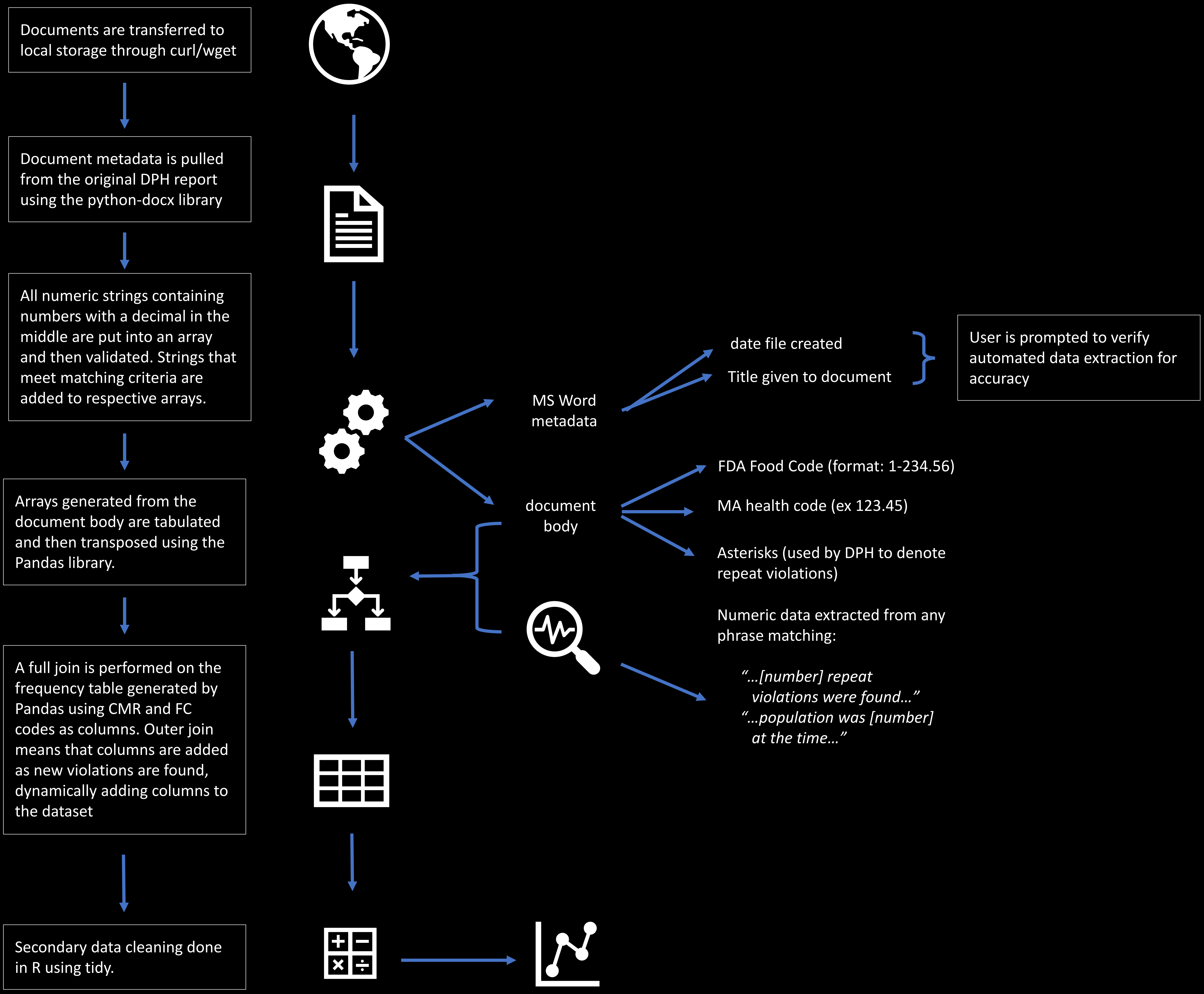

Documents are transferred to local storage through curl/wget

Document metadata is pulled from the original DPH report using the python-docx library

All numeric strings containing numbers with a decimal in the middle are put into an array and then validated. Strings that meet matching criteria are added to respective arrays.

Arrays generated from the document body are tabulated and then transposed using the Pandas library.

A full join is performed on the frequency table generated by Pandas using CMR and FC codes as columns. Outer join means that columns are added as new violations are found, dynamically adding columns to the dataset

Secondary data cleaning done in R using tidy.

MS Word metadata

date file created

Title given to document

User is prompted to verify automated data extraction for accuracy

document body

FDA Food Code (format: 1-234.56)

MA health code (ex 123.45)

Asterisks (used by DPH to denote repeat violations)

Numeric data extracted from any phrase matching:

"…[number] repeat violations were found…"
"…population was [number] at the time…"

jkant@bu.edu