Data Extraction

Most Federal agencies from which we obtain data make their content available through multiple methods. Although our table structure is designed to be similar to the source data and the content is publicly available, finding and formatting the data can sometimes be cumbersome. Historically, the ARC's approach has been to obtain the data and repackage it into files for a more seamless import into the database. As federal agencies have made their data more accessible, more and more states are obtaining their data directly from those agencies. WID-formatted files aren't without value, though – for states that rely on a contractor for managing their WID, the files are the core input.

We're attempting to support both approaches by adjusting how we provide data. While there's no plan to discontinue producing the widely-used files on our site, we're also adding Python utilities to create those same files in your own offices. This serves a few functions:

1) The data is always up to date, even if there have been revisions
2) The best way to access the data is defined in the script, saving time and effort in defining those criteria if a different approach is needed
3) The output is more easily customized – if only some states are needed, those can be excluded from the output file.

Background

While there are multiple ways to share processes, the option we've settled on is using Python scripts, made available publicly through GitHub. The software required is free and Python is widely used, so hopefully will be useful to a broad cross-section of states.

What do you need installed:

1) Python
2) Git (necessary to seamlessly get the package from Github)
3) Packages (e.g. pandas, requests, numpy, datatime, these are imported at the top of the script)
4) A means of executing the script – this can be done in many ways, including from the command line, and there is likely some preferred software in the organization. We use Visual Studio Code, a free Microsoft product.

Scripts can be extracted from GitHub seamlessly by connecting to the repository or can be copied as text.

Data Sources

Most data is sourced either from LABSTAT (a BLS file server which requires no authentication), or from an agency's API. When an API is used, the key or authentication information will not be included in the script. Users will need to obtain that for their own organization and insert it into the script. When that is needed, it will appear in the script as an empty variable labeled "api_key" and the script will fail unless it has been edited by the user to assign a valid key.

```
api_key = '' # insert API key here
```

Examples available

So far the JOLTS, BEA, and Demographics tables have available content. Those can be obtained directly from the GitHub repository here: https://github.com/CoralVertex/WID-Table-Extraction

Future

If there's demand, similar processes could be made available in R.

If there are questions or comments please direct them to arc.deed@state.mn.us.