Report ASMS 2022 Energy, Petroleum, and Biofuels Interest Group Workshop:
"Big Data Analytics for Energy, Petroleum and Biofuels"

Presiding: Leonard Nyadong and Yuri Corilo
Date & Time: Tuesday, June 7, 2022. 5:45 – 7:00 pm
Place: Minneapolis Convention Center, Room L100

Workshop Outline

1. Leonard Nyadong (Phillips 66 Energy Research and Innovation, Bartlesville OK, USA) and
   Yuri Corilo (Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland WA, USA.)
   - Welcome
   - Introduction and overview of the workshop

2. Two speakers (30 minutes each) on the topic of "Big Data Analytics for Energy, Petroleum and Biofuels."

   - Qingping Tao. President and COO of R&D, GC Image LLC, Lincoln, NE, USA.
     Topic: Thoughts on Data Ingestion for Analytical Data, Big or Small.

   - William Kew. Chemists, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA.
     Topic: FAIR FTMS Data at PNNL and EMSL

Workshop Summary

The ASMS Energy, Petroleum, and Biofuels Interest Groups held a live workshop after a two-years hiatus due to the pandemic on Tuesday evening (June 7th, 2022) as part of the ASMS Conference on Mass Spectrometry and Allied topics in Minneapolis MN.

Leonard Nyadong (Phillips 66 Energy Research and Innovation) and Yuri Corilo (EMSL, Pacific Northwest National Laboratory) organized the workshop which had less than 20 people in attendance.

The workshop proceeded with an introduction of the topic of the workshop, which was centered around big data analytics and what it means to petroleum and biofuels research. The discussions were tailored to focus on the challenges and enablers for implementing big data analytics in the petroleum and biofuels fields; paying particular attention to the merits for implementing big data analytics, which include, finding the right tools and platform, maintaining data quality, keeping data secure, ensuring metadata capture, and guidelines for FAIR data. Guest speakers were instructed to create an atmosphere to have a discussion

as opposed to a formal presentation. The speakers were urged to be open to interruption for questions or comments.

Qingping Tao shared thoughts on a workflow for data ingestion for analytical data, big or small. Data ingestion involves collecting and moving analytical data from multiple sources to a single shared data system to enable access and analysis. The goals of data ingestion include (1) enabling easy data archiving, sharing and searching, (2) characterizing samples and compounds for comparison or classification and (3) building models to improve or automate raw data processing. Some of the obstacles for implementing a shared data system include the need for different requirements and levels of willingness from different parties, inadequate IT infrastructure and limited budgeting. Some considerations for implementing data storage include, choosing a format that is somewhat future proof (e.g., plain text, CSV, spreadsheet etc.) and being organized.

William Kew led a discussion on application of FAIR data principles for optimizing the reuse of data. These include ensuring the data is findable, which can be achieved by having a unified data storage system with integrated metadata and a perpetual link to the data, which should also be searchable. The data should be accessible via publicly available authentication or via web access. The data should be interoperable using standard file formats or original vendor files that can be converted to open format. The data should be reusable by ensuing there is sufficient metadata and with fair licensing. These principles were demonstrated with the CoreMS software tool being developed at the pacific northwest national laboratory.

The discussions concluded with a question-and-answer sessions, though the speakers were also interrupted during their presentations with questions and comments.

Respectfully submitted,

Leonard Nyadong and Yuri Corilo.

June 2022.