Charity Scraper

CHARITY ORGANISATION SCRAPER

USER MANUAL

Version 1.0

Charity Scraper

Thank you for purchasing the Charity Organisation Scraper (further in text: COS). This software is meant for getting information out of Canadian Registry of Charity organisation. It has multi-process architecture and it uses Microsoft Access 2007 format database as its storing mechanism.

SYSTEM REQUIRMENTS.

Program is written using Microsoft .NET framework version 4.0, therefore minimum system requirments are as follows:

- Windows OS, at least xp sp3
- .NET framework 4.0
- Microsoft Access Connector (in case the Microsoft Access is not installed on the workstation)

THE FILE STRUCTURE

Program needs multiple files in its folder in order for it to work properly.

CharityScrape.exe.config – holds the connection string to database.

Database11_template.accdb The database template that holds structure for the datastore database.

chromedriver.exe Chrome browser that is used to re-fetch failed records

testMap.xml – Used when fetching/updating the records. Holds directions to where each of the Organisation Detail Field is located in webpages. Also holds default url and directions for csv file download.

testSelMap.xml – Used when re-fetching the records. Holds directions to where each of the Organisation Detail Field is located in webpages. Also holds default url.

OVERALL PROCESS.

When you first launch the application, it will create a file copy of database template and name it Database11.accdb. Further on this copy will be used as a storage mechanism for all the results obtained in the program work process.

Program has four modes of action:

UPDATE DATA

Gathers information from the website, starting with the resultspage number specified. Then saves the results to database.

SYNC WITH CSV

Downloads the csv information page and then updates record's real address information in database based on organisation registration number.

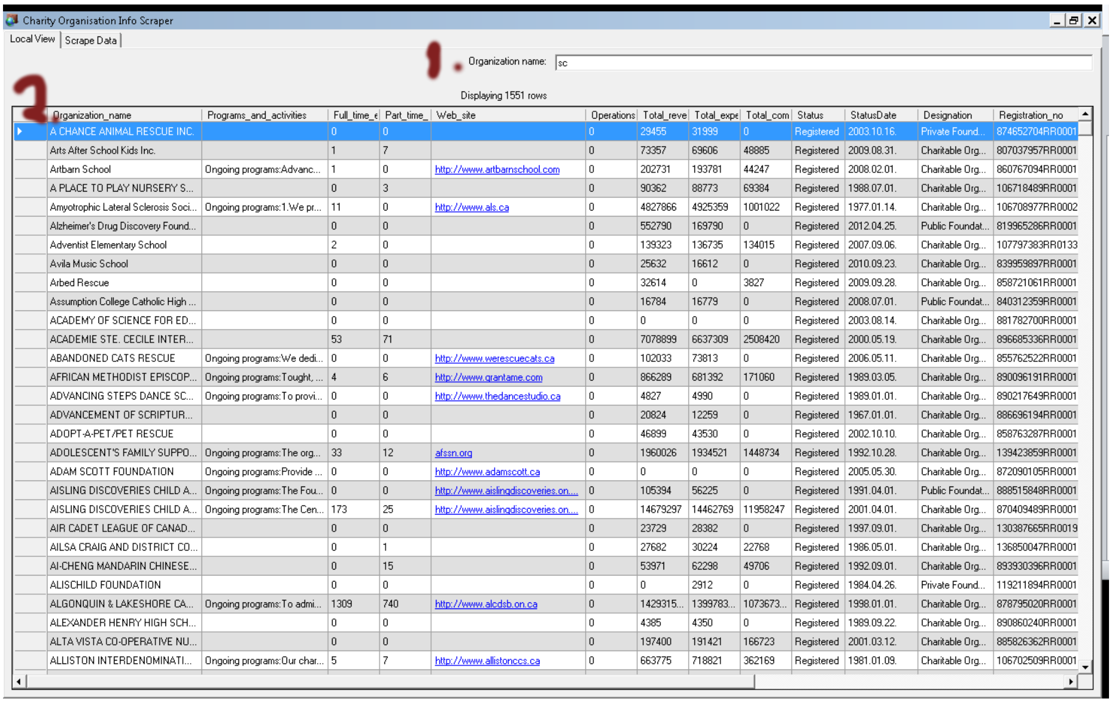
RE-FETCH FAILED ROWS

Sometimes Update data might fail migth fail more than 3 times. In such times program skips the record and continues to next ones. That results in a record inside database that has no Organisation name. This action seeks such records and tries to obtain the information using alternative method – by fully rendering the page using chromedriver and then applying the update data process on it.

Charity Scraper

THE LOCAL VIEW TAB

This section shows an overview of fetched results. It is filterable by organisation name. Clicking on the header of a column will sort the table by this column. Double clicking on any record will open the page from which this information was provided from. Clicking on an website column will open that particular link in a new system default browser window.



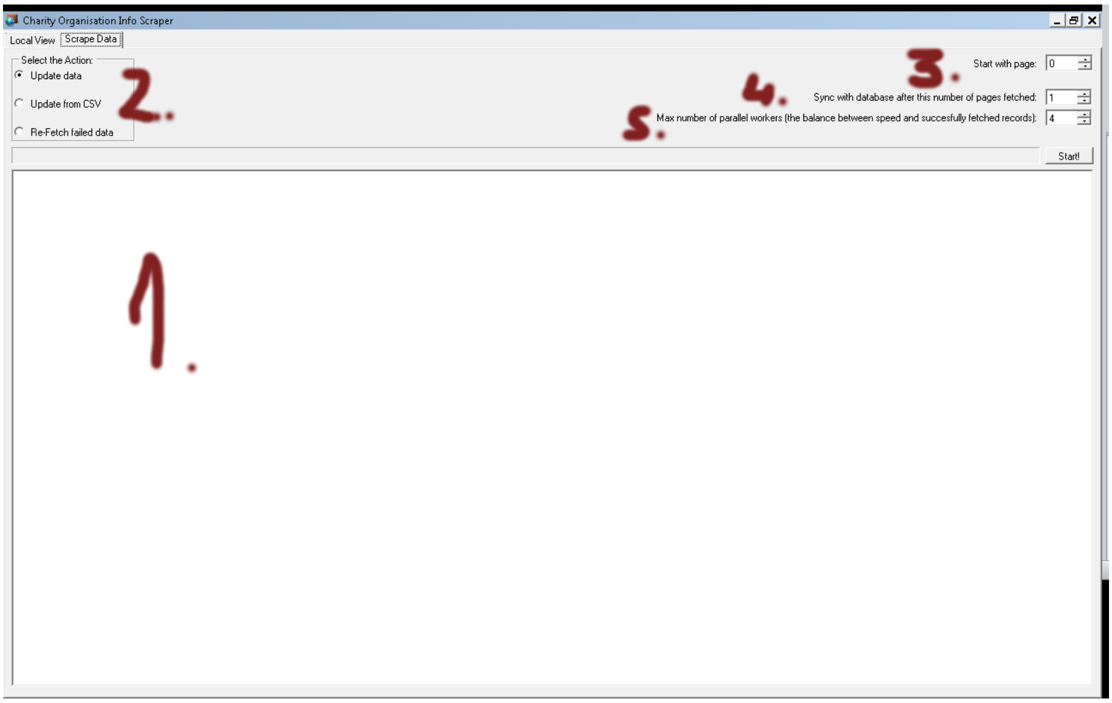
Organization name: sc

Displaying 1551 rows

Organization_name	Programs_and_activities	Full_time_e	Part_time_	Web_site	Operations	Total_reve	Total_expe	Total_com	Status	StatusDate	Designation	Registration_no
A CHANCE ANIMAL RESCUE INC.		0	0		0	29455	31999	0	Registered	2003.10.16.	Private Found...	874652704RR0001
Arts After School Kids Inc.		1	7		0	73357	69606	48885	Registered	2009.08.31.	Charitable Org...	807037957RR0001
Artbarn School	Ongoing programs:Advanc...	1	0	http://www.artbarnschool.com	0	202731	193781	44247	Registered	2008.02.01.	Charitable Org...	860767094RR0001
A PLACE TO PLAY NURSERY S...		0	3		0	90362	88773	69384	Registered	1988.07.01.	Charitable Org...	106718489RR0001
Amyotrophic Lateral Sclerosis Soci...	Ongoing programs:1.We pr...	11	0	http://www.als.ca	0	4827866	4925359	1001022	Registered	1977.01.14.	Charitable Org...	106708977RR0002
Alzheimer's Drug Discovery Found...		0	0		0	552790	169790	0	Registered	2012.04.25.	Public Foundat...	819965286RR0001
Adventist Elementary School		2	0		0	139323	136735	134015	Registered	2007.09.06.	Charitable Org...	107797383RR0133
Avila Music School		0	0		0	25632	16612	0	Registered	2010.09.23.	Charitable Org...	839959897RR0001
Arbed Rescue		0	0		0	32614	0	3827	Registered	2009.09.28.	Charitable Org...	858721061RR0001
Assumption College Catholic High ...		0	0		0	16784	16779	0	Registered	2008.07.01.	Public Foundat...	840312359RR0001
ACADEMY OF SCIENCE FOR ED...		0	0		0	0	0	0	Registered	2003.08.14.	Charitable Org...	881782700RR0001
ACADEMIE STE. CECILE INTER...		53	71		0	7078899	6637309	2508420	Registered	2000.05.19.	Charitable Org...	896685336RR0001
ABANDONED CATS RESCUE	Ongoing programs:We dedi...	0	0	http://www.werescuecats.ca	0	102033	73813	0	Registered	2006.05.11.	Charitable Org...	855762522RR0001
AFRICAN METHODIST EPISCOP...	Ongoing programs:Tought, ...	4	6	http://www.grantame.com	0	866289	681392	171060	Registered	1989.03.05.	Charitable Org...	890096191RR0001
ADVANCING STEPS DANCE SC...	Ongoing programs:To provi...	0	0	http://www.thedancestudio.ca	0	4827	4990	0	Registered	1989.01.01.	Charitable Org...	890217649RR0001
ADVANCEMENT OF SCRIPTUR...		0	0		0	20824	12259	0	Registered	1967.01.01.	Charitable Org...	886696194RR0001
ADOPT-A-PET/PET RESCUE		0	0		0	46899	43530	0	Registered	2002.10.10.	Charitable Org...	858763287RR0001
ADOLESCENT'S FAMILY SUPPO...	Ongoing programs:The org...	33	12	afssn.org	0	1960026	1934521	1448734	Registered	1992.10.28.	Charitable Org...	139423859RR0001
ADAM SCOTT FOUNDATION	Ongoing programs:Provide ...	0	0	http://www.adamscott.ca	0	0	0	0	Registered	2005.05.30.	Charitable Org...	872090105RR0001
AISLING DISCOVERIES CHILD A...	Ongoing programs:The Fou...	0	0	http://www.aislingdiscoveries.on....	0	105394	56225	0	Registered	1991.04.01.	Public Foundat...	888515848RR0001
AISLING DISCOVERIES CHILD A...	Ongoing programs:The Cen...	173	25	http://www.aislingdiscoveries.on....	0	14679297	14462769	11958247	Registered	2001.04.01.	Charitable Org...	870409489RR0001
AIR CADET LEAGUE OF CANAD...		0	0		0	23729	28382	0	Registered	1997.09.01.	Charitable Org...	130387665RR0019
AILSA CRAIG AND DISTRICT CO...		0	1		0	27682	30224	22768	Registered	1986.05.01.	Charitable Org...	136850047RR0001
AI-CHENG MANDARIN CHINESE...		0	15		0	53971	62298	49706	Registered	1992.09.01.	Charitable Org...	893930396RR0001
ALISCHILD FOUNDATION		0	0		0	0	2912	0	Registered	1984.04.26.	Private Found...	119211894RR0001
ALGONQUIN & LAKESHORE CA...	Ongoing programs:To admi...	1309	740	http://www.alcdsb.on.ca	0	1429315...	1399783...	1073673...	Registered	1998.01.01.	Charitable Org...	878795020RR0001
ALEXANDER HENRY HIGH SCH...		0	0		0	4385	4350	0	Registered	1989.09.22.	Charitable Org...	890860240RR0001
ALTA VISTA CO-OPERATIVE NU...		0	0		0	197400	191421	166723	Registered	2001.03.12.	Charitable Org...	885826362RR0001
ALLISTON INTERDENOMINATI...	Ongoing programs:Our char...	5	7	http://www.allistonccs.ca	0	663775	718821	362169	Registered	1981.01.09.	Charitable Org...	106702509RR0001

THE SCRAPE DATA TAB

This section allows to obtain the data from the Canadian Charity Organisation homepage and save it to database.



Select the Action:
● Update data
○ Update from CSV
○ Re-Fetch failed data

Start with page: 0
Sync with database after this number of pages fetched: 1
Max number of parallel workers (the balance between speed and succesfully fetched records): 4

With (5) it is possible to specify maximum level of concurrent requests.

XML MAPPINGS DESCRIPTION

The program needs two xml files for it to understand where the information is located in webpage. testMap.xml holds information for Update data process, while testSelMap holds same information for re-fetch failed rows process.

LISTINGS PAGE MAPPINGS AND SOME SETTINGS

FirstResultPageFormat - Holds an URL to first page of advanced search results. Please notice that there must be a {0} written inside it in order for program to know where to put page number so to navigate between results pages.

URLPrefix - Holds Url that is prepended to every link in search results page so to open the corresponding details page. It must not include trailing slash (/)

NumHolder - Holds XPath to element that contains information about total records count in search results page.

NumHolderRegEx - Holds Regular Expression that is applied to NumHolder in order to extract number from its text.

ResultLinks - Holds XPath to elements (href attribute) that contains links to detail pages.

DownloadResults - Holds XPath to element that contains link (href attribute) to csv when geting addresses out of their csv.

COMMON MAPPINGS

Organization_name - Holds XPath to element that contains information about the name of the organisation.

<!-- Match at groups[2] means quick info, groups[4] - Details page (like for 849923438RR0001)-->

Organization_name_RegEx - Holds Regular Expression that is applied to Organization_name in order to extract organisation name from its text.

QUICK VIEW PAGE MAPPINGS

Registration_no - Holds XPath to element that contains information about the Registration number of the organisation.

Designation - Holds XPath to element that contains information about the Designation of the organisation.

Web_site - Holds XPath to element that contains information about the Website of the organisation.

Programs_and_activities - Holds XPath to element that contains information about the Programs and activities of the organisation.

Status - Holds XPath to element that contains information about the Status of the organisation.

StatusDate - Holds XPath to element that contains information about the StatusDate of the organisation.

StatusDateFormat - Holds XPath to element that contains information about the StatusDateFormat of the organisation.

Operations_outside_Canada - Holds XPath to element that contains number of Operations outside Canada of the organisation.

Operations_outside_Canada_RegEx - Holds Regular Expression that is applied to Operations_outside_Canada in order to extract number from its text.

Total_revenue - Holds XPath to element that contains information about the Total revenue of the organisation.

Total_revenue_RegEx - HHolds Regular Expression that is applied to Total revenue in order to extract number from its text.

Total_expenses - Holds XPath to element that contains information about the Total expenses of the organisation.

Total_expenses_RegEx - Holds Regular Expression that is applied to Total_expenses in order to extract number from its text.

Total_compensation_for_all_positions - Holds XPath to element that contains information about the Total_compensation_for_all_positions of the organisation.

Full_time_employees - Holds XPath to element that contains number of Full time employees in the organisation.

Part_time_employees - Holds XPath to element that contains number of Part time employees in the organisation.

DETAILS VIEW PAGE MAPPINGS

DRegistration_no - Holds XPath to element that contains information about the Registration number of the organisation.

DDesignation - Holds XPath to element that contains information about the Designation of the organisation.

DWeb_site - Holds XPath to element that contains information about the Web site of the organisation.

DPrograms_and_activities - Holds XPath to element that contains information about the Programs and activities of the organisation.

DStatus - Holds XPath to element that contains information about the Status of the organisation.

DStatusDate - Holds XPath to element that contains information about the StatusDate of the organisation.

DStatusDateFormat - Holds XPath to element that contains information about the StatusDateFormat of the organisation.

DOperations_outside_Canada - Holds XPath to element that contains number of Operations outside Canada of the organisation.

DOperations_outside_Canada_RegEx - Holds Regular Expression that is applied to DOperations_outside_Canada in order to extract number from its text.

DTotal_revenue - Holds XPath to element that contains information about the Total_revenue of the organisation.

DTotal_revenue_RegEx - Holds Regular Expression that is applied to Total_revenue in order to extract number from its text.

DTotal_expenses - Holds XPath to element that contains information about the Total_expenses of the organisation.

DTotal_expenses_RegEx - Holds Regular Expression that is applied to Total_expenses in order to extract number from its text.

DTotal_compensation_for_all_positions - Holds XPath to element that contains information about the Total compensation for all positions of the organisation.

DFull_time_employees - Holds XPath to element that contains information about the Full time employees of the organisation.

DPart_time_employees - Holds XPath to element that contains information about the Part time employees of the organisation.