POSTER 23

CMAC FUTURE MANUFACTURING RESEARCH HUB

Making Pharmaceutical Manufacturing Data Ready for AI

Tabbasum Naz[1*], Blair Johnston[1], Murray Robertson[1], Antony Vassileiou[1], Sophie Bailes[2], Neil Dawson[3], Simone Zomer[4], Tiffany Lai[3], Rachel Findlay[5], Gavin Reynolds[2], Amy Robertson[2]

tabbasum.naz@strath.ac.uk

1 CMAC Future Manufacturing Hub, University of Strathclyde, Glasgow, UK, 2 AstraZeneca, Macclesfield, UK, 3 Pfizer, Sandwich, UK,
4 GlaxoSmithKline, Ware, UK, 5 Centre of Process Innovation, Sedgefield, UK

Introduction

Large volumes of pharmaceutical manufacturing data have been generated in recent years. A lot of time and effort has been spent producing data but they are, for the most part, scattered, unstructured, not machine readable and in heterogeneous formats. The work presented here provides integrated management and access to these valuable datasets. The Digital Design Accelerator Platform (DDAP) Extract-Transform-Load (ETL) tool has been developed to derive maximum value from the data acquisition effort to date and to allow future data to be integrated easily. DDAP ETL with multiple components can be used for automatic extraction, transformation and loading of heterogeneous pharmaceutical manufacturing data from multiple instruments. It is a collaborative effort to digitalise and make data Findable, Accessible, Interoperable and Reusable (FAIR). It also provides an opportunity to explore semantic heterogeneity across partners for standardisation efforts and ontology development in the medicine manufacturing domain. DDAP ETL can help domain experts to reap the benefits of the digital age and extract more value from organised data. It provides a foundation for future analytics and data-driven projects across the sector. In future AI, predictive analysis, statistical analysis, data visualization, data mining and machine learning techniques can be applied on the extracted data.

Methodology

ETL Extractor

DDAP Extractor is responsible for extracting schema/data from different raw data sources i.e., Brookfield Powder Flow Tester (PFT), Freeman Technology FT4 Powder Rheometer, Dr Dietmar Schulze Ring Shear Tester (RST-XS) and Micromeritics GeoPyc. There is heterogeneity in data at format, schema and data level and data is not machine readable. Data is available in multiple formats i.e., MS Excel, text files etc. Format heterogeneity makes it difficult to integrate so our DDAP-Extractor component resolves the problem of format heterogeneity by accessing multiple data formats and convert the data to XML format.

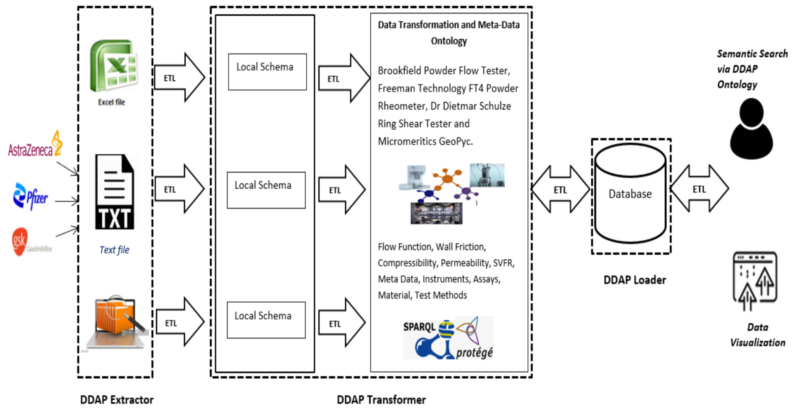
ETL Transformer

With the help of DDAP Transformer, our key requirement is to provide automatic techniques for schema/data transformation. DDAP Transformer is responsible for resolving syntactic and semantic heterogeneity.

ETL Loader

Once the extraction and transformation for schema and data is completed, the data is ready to send to the central repository. The DDAP-Loader component is responsible to load the data into DDAP repository.

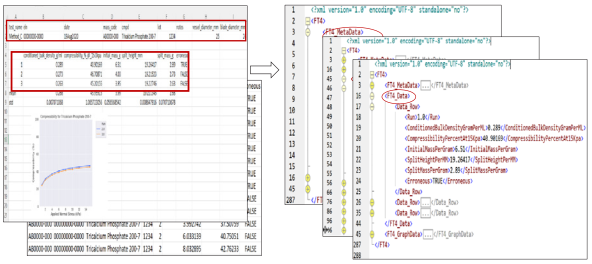
Software Architecture

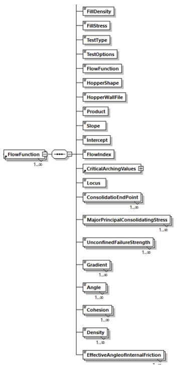


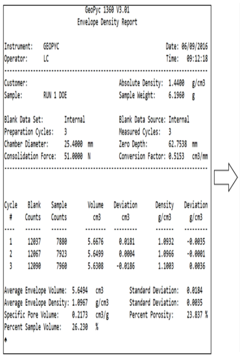
AstraZeneca
Pfizer
GSK
Excel file
Text file
ETL
Local Schema
Local Schema
Local Schema
ETL
Data Transformation and Meta-Data Ontology
Brookfield Powder Flow Tester, Freeman Technology FT4 Powder Rheometer, Dr Dietmar Schulze Ring Shear Tester and Micromeritics GeoPyc.
Flow Function, Wall Friction, Compressibility, Permeability, SVFR, Meta Data, Instruments, Assays, Material, Test Methods
SPARQL protégé
ETL
Database
ETL
DDAP Loader
Semantic Search via DDAP Ontology
Data Visualization
DDAP Extractor
DDAP Transformer

Schema and Data Extraction Process

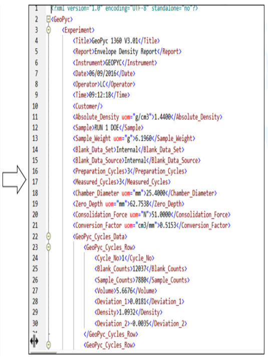
- Freeman Technology FT4 Powder Rheometer's Compressibility schema and data from MS Excel to xml by DDAP ETL
- Schema for Flow Function from Brookfield Powder Flow Tester.
- Schema and data from Micromeritics GeoPyc text file to xml by DDAP ETL.





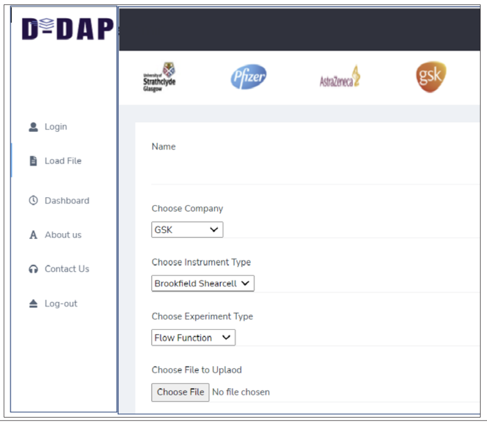


```
                    GeoPyc 1360 V3.01
                 Envelope Density Report

Instrument:   GEOPYC                              Date: 06/09/2016
Operator:     LC                                  Time:   09:12:18
--------------------------------------------------------------------
Customer:                          Absolute Density:  1.4400   g/cm3
Sample:       RUN 1 DOE            Sample Weight:     6.1960   g

Blank Data Set:       Internal     Blank Data Source: Internal
Preparation Cycles:   3            Measured Cycles:   3
Chamber Diameter:     25.4000  mm  Zero Depth:        62.7538  mm
Consolidation Force:  51.0000  N   Conversion Factor: 0.5153   cm3/mm
--------------------------------------------------------------------


Cycle  Blank   Sample     Volume   Deviation     Density   Deviation
  #    Counts  Counts      cm3       cm3          g/cm3      g/cm3
-----  ------  ------   --------  ---------     -------  ---------
  1    12037    7880     5.6676    0.0181        1.0932    -0.0035
  2    12067    7923     5.6499    0.0004        1.0966    -0.0001
  3    12090    7960     5.6308   -0.0186        1.1003     0.0036

Average Envelope Volume:  5.6494   cm3      Standard Deviation:   0.0184
Average Envelope Density: 1.0967   g/cm3    Standard Deviation:   0.0035
Specific Pore Volume:     0.2173   cm3/g    Percent Porosity:     23.837 %
Percent Sample Volume:    26.230   %
```



```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<GeoPyc>
    <Experiment>
        <Title>GeoPyc 1360 V3.01</Title>
        <Report>Envelope Density Report</Report>
        <Instrument>GEOPYC</Instrument>
        <Date>06/09/2016</Date>
        <Operator>LC</Operator>
        <Time>09:12:18</Time>
        <Customer/>
        <Absolute_Density uom="g/cm3">1.4400</Absolute_Density>
        <Sample>RUN 1 DOE</Sample>
        <Sample_Weight uom="g">6.1960</Sample_Weight>
        <Blank_Data_Set>Internal</Blank_Data_Set>
        <Blank_Data_Source>Internal</Blank_Data_Source>
        <Preparation_Cycles>3</Preparation_Cycles>
        <Measured_Cycles>3</Measured_Cycles>
        <Chamber_Diameter uom="mm">25.4000</Chamber_Diameter>
        <Zero_Depth uom="mm">62.7538</Zero_Depth>
        <Consolidation_Force uom="N">51.0000</Consolidation_Force>
        <Conversion_Factor uom="cm3/mm">0.5153</Conversion_Factor>
        <GeoPyc_Cycles_Data>
            <GeoPyc_Cycles_Row>
                <Cycle_No>1</Cycle_No>
                <Blank_Counts>12037</Blank_Counts>
                <Sample_Counts>7880</Sample_Counts>
                <Volume>5.6676</Volume>
                <Deviation_1>0.0181</Deviation_1>
                <Density>1.0932</Density>
                <Deviation_2>-0.0035</Deviation_2>
            </GeoPyc_Cycles_Row>
            <GeoPyc_Cycles_Row>
```

DDAP ETL Interface



D=DAP

Login
Load File
Dashboard
About us
Contact Us
Log-out

Name

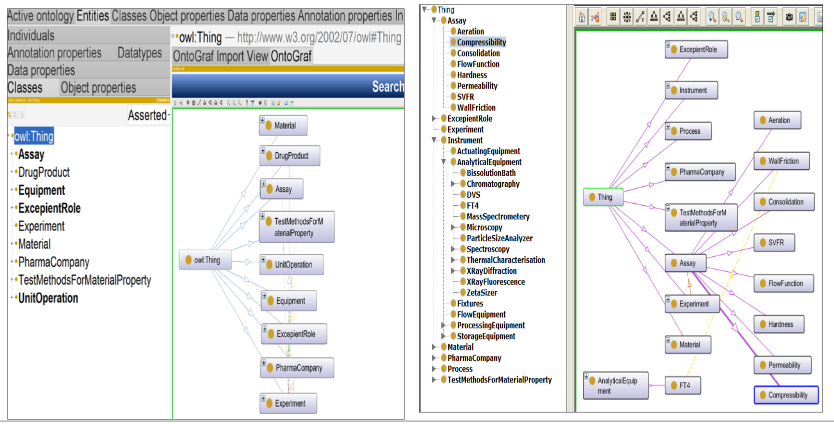
Choose Company: GSK

Choose Instrument Type: Brookfield Shearcell

Choose Experiment Type: Flow Function

Choose File to Upload: Choose File No file chosen

DDAP Ontology for Pharmaceutical Manufacturing Domain



Active ontology Entities Classes Object properties Data properties Annotation properties In
Individuals
Annotation properties Datatypes
Data properties
Classes Object properties
owl:Thing — http://www.w3.org/2002/07/owl#Thing
OntoGraf Import View OntoGraf
Search
Asserted

- owl:Thing
  - Assay
  - DrugProduct
  - Equipment
  - ExcepientRole
  - Experiment
  - Material
  - PharmaCompany
  - TestMethodsForMaterialProperty
  - UnitOperation

- Thing
  - Assay
    - Aeration
    - Compressibility
    - Consolidation
    - FlowFunction
    - Hardness
    - Permeability
    - SVFR
    - WallFriction
  - ExcepientRole
  - Experiment
  - Instrument
    - ActuatingEquipment
    - AnalyticalEquipment
      - BissolutionBath
      - Chromatography
      - DVS
      - FT4
      - MassSpectrometery
      - Microscopy
      - ParticleSizeAnalyzer
      - Spectroscopy
      - ThermalCharacterisation
      - XRayDiffraction
      - XRayFluorescence
      - ZetaSizer
    - Fixtures
    - FlowEquipment
    - ProcessingEquipment
    - StorageEquipment
  - Material
  - PharmaCompany
  - Process
  - TestMethodsForMaterialProperty

Conclusion and Future Work

We have developed DDAP ETL (Extract, Transform and Load) – a tool that can extract pharmaceutical manufacturing data from different sources and develop a mechanism to translate between different concepts and data from multiple schemas. We have started to develop domain ontology in the pharmaceutical manufacturing domain as a way to represent meta-data and semantics. In future, data from DDAP ETL can be reused by experts in AI, predictive analysis, statistical analysis, data visualization, data mining and machine learning.

Innovate UK | University of Strathclyde Glasgow | The University Of Sheffield. | University of Leeds | AstraZeneca | CCDC advancing structural science | cpi | gsk | Perceptive Engineering an Applied Materials company | Pfizer | PSE A Siemens Business