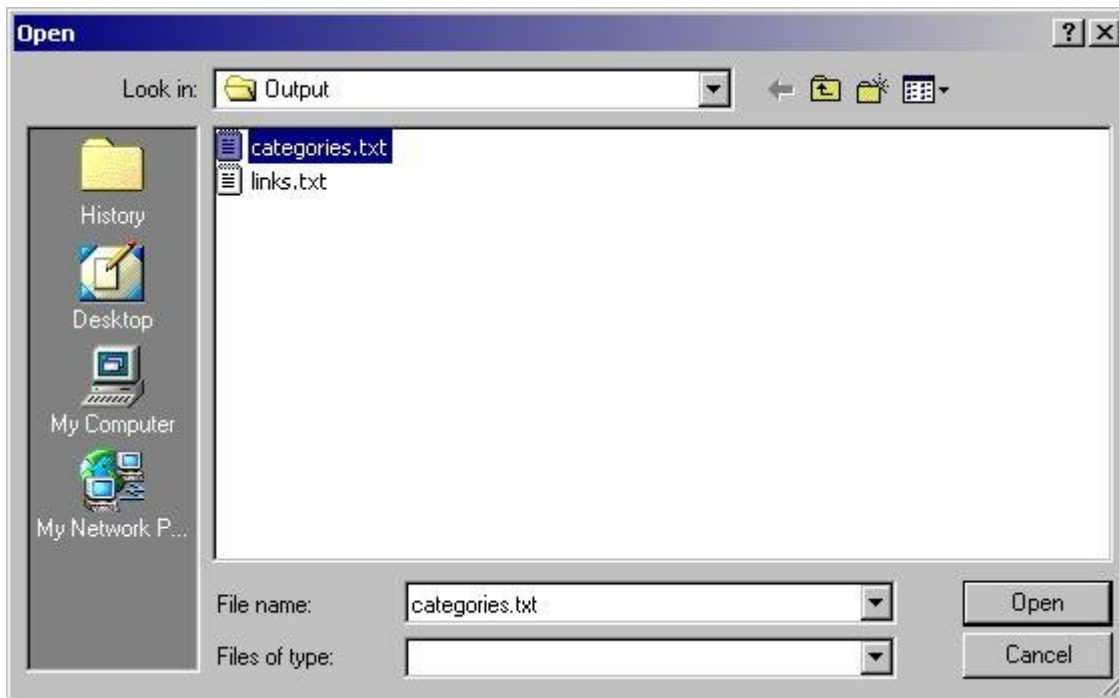


## Import Categories

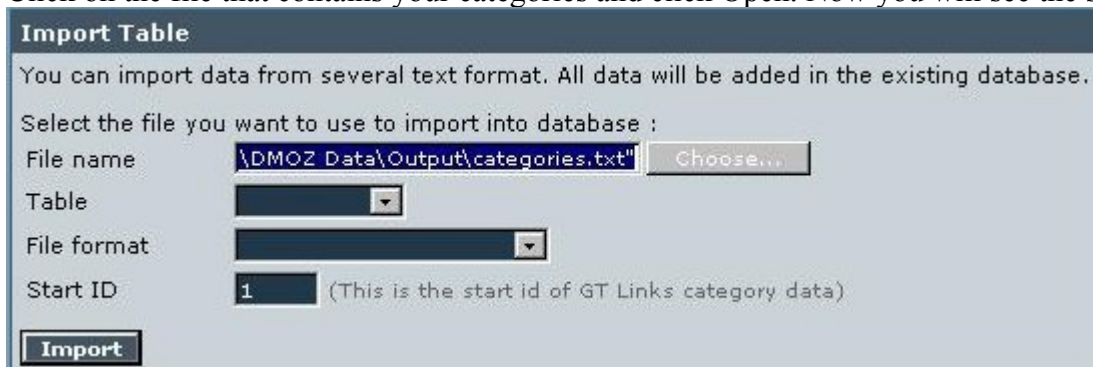
This is what you will see on the right hand side on the screen now.



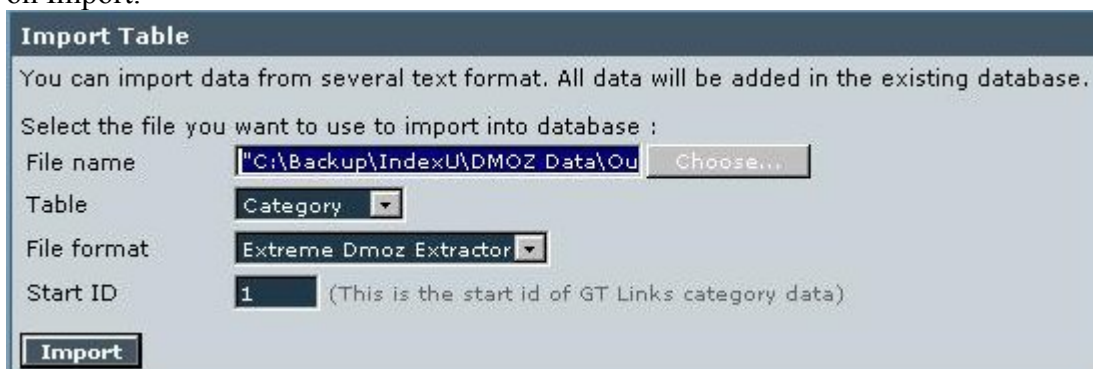
Click Choose, browse to the folder that you specified earlier where you saved the output.



Click on the file that contains your categories and click Open. Now you will see the screen below.



Now you need to select the data for the rest of the fields as shown below. Once that data is entered click on Import.

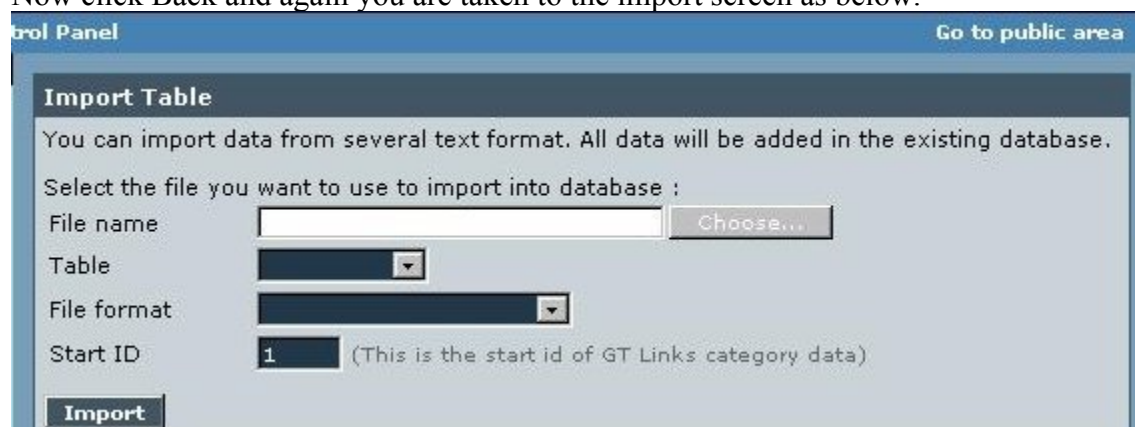


After a bit of time you will see the following screen. This means you have successfully imported your database!

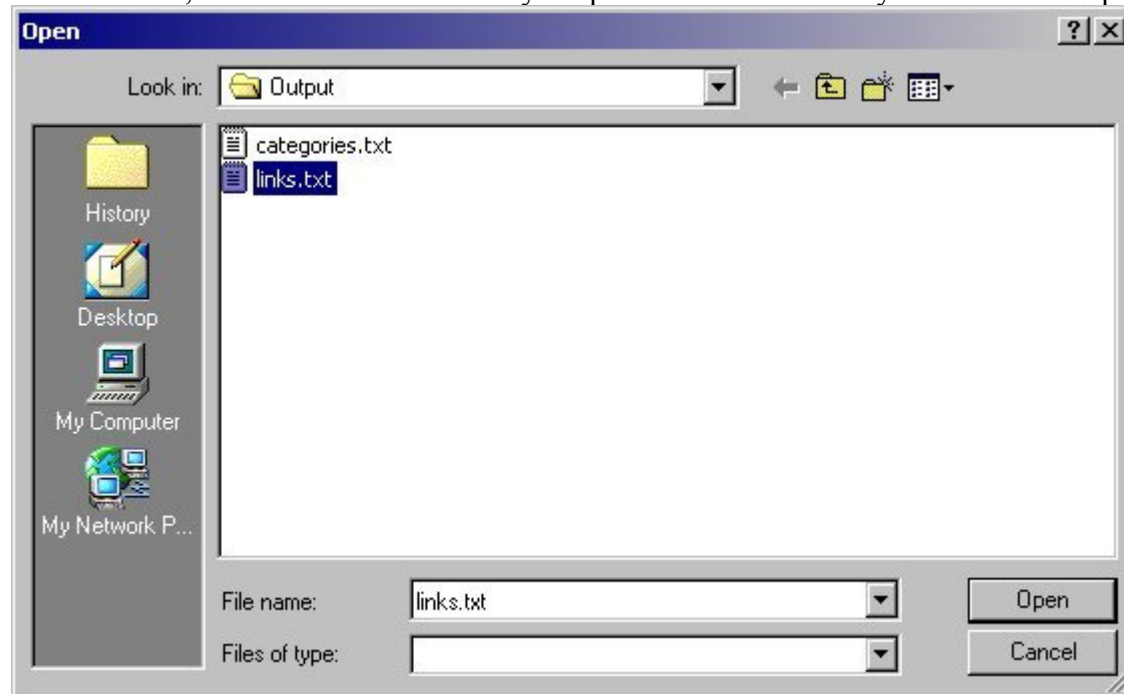


## Import Links

Now click Back and again you are taken to the import screen as below.



Click Choose, browse to the folder that you specified earlier where you saved the output.



Click on the file that contains your links and click Open. Now you will see the screen below.

**Import Table**

You can import data from several text format. All data will be added in the existing database.

Select the file you want to use to import into database :

File name

Table

File format

Start ID  (This is the start id of GT Links category data)

Now you need to select the data for the rest of the fields as shown below. Once that data is entered click on Import.

**Import Table**

You can import data from several text format. All data will be added in the existing database.

Select the file you want to use to import into database :

File name

Table

File format

Start ID  (This is the start id of GT Links category data)

After a bit of time you will see the following screen. This means you have successfully imported your database!

**Control Panel** [Go to public area](#)

Done

[Back](#)

**NOTE** - If you see the Done message but no categories or links appear in your database it means you uploaded a file that was too large. You will need to split to your categories and/or links in order to upload them. Most web servers have a 2MB size limit that you can upload! This is why the Extreme DMOZ Extractor allows you to split files when extracting the categories and links. You will need to split the file manually or extract them again and use the split feature.

[Return to the top](#)

## Part 7 - Completing your import

At this point do not forget to update your category paths and update your link counts. You are now complete! Enjoy your new data.

**NOTE** - Failure to update category paths or update link counts will cause problems with your directory.

[Return to the top](#)

## Part 8 - Special Note

A special note is required as some users have issues during the import of categories. If you happen to see an error similar to the one below then read the notes below the screenshot.

```
ERROR: Parent not found (Business_and_Economy)
ERROR: Parent not found (Education)
ERROR: Parent not found (Government)
ERROR: Parent not found (Guides_and_Directories)
ERROR: Parent not found (Health)
ERROR: Parent not found (Islands)
ERROR: Parent not found (Localities)
ERROR: Parent not found (Maps_and_Views)
ERROR: Parent not found (News_and_Media)
ERROR: Parent not found (Prefectures)
ERROR: Parent not found (Recreation_and_Sports)
ERROR: Parent not found (Regions)
ERROR: Parent not found (Science_and_Environment)
ERROR: Parent not found (Society_and_Culture)
ERROR: Parent not found (Transportation)
ERROR: Parent not found (Travel_and_Tourism)
ERROR: Parent not found (Weather)

Done
Back
```

This issue is a path issue and is quite easy to fix. In the case of the above example using the Root of Top/Regional/Asia/Japan/ you must open the category output file and remove the path from all category lines.

In this case simply search for "Top/Regional/Asia/Japan/" and replace it with "" (that's correct, replace it with nothing). Once you do that your import will now work properly.

[Return to the top](#)