How to execute DQA on a PEDSnet Site Dataset

This article could be used when executing DQA on Site's dataset version *x* for data cycle *C*, where *x* is an integer and *C* is represented as "Month YYYY", linked to the submission or resubmission for PEDSnet CDM v *a.b.*

Prerequisites

1. Both DQA repositories should be cloned: DQA Analysis (public), DQA Results (private)

2. User must have access to the site database (Postgres) version *x* for the data cycle *C*
3. User must be familiar with the data cycle version document
4. Required downloads
    1. R: https://github.com/PEDSnet/Data-Quality-Analysis#required-downloads
    2. Python and Golang: https://github.com/PEDSnet/Data-Quality-Analysis/blob/master/Tools/dqa/README.md#required-downloads

5. Create RxNorm Reference Table using this DDL script
6. Prepare previous data cycles' summary data files (total 8) to be placed into the PreviousDataSummary folder on the DQA Analysis repo
    1. total_counts.csv and total_fact_type_counts.csv could be retrieved from the (*C-1)th* data cycle's results for the site. For example, if the user is currently executing DQA on CHOP v21, then these files should be retrieved from CHOP v20 results' data folder.
    2. the 6 top50 reports could be retrieved from the DQA results repo. These files were generated during the end-of-cycle operations for the previous data cycle.

Apply DQA Checks

1. Set up the config files
2. Set up R
3. Execute Level 1 DQA
4. Execute Level 2 DQA
5. The result will include 4 folders: data, images, reports, issues. We refer to these results are primary DQA reports. Upload the results on the site's repo, e.g. the results for CHOP v21 DQA are posted here.

Generate Feedback

1. Get familiar with the terminology

```
ETLv{x}:
   {x} = Current ETL Script Version
   Example: ETLv14

{Previous Cycle Name}:
   Month and Year of previous cycle in quotes
   Example: "August 2017"

{Current Cycle Name}"
   Month and year of current cycle in quotes
   Example: "November 2017"

{PEDSnet Convention Version}
   Semantic version number of current PEDSnet Convention in quotes
   Example: "2.7.0"

{GitHub Personal Access Token}
   Personal access token generated from GitHub
```

2. Investigate Differences with previous data cycle
    1. Note: some conflicts may need to be resolved manually
    2. Please commit any changes made to the DQA results repo
3. Review New Issues and Post Feedback
    1. Carefully review new issues before posting them on the site's repo to avoid false alarms
    2. Please commit any changes made to the DQA results repo

Releasing DQA Scripts for Remotes Sites

Note that the remote sites only perform the "Apply DQA Check" step of the DQA toolkit. DCC is responsible for generating the feedback on GitHub

1. Make an official release on the DQA Analysis repo

2. Prepare the previous data summary files (total 8) for the remote site
3. Prepare the wiki document with detailed instructions
4. Once the site has executed the scripts and uploaded their results on the respective repo, DCC will execute the "Generate Feedback" step as mentioned above.

End-of-cycle Activities

1. Update the DCC data cycle version document.
    1. Commit changes to the DQA results repo
2. Generate end-of-cycle summaries to be used in the next data cycle.
    1. Commit changes to the DQA results repo
3. Prepare the DQA summary presentation for the data committee using the instructions here.

Ongoing Activities

1. To process site responses to DQA related GitHub issues, please refer to the labeling guidelines.
2. Log all new DQA issues into JIRA, and convert them into checks based on priority
3. (Remind Levon to) Review no matching conditions (posted in the data folder of primary DQA results)
4. Maintain DQA Dashboard for each data cycle e.g. https://github.com/PEDSnet/Data-Quality-Results/wiki/v2.9-Data-Cycle

Open Technical Challenges

1. Issue Ranking (almost done!)
2. Rewrite Level1 checks using dplyr grammar
3. Migrate secondary DQA reports from csv to relational database
4. Write resolvers for all check types
5. Review persistent issues
6. Write a shell script with a config file for the "Generate Feedback" step

Related articles

Page:How to execute DQA on a PEDSnet Site Dataset