CZO: Geostreaming Data Framework Integration

Goals

- CSV files uploaded to Clowder are annotated with information about the variables contained within the file using standard vocabularies.
- This metadata, together with metadata about the location or sensor attached to a dataset is used to automatically ingest data into the Geostreaming API.
- Given an annotated tabular file, apply format unit conversion to specific columns and create a new version of the tabular data.

Components

- Clowder
    - Dataset is annotated with sensor information
        - Reuse existing relationship between dataset and sensor
        - Or... add metadata to dataset
- Variable Annotation Extractor (VAE) ☑ **~~BD-2315~~** - czo semantic extractor **DONE**
    - Annotate files with entries from standard vocabularies
        - Col. 3 contains term http://odm2/precipitation
        - Multiple mappings can be provided, each with their own likelihood
            - For example, if only 9 out of 10 columns match a prior mapping, likelihood is 90%
            - Or percentage of files seen with this type of mapping
- Variables Mapping Service (VMS) ☑ **~~BD-2310~~** - Variables Mapping Service **DONE**
    - POST/GET/PUT/DELETE mappings
    - The collection in MongoDB contains documents that represent mappings
        - Each mapping is a collection of mappings between strings (column headers) and standard vocabularies (uri terms)
        - How many times have seen a particular mapping (how many unique files)
        - When a mapping is not complete, i.e. we can only identify a subset of the columns, we should keep track of how many we columns we successfully identified
            - let's say a csv file has 10 columns, but we can only tag 4, we would have 40% accuracy
    - Maybe keep a collection of what files match what mapping
    - SEARCH for mappings that match a set of CSV headers and return them in order of accuracy
        - Client submits one list of CSV column names, service returns a list of potential mappings including accuracies.
    - Dockerize the service:
        - ⚠ BD-2318 - Jira project doesn't exist or you don't have permission to view it.
- Semantic Annotation Service (SAS)
    - http://ecgs.ncsa.illinois.edu/SAS.html
    - We should build a simpler version of this as a Flask application storing info in MongoDB
- Datapoints Extractor (DPE)
    - Creates datapoints in the Geostreaming API based on rows in the CSV input file
    - Requires mapping from Variable Annotation Extractor
    - Site information as metadata on dataset
- Geostreaming Data Framework
    - Store and visualize datapoints
    - https://geodashboard.ncsa.illinois.edu/
    - Geostreaming API (GSAPI)
- Unit Conversion Extractor
    - Given a CSV file and information about what units to convert ??? return a new file with the specific column converted to new units
    - Requires ability to show derived files in GUI
    - How does the user specify what units they want?

Workflow

- File F1 (CSV) uploaded to dataset D1
- VAE reads headers in
- VAE requests matching mappings from mapping service VMS
- VAE adds metadata entries to file F1
- DPE extracts datapoints from CSV and adds them to GSAPI

Tasks

- ~~Update https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-csv to store more information~~ (**Decided as Won't Do.**)
    - ~~which column has which header~~
    - ~~include column number and label, for example (3, "temperature)~~
- Develop Variables Mapping Service (VMS)

    - Simple flask app with mongodb back end
- Variable Annotation Extractor (VAE)
    - En extension of the extractor-csv that queries the VMS and stores standard names in metadata
        - We should support multiple mappings added to metadata
- Figure out where the frontend should be
    - Standalone client
    - Clowder add metadata widget