Data Sourcing Needs for Project

**High Level Goal:**

Our goal is to specific data from Guidestar Pro. There are just under **2 million nonprofits** across the United States (U.S). Note- There are multiple limitations put in place to prevent scraping from this site.

**Data or Fields to collect:**

See the files:

* Guidestar NonProfit DataFields Needed\_v1.xlsx - The fields we are interested in.
* Data Mapping for Guidestar.org EandP -v1 02202023.docx -(Where to get the data for the fields listed in the spreadsheet)

**Websites or Public Sources to get data from:**

1. <https://www.guidestar.org/>

**(NOTE – YOU NEED TO LOG IN TO GET THIS DATA)**

**Skillset/Technology Requirements:**

* Experience with Python and Webscraping libraries
* Experience using Mysql database
* Experience doing similar Webscraping projects.
* Experience using IP proxies, and other similar methods of bypassing rate limiting/ restrictions.

**Deliverables:**

1. The data needs to be stored in a simple database (mysql).

* What should be provided to us is a simple SQL data dump (create statements and insert statements) which we can use to recreate your database and tables and also re-create the data saved.

Please ensure that the data saved is mapped to the identified data set (fields) in the files:

* Guidestar NonProfit DataFields Needed.xlsx

Also save in the database, the exact URLs where the data was primarily pulled from.

1. Note – it is important you show the first **100 datasets (in the delivery format)** before completion of the project so that we can provide feedback on initial data quality, if needed, before completion.

**Other Information:**

More specific data format requirements will be provided once we start the project.