We are in need of a data extraction (E) + transformation (T) tool -- the importer -- that converts information stored in various file types (such as .csv, .xls, .xlsx, .mdb, .accdb, etc.) and structures (wide, long, dyadic, etc.) into a standardized .json format based on an already established data schema and built off an existing prototype. Information on the json data schema can be found here:

<https://github.com/DataGator/contrib/tree/master/data>.

The data extraction ( E ) logic is how to generate the flow of events from data within each source file; and the data transformation ( T ) logic is how to serialize the data in the flow to separate JSON objects. The JSON output of the E+T tool will eventually be uploaded into a cloud-based data aggregation tool, aka. DataGator, that is currently under development. The full-fledged importer tool needs to be robust enough to identify different variants of standard data fields and records, including non-standard characters within the data as well as within the file names, and reformat and package these into the final .json format that the system, aka. DataGator, recognizes.

We are looking for developers with expertise in Java and data processing. Familiarity with Jackson/Apache POI would also be advantageous. This project has the possibility of becoming a longer-term arrangement for the development of additional auxiliary components for the DataGator project. The final iteration of this tool will be included in a public rollout for the DataGator system.

To facilitate candidate developers to understand our requirements, we have created a prototype of the importer tool. A working example is provided within the prototype that can convert a single .csv file to JSON output. The prototype of the importer tool can be found in our public GitHub repository,

<https://github.com/DataGator/contrib/tree/master/samples/importer>

The work that needs to be completed would be to extend this prototype both horizontally (i.e. extract data from other file types such as .xls, .xlsx, .mdb, .txt) and vertically (i.e. support data filtering, merging, and hierarchical extraction, i.e. selecting multiple work sheets within a single Excel files, one sheet in multiple Excel files, multiple sheets within multiple Excel files etc.).

There are multiple modules that need to be constructed based on file types and additional functionality. **Please note that this project is only for Module 1 (XLSX Extractor)** and all additional modules will be completed under follow-on projects dependent on successful completion of **Module #1**.

**Module 1: Build E+T functionality for .xlsx file**

**Module 2:** Build E+T functionality for .xls file

**Module 3:** Build E+T functionality for .mdb file

**Module 4:** Build E+T functionality for .txt file

**Module 5:** Allow for processing of multiple files

**Module 6:** Create a Java GUI for client-side operation

**Note:** The logic for multi-sheet xls / xlsx/mdb files can have three alternative options (i) extract data from one of the sheets (ii) extract data from multiple sheets and combine them into a single JSON object (iii) extract data from multiple sheets and output them as individual JSON objects.

Our expectation is that **Module 1 (XLSX Extractor)** will be able to be completed within **3-4 days** of being awarded the project and the freelancer is qualified to complete **Modules 2** through **6**. We will provide several sample files that will need to be converted to JSON format and eventually uploaded, however it must be understood that these sample files are not an exhaustive list of characters within a file or file structures. The ideal freelancer will be willing and able to identify potential issues that may arise as other users begin to use it and different file structures are loaded into the tool. We are looking for freelancers that are passionate about this project and interested in a longer term working relationship.