Integrating TCGA Clinical Data Using Metadata-driven Tools and NLP

Richard C. Kiefer[1], Guoqian Jiang, MD, PhD[1], Michael K. Davis[2], Melissa Castine[2], Girish Chavan[2], Guergana Savova, PhD[3], Rebecca Jacobson, MD, MS[2]

[1] Mayo Clinic, Rochester, MN; [2] University of Pittsburgh, Pittsburgh, PA; [3] Harvard Medical School/Boston Children's Hospital, Boston, MA.

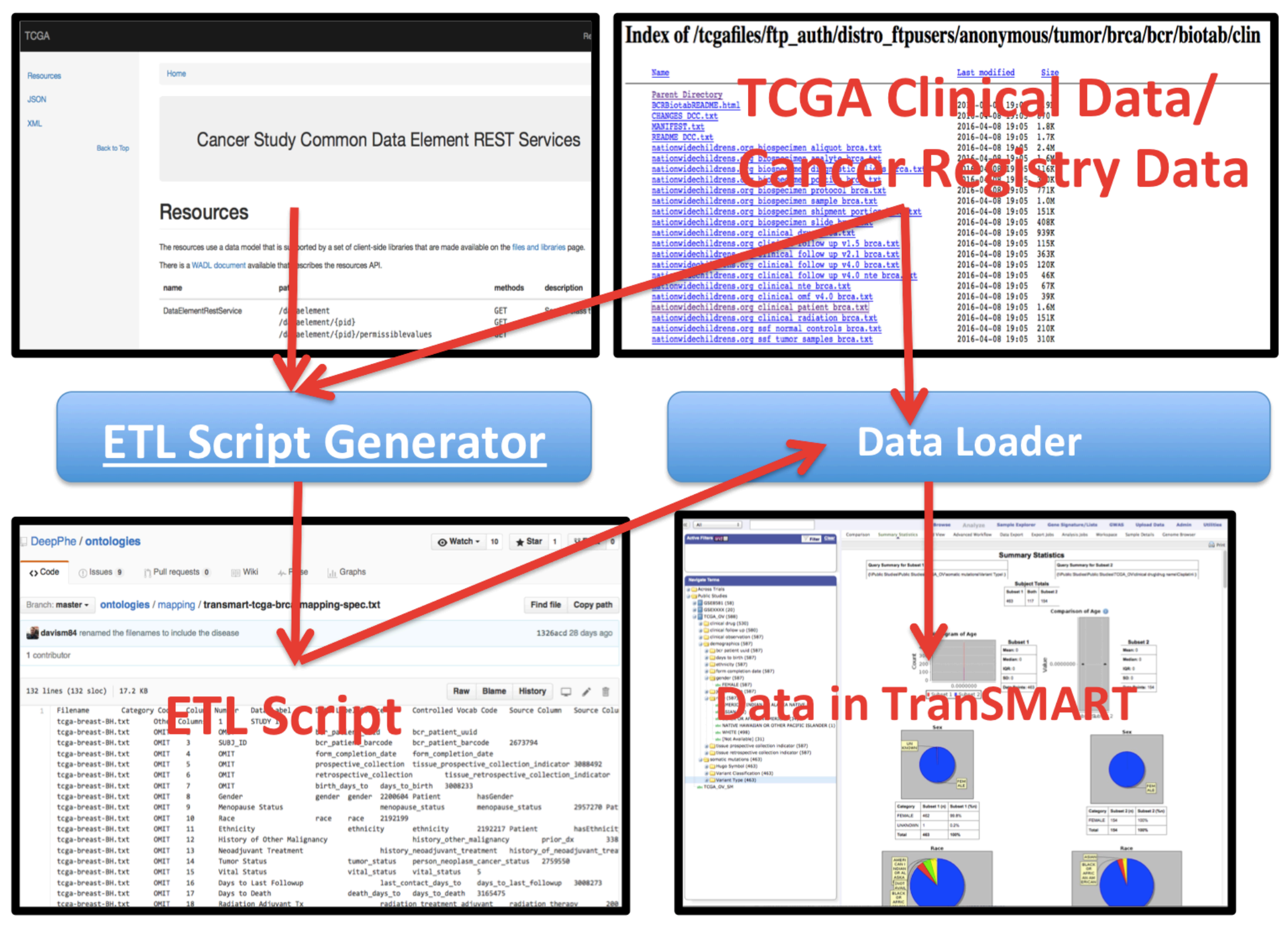
The Integration Problem

Background: Integration of cancer clinical data from EMRs and cancer registries is a major impediment to the development of integrated data repositories (IDRs) for translational cancer research because of the wide range of data models used by these systems.

Objective: We sought to leverage work ongoing in multiple ITCR groups to transcend challenging formats silos that limit integration of The Cancer Genome Atlas (TCGA) clinical text, cancer registry data export files (North American Association of Central Cancer Registries - NAACCR standard based), images and clinical text. A goal of this collaboration was to be able to better integrate TCGA clinical data with molecular data to support translational research projects.

Implementations: This is a collaboration between two NCI ITCR projects: caCDE-QA U01 project, the DeepPhe U24 project and the TIES U24 project. We collaboratively developed a metadata-driven framework to facilitate the ETL process (Figure 1). The caCDE-QA team developed the cancer study common data elements (CDE) RESTful services whereas the DeepPhe team implemented a ETL script generator that consumes the metadata RESTful services (Figure 2) to load TCGA data and Cancer Registry data into tranSMART using the DeepPhe domain ontology to harmonize these datasets. The TIES team developed a TCGA integrated clinical repository that enables users to create cohorts using a combination of NLP features in the TCGA pathology report, as well as TCGA clinical data.

ETL Process

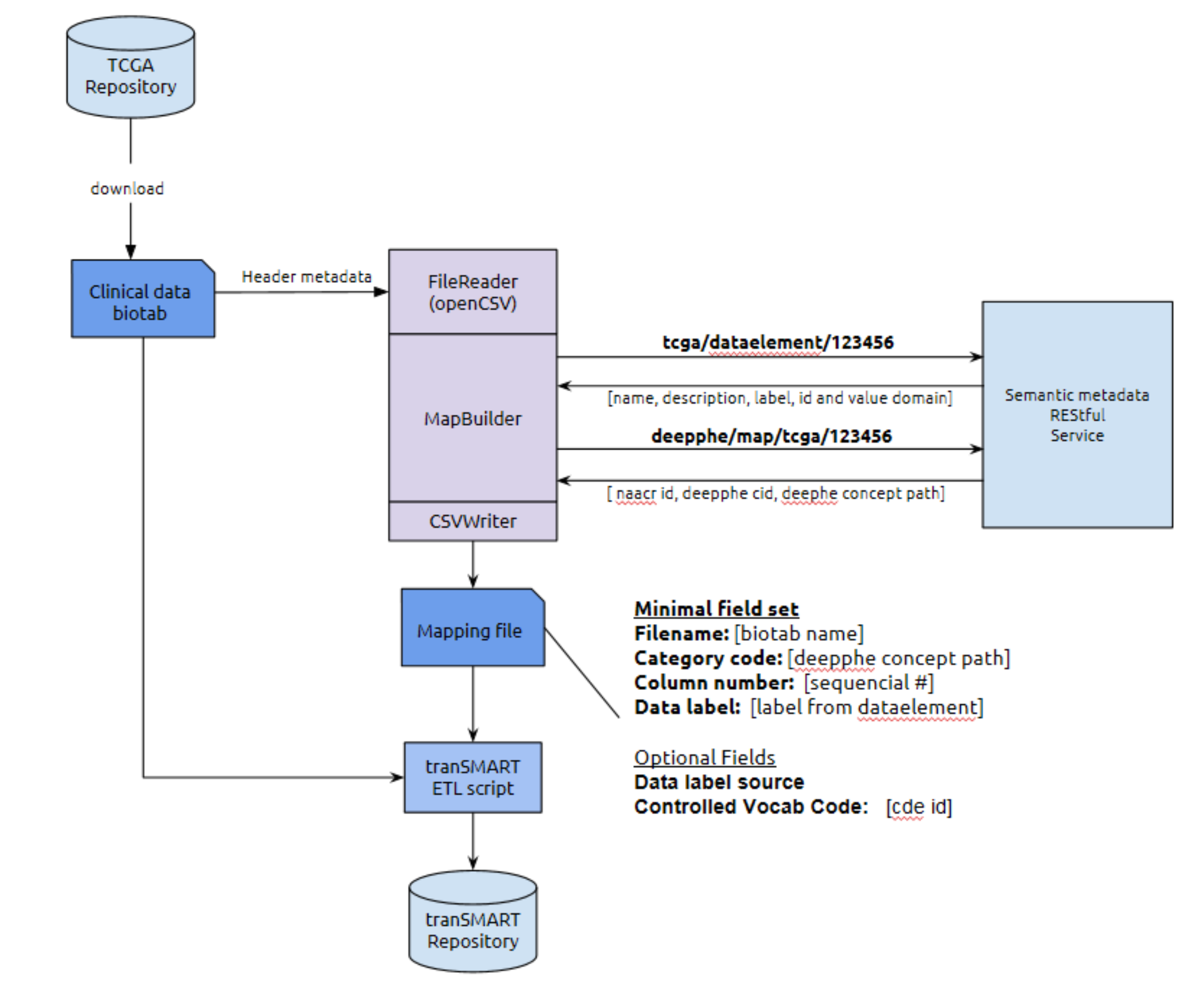


A Metadata-Driven ETL Framework

Metadata Methods

In the repository layer, we leverage the W3C standard Resource Description Framework (RDF) and the meta-data standard ISO/IEC 11179. The data elements from TCGA and NAACCR were converted into RDF triples along with their attribute values. Mappings between TCGA, NAACCR and DeepPhe cancer phenotype models were also written in RDF Turtle format. The RDF Turtle files were then imported into a 4store repository. The REST service APIs were developed so that users of the repository would not have to craft their own SPARQL queries. Using calls to the service APIs, the data elements are returned in JSON format. Calls return all elements from an ontology or a data dictionary, or return a single specified element. The permissible values for an element are also available through the semantic metadata REST services.

ETL Script Generator

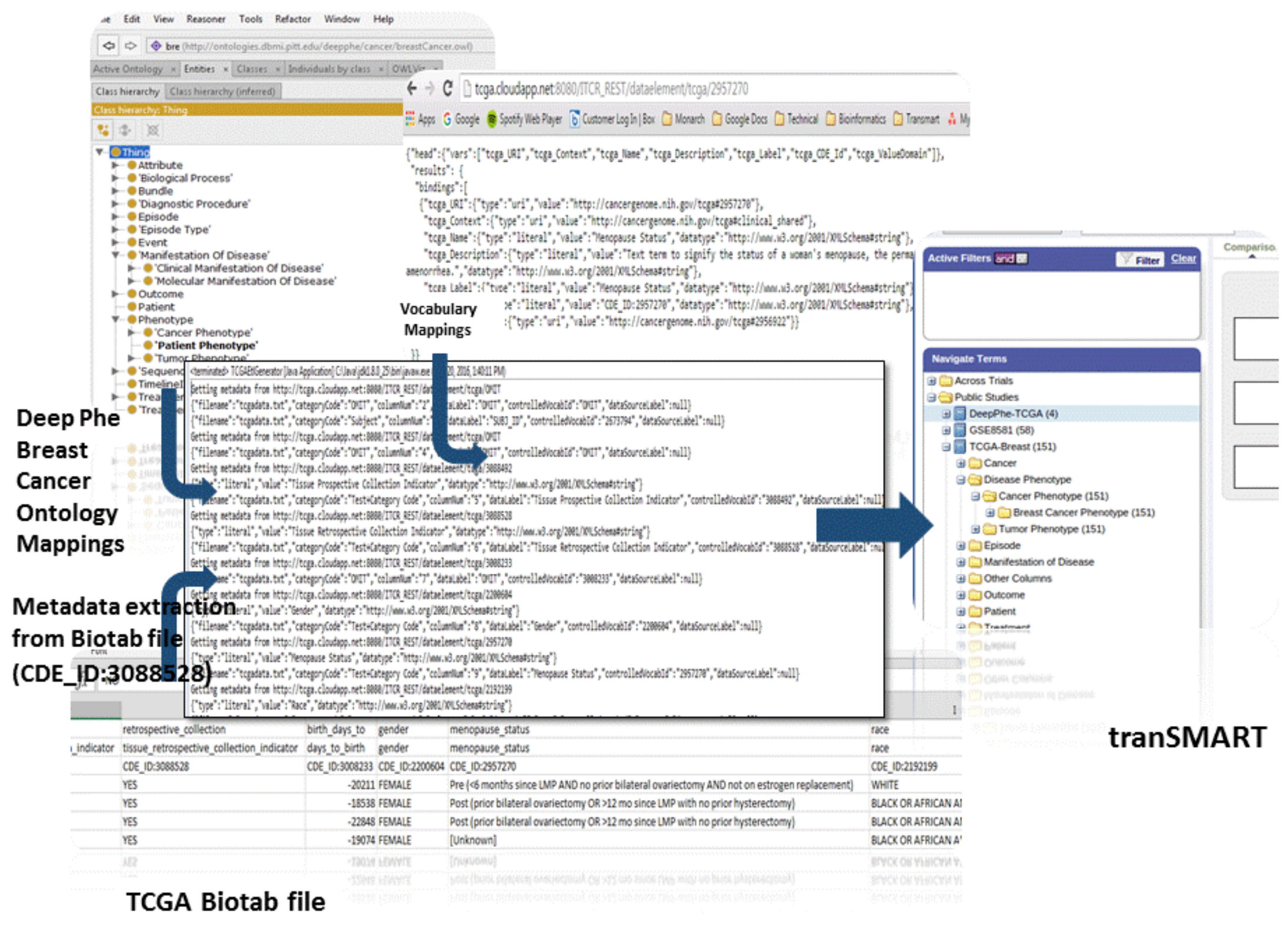


Minimal field set
Filename: [biotab name]
Category code: [deepphe concept path]
Column number: [sequencial #]
Data label: [label from dataelement]

Optional Fields
Data label source
Controlled Vocab Code: [cde id]

System Architecture of an ETL Script Generator

tranSMART ETL


Deep Phe Breast Cancer Ontology Mappings
Metadata extraction from Biotab file (CDE_ID:3088828)
tranSMART
TCGA Biotab file

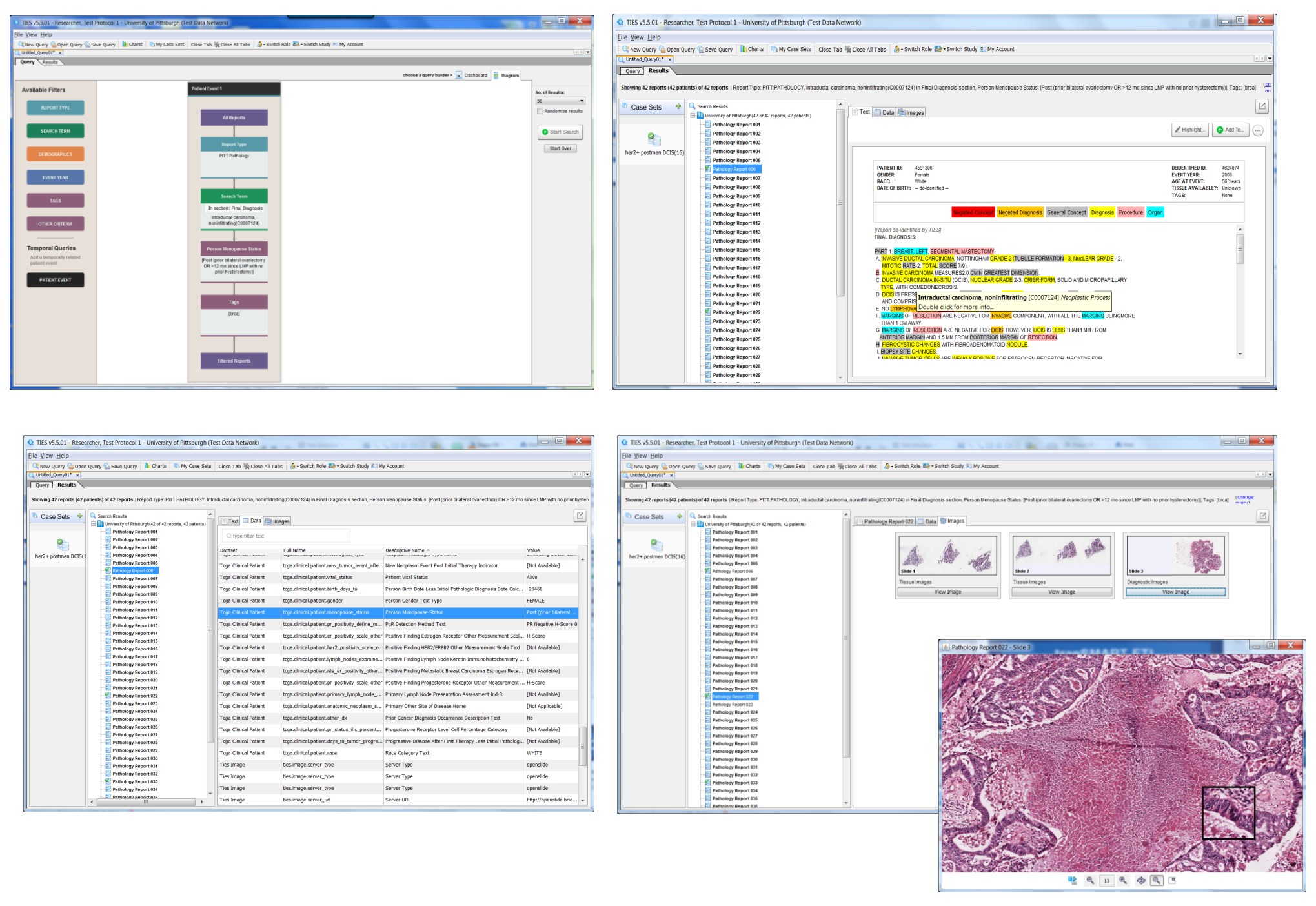
TCGA collects clinical and biospecimen information for all qualified patients which are provided as XML files and tab-delimited text files (biotab). Each clinical biotab file contains a header row of metadata for each column of data (e.g. field name and common data element ID (CDE_ID) or public id). This metadata provides structured data used as a source to obtain the reference mappings. Using this information, the translation process parses the metadata into discrete parts. The CDE_ID is used as a key reference, and a RESTful call is made to the ITCR semantic metadata mapping service. This service contains mapped vocabularies for all the TCGA data elements, the North American of Central Cancer Registries (NAACCR) data elements as well as mappings to their respective DeepPhe ontology mappings. This process ensures that related data elements are consistently mapped across these common sources and provides further consistency across cancers in general. The returned mapping data, from the ITCR service includes information such as preferred data label, permissible values, and ontology classifier, among various other provided information. Returned metadata is used to describe key elements (i.e., category code, data label and controlled vocab code) to map the clinical biotab file. A tab-delimited text file is generated which provides the specification used by the standard ELT scripts, provided by the tranSMART community, to read the clinical biotab files and load data directly into the tranSMART repository.

TCGA Data in TIES

Use Case: Find TCGA cases with DCIS in postmenopausal women, verify by checking WSI, collect TCGA bar codes to identify necessary NGS files

Problems:

- Presence/Absence of DCIS was not routinely collected as part of TCGA structured data
- Requires combination of structured (e.g. pre or post menopause) and unstructured data (e.g. clinical text and images)
- Currently a manual, laborious and error-prone task



Pathology reports converted to text and processed with TIES, images and structured loaded to create unique resource for identifying TCGA cohorts.

Try it out: http://ties.dbmi.pitt.edu/live-demo

Conclusions

- Metadata-driven ETL tools enable reusable ETL script generation and cancer-specific phenotype model bindings.
- DeepPhe ontology was sufficient to generate mappings among multiple different sources of data including Cancer Registry, TCGA clinical data, and phenotype extractions emanating from text.
- Tools and approaches developed can be shared and used by translational investigators
- Collaboration between ITCR groups helps diffuse ITCR innovations among groups, helps us develop trans-ITCR use cases that combine tools, facilitates program level achievements.

References

Code and Services:

- caCDE-QA: https://github.com/caCDE-QA
- DeepPhe: https://github.com/DeepPhe
- TIES https://sourceforge.net/projects/caties/
- Metadata Services: http://informatics.mayo.edu/ITCR

Project Websites:

- caCDE-QA: http://informatics.mayo.edu/caCDE-QA
- DeepPhe: http://cancer.healthnlp.org
- TIES: http://ties.pitt.edu

caCDE-QA DeepPhe TIES