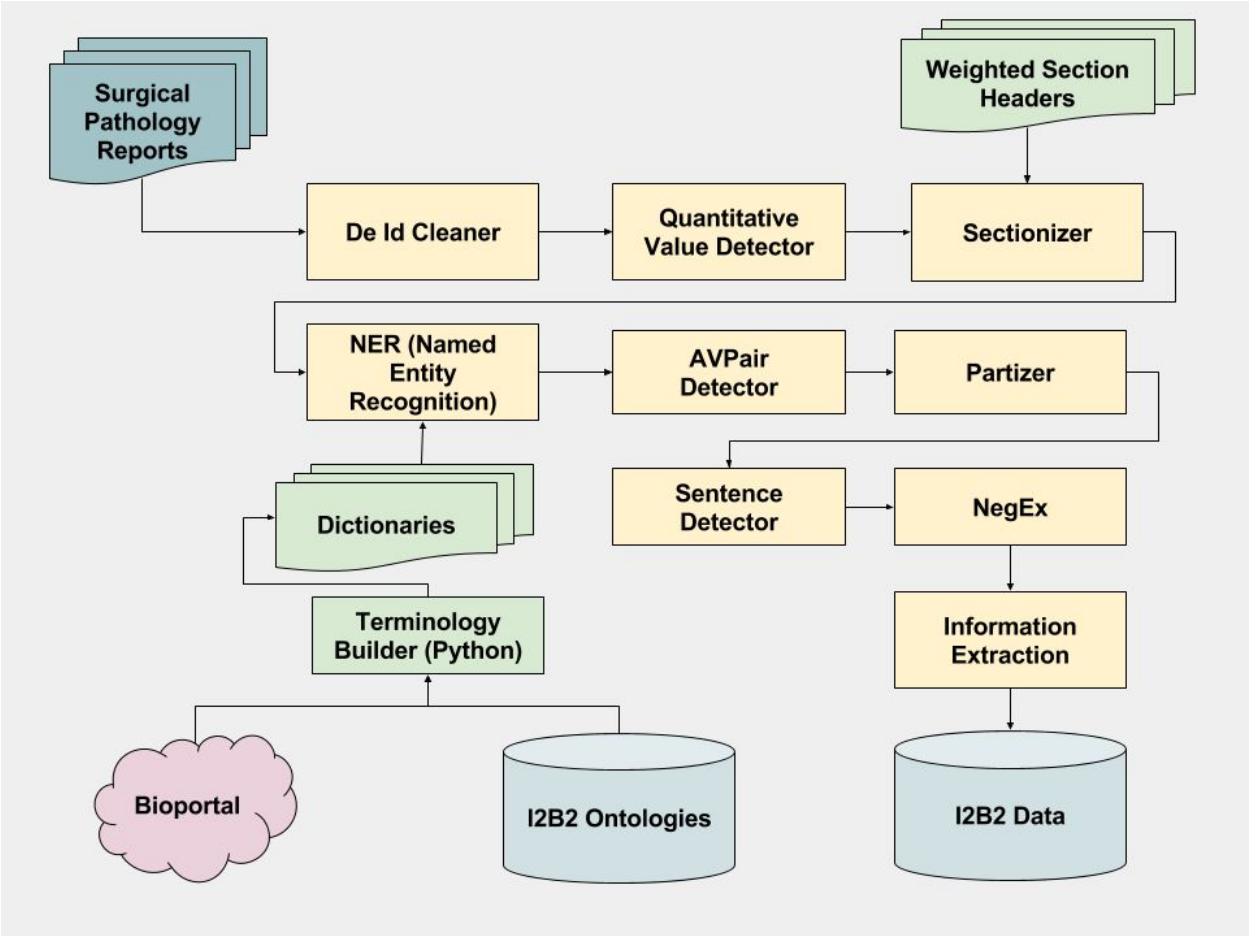
Xmeso User and Installation Guide

Xmeso Description

Xmeso is an Information Extraction Program designed to extract information predominantly from the Mesothelioma Surgical Pathology Reports. It has been written to bolster traditional Extract Translate Load (ETL) approaches for populating Mesothelioma data for the National Mesothelioma Virtual Bank (NMVB) project. NMVB consists of a federated network of four cooperating institutions PITT, NYU, RPCI, and UPENN. The underlying architecture for federated query is based on Shrine connectivity between i2b2 instances housed within each participating covered site. This architecture is similar to that used for ACT and PCORI networks designed by PITT Bioinformatics.

Xmeso Architecture



Xmeso Components

The components of the Xmeso aggregate analysis engine are enumerated in the following table. Xmeso is implemented as a hybrid system of Apache Ruta Scripts and Apache Uima Java Text Annotators. Ruta is essentially a Java Annotator under the hood but the Ruta Scripting language is unique in its expressivity.

Analysis Engine	Function	Implementation
De Id Cleaner	Disables all text generated by the PITT DeIdentifier which is assumed to have been run on the Surg Path input deck	Ruta Script
Quantitative Value Detector	Conflates Floating point values, Determines Sizes as anchored by UOM. Important cleansing step to increase accuracy of part number detection later.	Ruta Script
Sectionizer	Uses Mesothelioma derived gazetteer of section headers to delineate section text with the report. Sections are weighted by “importance”	Ruta Script and hand crafted gazetteer of section headers Additional Java Annotator called Section Creator Annotation Engine
NER (Named Entity Recognition)	A series of Gazetteers generated via on off Python programs against the target i2b2 ontology and bolstered for synonymy and cooccurence via traversal of the Bioportal Rest Services	Ruta Scripts In particular using the MARKTABLE action of the Ruta language. All words near to the “seed” words are assigned distance weights. For now most of these weights are arbitrary but future versions of Xmeso will score decisions by considering them
AVPair Detection	Heuristic rules designed to fill in Yes/No slots of the target templates. Currently Lymph Nodes Involved and Special Stains Used?	Ruta Scripts

Analysis Engine	Function	Implementation
Partizer	Detects part specific text within each Section.	Ruta Scripts and Java Annotation Engine called
Sentence Detector	LINGPipe's Sentence Detector along with tokenization beneath as a prerequisite for NegEx implementation.	Java Annotation Engines
NegEx	Implementation of Chapman's NegEx algorithm	Java Annotation Engine
Information Extraction	Assigns slots values to extraction target template. These are based on appearance of named entities in important sections.	Java Annotation Engine

Xmeso Data Elements

On this cycle we wll extract six Data Elements over the report set.

Case Level:

- Ultrastructural Findings
- Lymph Nodes Examined
- Special Stain Profile

Part Level:

- Histopathologic Type
- Tumor Configuration
- Tumor Differentiation

Xmeso Data Directory

The Xmeso data directory can be anywhere on the file system accessible from the executable jar.
We will explain how to install and run the jar itself in the next section.

Each node will be populated from an NLP acquisition directory that contains a list of files enumerated as MVB0000_123456.txt. Each directory will have these kinds of free text reports along with a properties file and a linkage file

```
REPORT_ID,NMVB_ID,PATIENT_NUM,EVENT_DATE
15869,MVB0002,0002,1991-12-31
15887,MVB0003,0003,1984-05-10
17555,MVB0004,0004,1987-08-08
15979,MVB0006,0006,1979-02-28
15891,MVB0010,0010,1979-01-10
17617,MVB0018,0018,1979-12-31
17623,MVB0019,0019,1980-01-01
17631,MVB0020,0020,1978-08-24
```

The final module will read the %XMESO_HOME%/xmeso.properties file which will contain the dataSetName usually PITT,RPCI,UPENN, or NYU. It will also contain the i2b2 jdbc connection parameters and i2b2 location ontology path and code. There is also a linkage file called

“nmvb_path_report_event_date.csv” that contains linkage from the patient report to visit number and visit date. The surgical pathology reports will be in %XMESO_HOME%/reports with names of the form MVB00002_0002.txt

Files will be read and free text will be piped through the UIMA Ruta annotators. Resulting synoptics will populate i2b2 observation_fact table as well as appropriate dimension tables.

Java source will be packaged into an executable jar file for delivery to the nodes.

Installing Xmeso

Prerequsite to installation and xmeso run are

Java version 7 or higher
git.

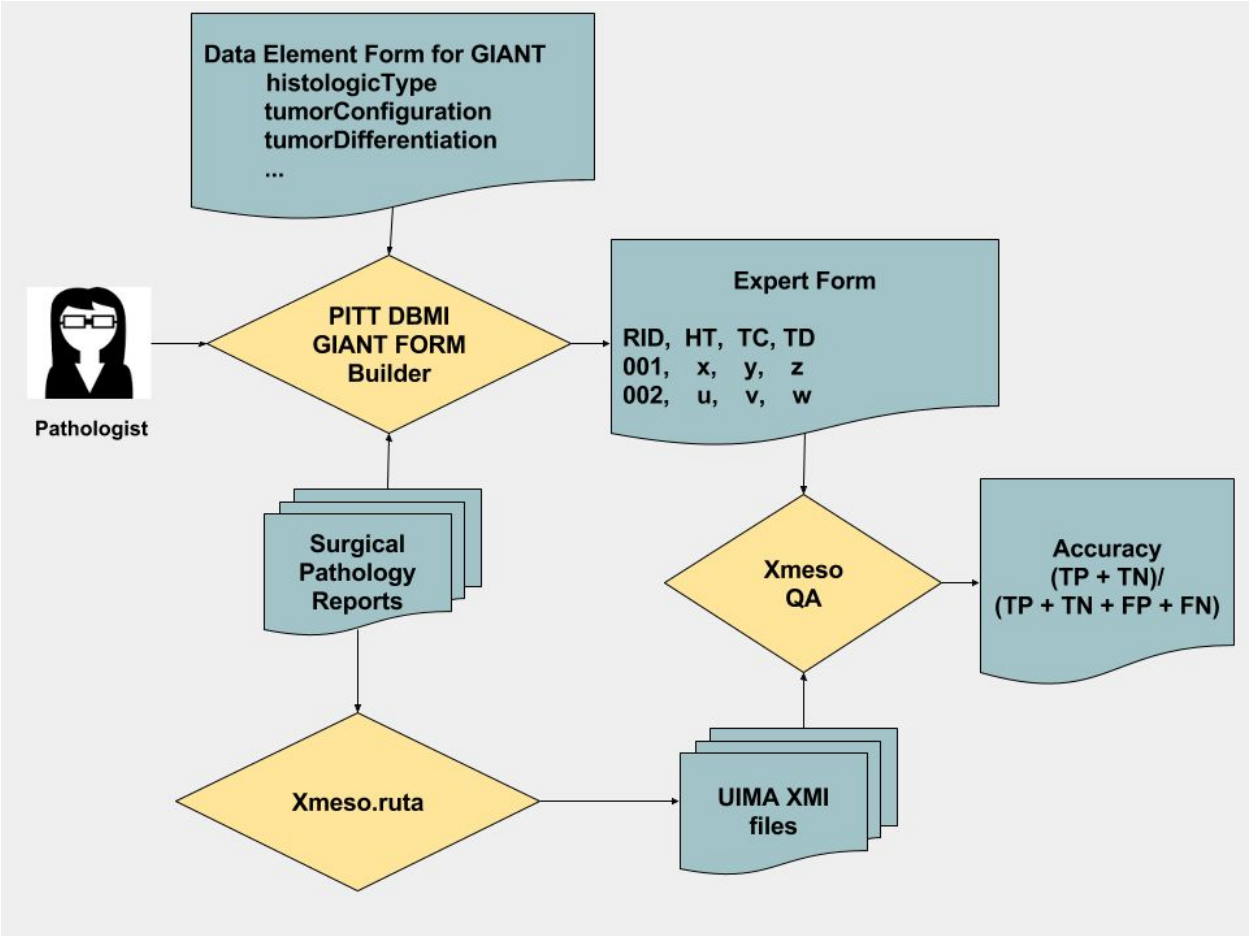
C:\> mkdir C:\run_xmeso
C:\> cd C:\run_xemso
C:\> git clone https://github.com/dbmi-pitt/xmeso.git
C:\> cd xmeso
C:\> set XMESO_HOME=<xmeso_data_directory>

Running Xmeso

C:\> xmeso

Xmeso Quality Assurance

Before Xmeso could be delivered to participating NMVB network sites it was required to attain an accuracy plateau of 75%. The Xmeso Quality Assurance framework is described here.



Xmeso uses the University of Pittsburgh DBMI GIANT tool and a GIANT form definition file with the key prototype Xmeso Data Elements. For Part Form Data Elements each data element is repeated with a numeric suffix up to eight times. Later during accuracy measurements only the expert designated parts are used. For the few cases that have more than eight part mentions only the first eight part mentions are considered by the expert annotator. The result on the Expert GIANT annotation step is a Report# x ((3 DEs for Cases) + (3 DEs for Parts) x (8 part slots)) cell spreadsheet. The Xmeso QA standalone Java program reads a set of UIMA XMI files that have been generated on the test deck by calling Xmeso. The set of generated CaseForms and PartForms are read and compared against the spreadsheet slots. A simple accuracy disagreement is reported.

Initial Xmeso expert analysis for our six elements indicates that the mined results are sparse and a default background setting of “unknown” for both the expert and Xmeso along with Xmeso’s reluctance to assign spurious quantities have led to a rather impressive 85% accuracy overall.

Nevertheless, we assign a special data source for Xmeso in the final I2b2 output tables. This special data source let’s an NMVB end user have an option as to whether NLP results are included in their results or not.