Proposed Features

The following are additional features proposed by T.M. for inclusion in the QC workflow. Short explanations of why each might be useful are included

1. When MongoDB query is used to provide input to workflow, option to write out that input data set as a CSV file.

    - When the workflow is run using data in MongoDB as input, no record is made of the actual data passed into the workflow.
    - Given that the data in MongoDB could change following the workflow run, the provenance of workflow outputs can can be lost.
    - It also could be useful for users to subset their input data set manually using a CSV file and then run the workflow again using this subset (see 1 above).
    - Simple CSV output is desirable for some use cases, e.g. using a scientific name validator to lookup GUIDs for nomenclatural acts from a global authority (e.g. IPNI) for records extracted from a local taxonomic authority file.

2. Preservation of original data values in records passed between actors. KURATOR-119

    - Data validation actors in the QC workflow currently overwrite the original values in the record fields for which they propose updated values.
    - Although comments added as new fields into the records record the original values, these are not as easily read programmatically, e.g. by a user of the report spreadsheet.
    - Overwriting values also means that downstream actors cannot access the original values and propose alternative values based on the originals.

3. Actor for outputting the results spreadsheet (or CSV file) automatically at the end of the workflow run.

    - Currently the QC workflow writes its output to a MongoDB instance. A separate program is used to generate the report spreadsheet from these results in MongoDB.
    - Users may want to use the results of a workflow run without having to query a MongoDB database (manually or using the report-generating program) to evaluate the results of a workflow run.
    - Based on command line options the QC workflow itself could output a report spreadsheet, a CSV file, or both.
    - For large (>60k records) data sets, output as a spreadsheet is impractical.
    - Some classes of advanced users can more effectively work by using command line tools (grep, sed, awk) to isolate portions of a flat text report and either pass them on as QC reports for specialists to resolve, or edit them into sql statements that can directly update a source database (e.g. add GUIDs from a global authority to a local taxon authority file).
    - **Note**: KURATOR-82 relates to part of this, there is a CSV writer actor in FP-Akka that can be composed into a workflow (CSVWorkflow) to write a CSV output report.

Proposed Features that have been added

Option to provide input data to QC workflow using a CSV file (e.g., from a DwC archive) as input. Exists in FP-Akka 1.4.0, under refinement for direct read from DwC archive in FP-Akka 1.5.0

- Embedded in FilteredPush nodes, the QC workflow uses a query provided as a command-line option to retrieve input data from a MongoDB database.
- Users may want to provide data in the form of CSV file so that loading input data into MongoDB instance is not needed.