National Center for Advancing Translational Sciences

Metadata Provider

Stanford Center for Biomedical Informatics Research

BMIR
Stanford Center for Biomedical Informatics Research
CONNECTING DATA TO HEALTH

Full Send to:

JHH-2, human cell line STR and SNP profiles from GNE, Genentech

Identifiers BioSample: SAMN03473249; GNE: GNE Tracking ID: 586138

Organism Homo sapiens (human)

Attributes

cell line	JHH-2
culture collection	GNE:586138
repository	Genentech (GNE)
tissue	liver
disease	carcinoma hepatocellular
sex	male
ethnicity	japanese
age	57 year
development stage	adult
canonical name	JHH-2
human cell line STR profile	yes
human cell line STR profile status	repository authenticated
human cell line SNP profile	yes

BioProject PRJNA271020 Homo sapiens
Retrieve all samples from this project

Related information

BioProject

BioCollections

Taxonomy

Recent activity

Turn Off Clear

JHH-2, human cell line STR and SNP profiles from GNE, sul biosample

Human sample from Homo sapiens biosample

"disease=carcinoma hepatocellular"[attr] (28) BioSample

Con rep1 biosample

carcinoma hepatocellular (11749) BioSample

See more...

LinkOut to external resources

JHH-2 (CVCL_2786)

Metadata can answer unique questions

- Has anyone ever performed an **experiment** using methods like these?
- Is there an **investigator** from whom I can obtain a particular cell line?
- What **centers** are studying a given disease?
- Has anyone performed a **study** using these materials?

Metadata from Online Repositories Are Terrible! Look at NCBI BioSample:

- 73% of “Boolean” metadata values are not actually *true* or *false*
  nonsmoker, former-smoker
- 26% of “integer” metadata values cannot be parsed as numbers
  JM52, UVPgt59.4, pig
- 68% of metadata values that are supposed to represent terms from biomedical ontologies do not actually do so.
  presumed normal, wild_type

Gonçalves, R.S., and Musen, M.A. *Scientific Data* **6**:19001, 2019

Full Send to:

JHH-2, human cell line STR and SNP profiles from GNE, Genentech

Identifiers BioSample: SAMN03473249; GNE: GNE Tracking ID: 586138

Organism Homo sapiens (human)

Attributes

cell line	JHH-2
culture collection	GNE:586138
repository	Genentech (GNE)
tissue	liver
disease	carcinoma hepatocellular
sex	male
ethnicity	japanese
age	57 year
development stage	adult
canonical name	JHH-2
human cell line STR profile	yes
human cell line STR profile status	repository authenticated
human cell line SNP profile	yes

BioProject PRJNA271020 Homo sapiens
Retrieve all samples from this project

Related information

BioProject

BioCollections

Taxonomy

Recent activity

Turn Off Clear

JHH-2, human cell line STR and SNP profiles from GNE, sul biosample

Human sample from Homo sapiens biosample

"disease=carcinoma hepatocellular"[attr] (28) BioSample

Con rep1 biosample

carcinoma hepatocellular (11749) BioSample

See more...

LinkOut to external resources

JHH-2 (CVCL_2786)

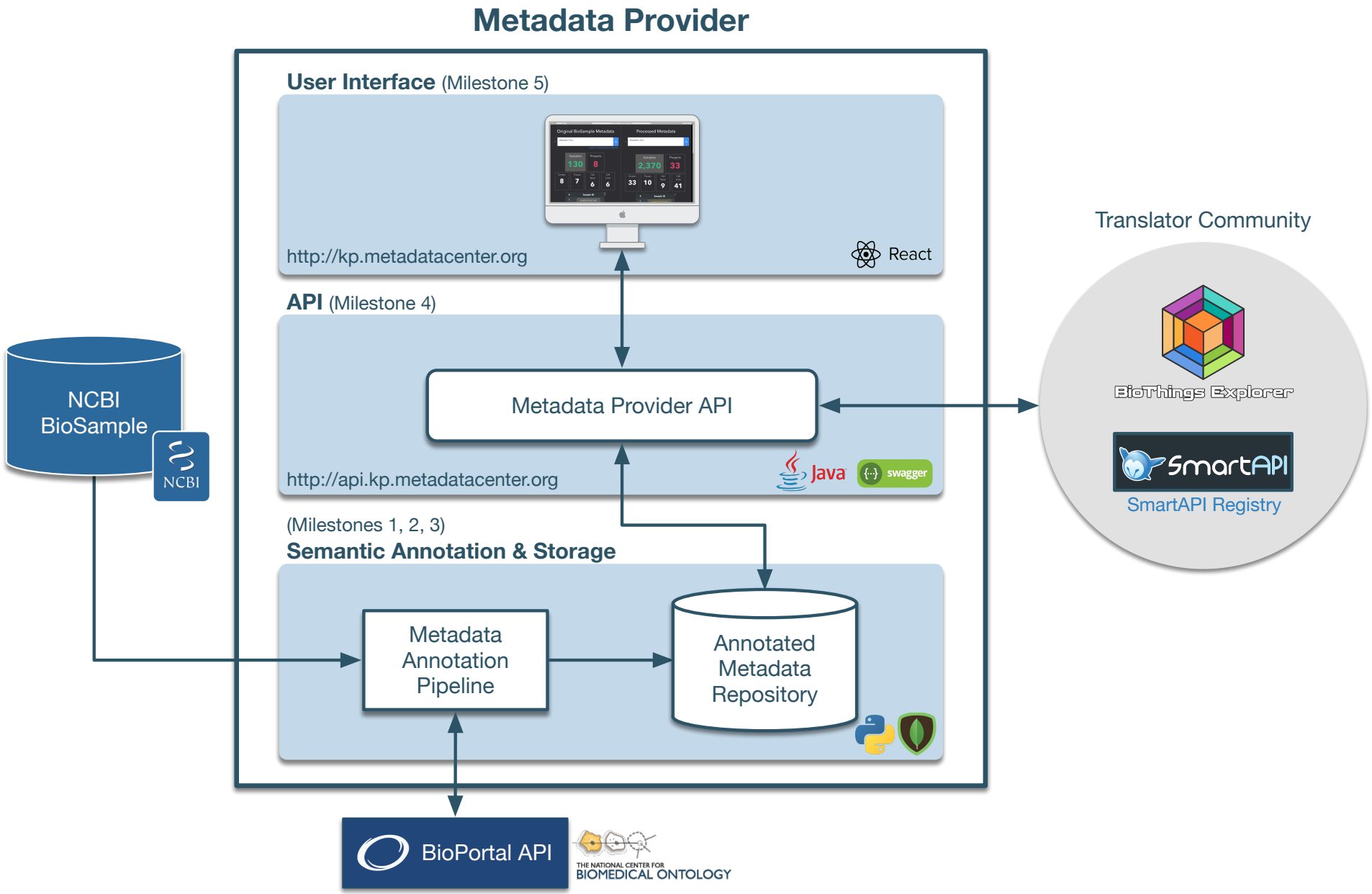
An Automated Pipeline to Enhance Metadata for Use by Translator

- **Milestone 1:** Develop software to annotate metadata attribute names with ontology terms
- **Milestone 2:** Develop software to annotate metadata attribute values with ontology terms
- **Milestone 3:** Apply our prototype software to a subset of NCBI BioSample metadata records
- **Milestone 4:** Make processed BioSample metadata available to Translator
- **Milestone 5:** Demonstrate enhanced query capabilities made possible by our work

An Automated Pipeline to Enhance Metadata for Use by Translator

- **Milestone 1: Develop software to annotate metadata attribute names with ontology terms**
- **Milestone 2: Develop software to annotate metadata attribute values with ontology terms**
- **Milestone 3:** Apply our prototype software to a subset of NCBI BioSample metadata records
- **Milestone 4:** Make processed BioSample metadata available to Translator
- **Milestone 5:** Demonstrate enhanced query capabilities made possible by our work

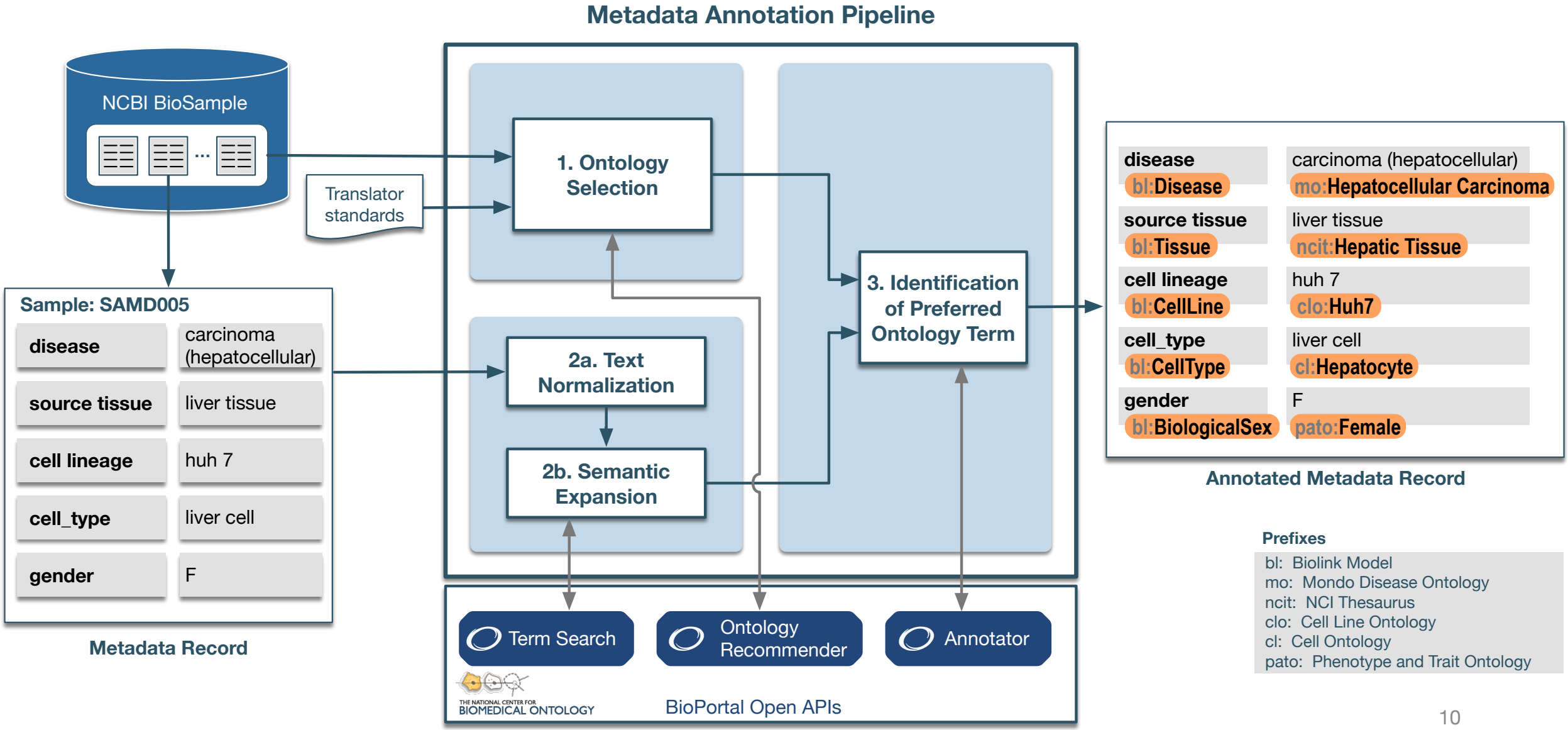
Architectural Overview



Metadata Annotation Pipeline

1. **Ontology Selection:** Find the most appropriate ontologies to annotate metadata attributes (e.g., Mondo Disease Ontology for the attribute *disease*)
2. **Text Preprocessing:** Prepare text for Annotation
    a. **Text Normalization—**Remove special characters and extra spaces
    Carcinoma_[Hepatocellular] → carcinoma hepatocellular
    b. **Semantic Expansion—**Generate term variations using synonyms and word permutations
    carcinoma hepatocellular → *carcinoma hepatocellular, hepatocellular carcinoma, HCC, hepatoma, etc.*
3. **Identification of Preferred Ontology Term:** Identify standardized terms for attribute names and values using the BioPortal Annotator
    hepatocellular carcinoma → Hepatocellular Carcinoma from Mondo

Metadata Annotation Pipeline



Prefixes

bl: Biolink Model
mo: Mondo Disease Ontology
ncit: NCI Thesaurus
clo: Cell Line Ontology
cl: Cell Ontology
pato: Phenotype and Trait Ontology

An Automated Pipeline to Enhance Metadata for Use by Translator

- **Milestone 1:** Develop software to annotate metadata attribute names with ontology terms
- **Milestone 2:** Develop software to annotate metadata attribute values with ontology terms
- **Milestone 3:** **Apply our prototype software to a subset of NCBI BioSample metadata records**
- **Milestone 4:** Make processed BioSample metadata available to Translator
- **Milestone 5:** Demonstrate enhanced query capabilities made possible by our work

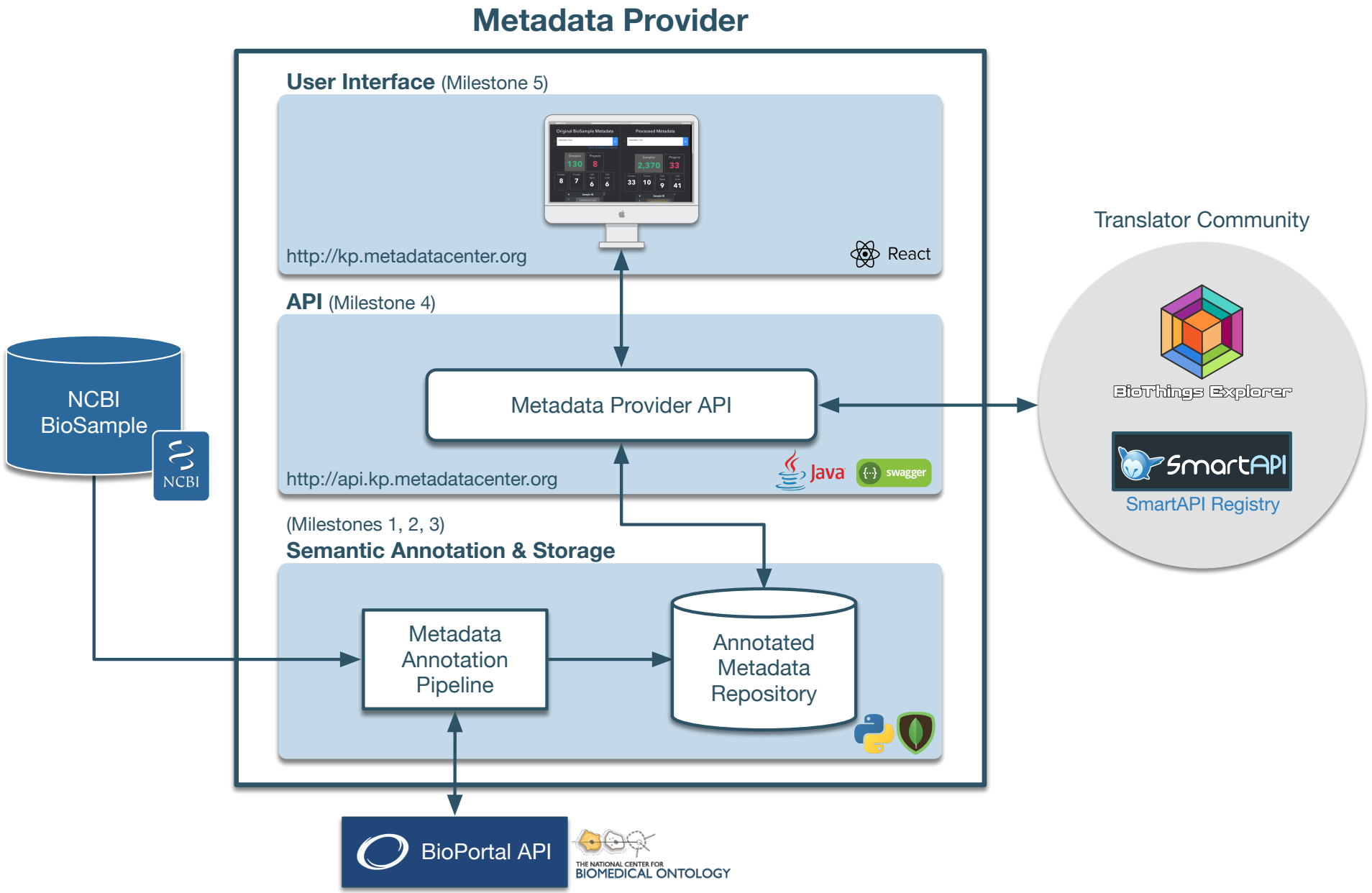
NCBI BioSample Extract

- 4,346 metadata records downloaded on Feb 20, 2020
- 3 representative diseases:
  - *Hepatocellular carcinoma*
  - *Myelodysplasia*
  - *Systemic lupus erythematosus*
- 5 representative attributes:
  - *Disease*
  - *Tissue*
  - *Cell type*
  - *Cell line*
  - *Sex*

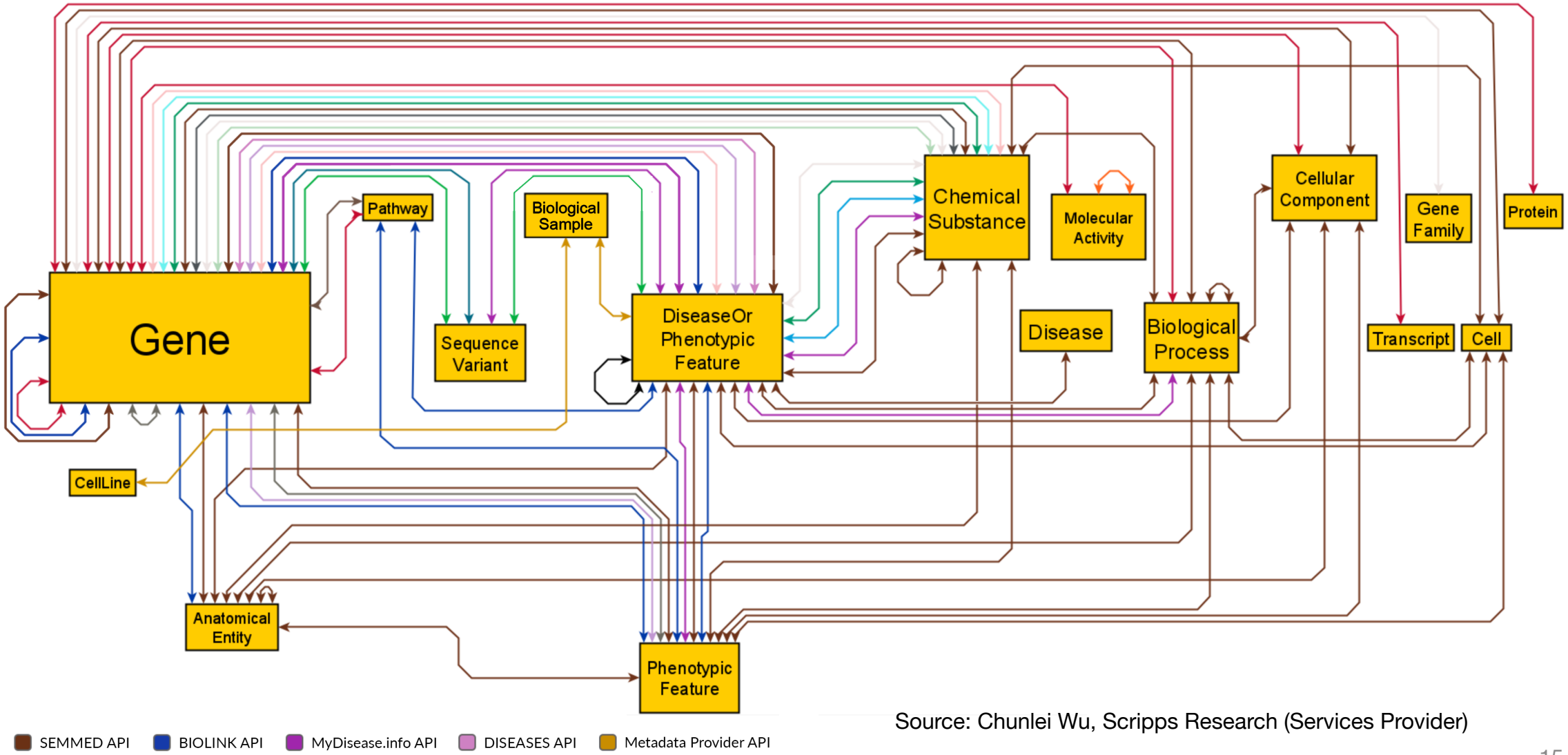
An Automated Pipeline to Enhance Metadata for Use by Translator

- **Milestone 1:** Develop software to annotate metadata attribute names with ontology terms
- **Milestone 2:** Develop software to annotate metadata attribute values with ontology terms
- **Milestone 3:** Apply our prototype software to a subset of NCBI BioSample metadata records
- **Milestone 4:** **Make processed BioSample metadata available to Translator**
- **Milestone 5:** Demonstrate enhanced query capabilities made possible by our work

Architectural Overview

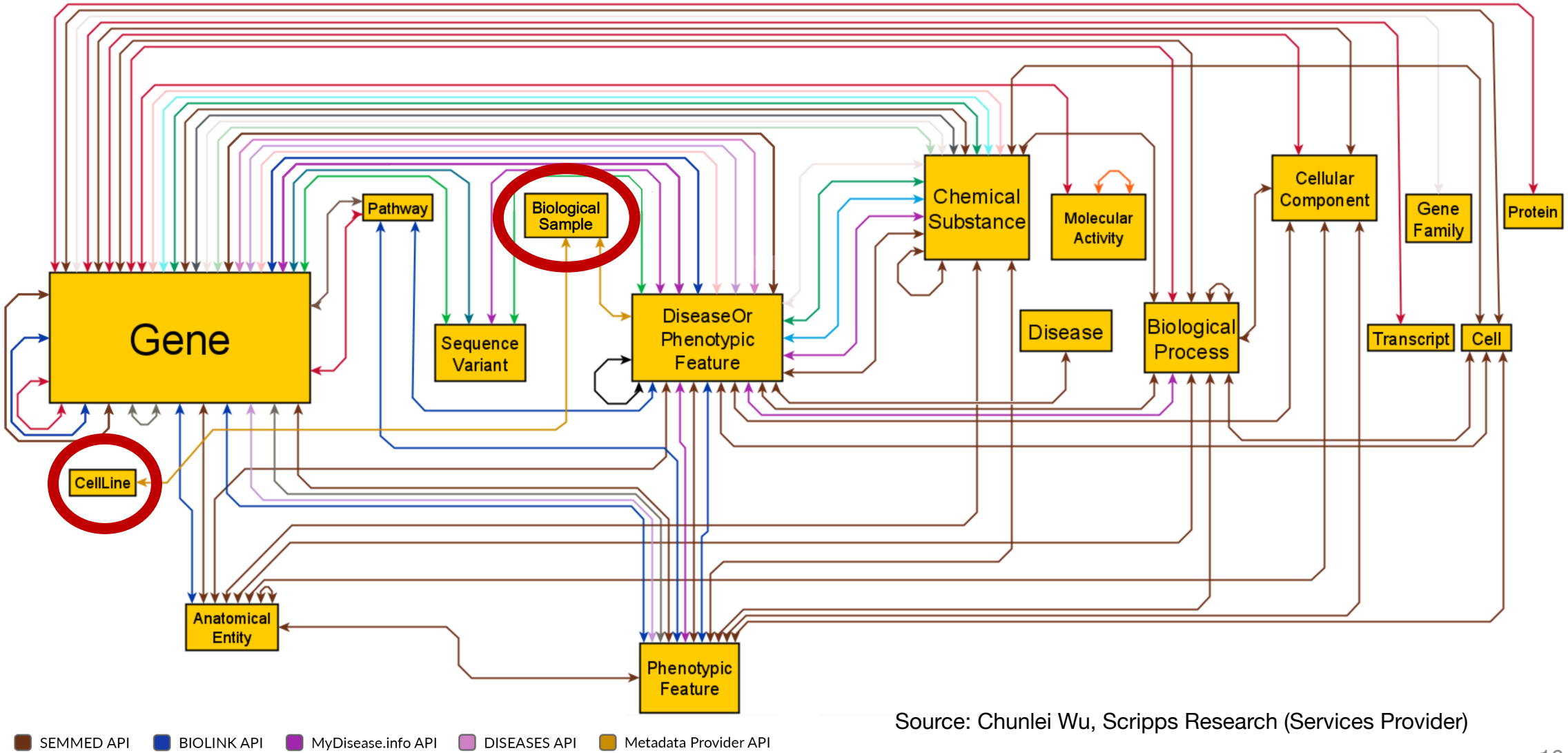


BioThings Explorer Knowledge Graph



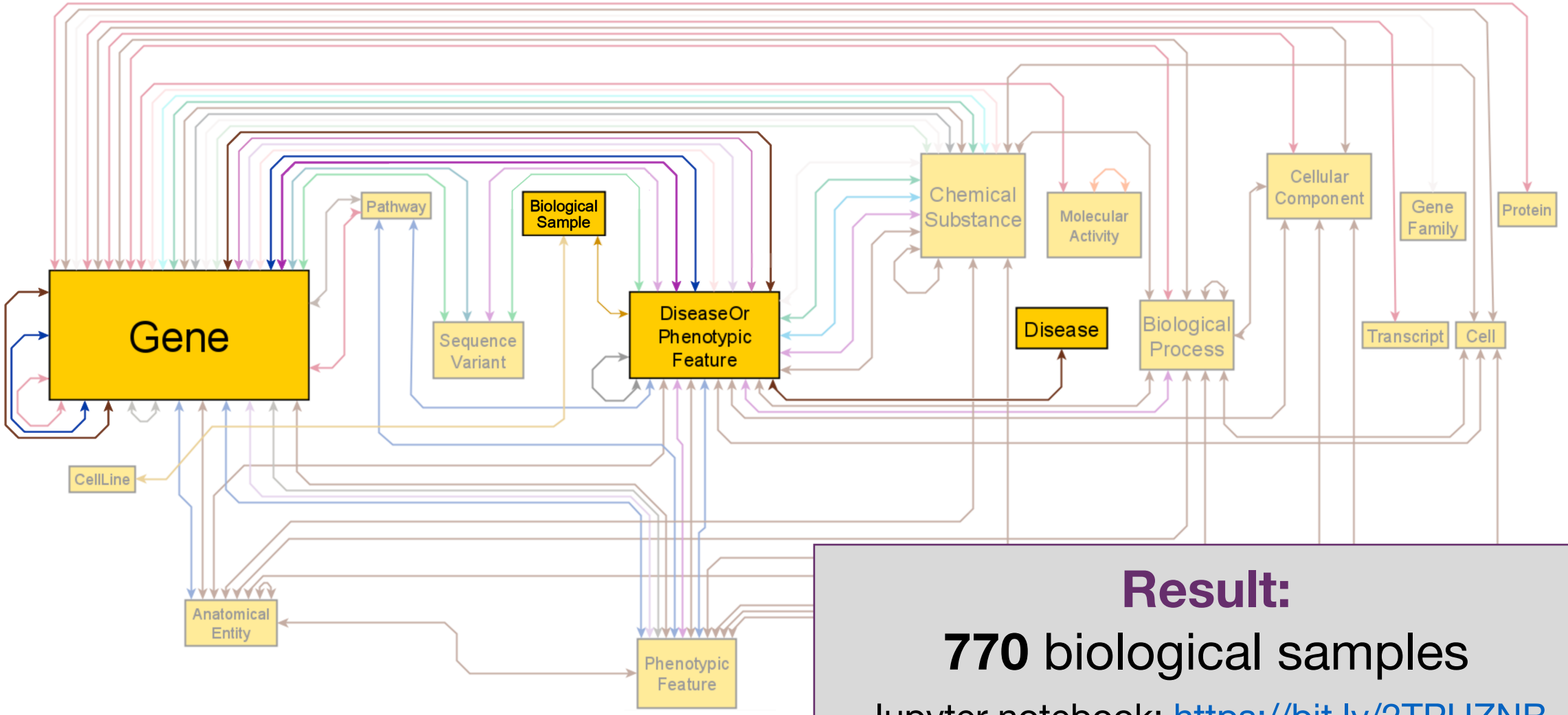
Source: Chunlei Wu, Scripps Research (Services Provider)

BioThings Explorer Knowledge Graph



Source: Chunlei Wu, Scripps Research (Services Provider)

What biological samples are associated with diseases related to gene SLC15A4 ?



Result:

770 biological samples

Jupyter notebook: https://bit.ly/2TPUZNB

Integration with BioThings Explorer - https://bit.ly/2TPUZNB

biothings / biothings_explorer

Unwatch 3 | Star 1 | Fork 4

Code | Issues 17 | Pull requests 3 | Actions | Projects 0 | Wiki | Security | Insights

Branch: master | biothings_explorer / jupyter notebooks / Demo of Integrating Stanford BioSample API into BTE.ipynb | Find file | Copy path

kevinxin90 add demo for integrating stanford biosample api d139872 2 days ago

1 contributor

599 lines (599 sloc) | 25.6 KB | Raw | Blame | History

Introduction

This notebook demonstrates how BioThings Explorer can be used to answer the following query:

"*What biosamples are associated with diseases related to gene SLC15A4*"

Background: BioThings Explorer can answer two classes of queries -- "EXPLAIN" and "PREDICT". EXPLAIN queries are described in EXPLAIN_demo.ipynb, and PREDICT queries are described in PREDICT_demo.ipynb. Here, we describe PREDICT queries and how to use BioThings Explorer to execute them. A more detailed overview of the BioThings Explorer systems is provided in these slides.

An Automated Pipeline to Enhance Metadata for Use by Translator

- **Milestone 1:** Develop software to annotate metadata attribute names with ontology terms
- **Milestone 2:** Develop software to annotate metadata attribute values with ontology terms
- **Milestone 3:** Apply our prototype software to a subset of NCBI BioSample metadata records
- **Milestone 4:** Make processed BioSample metadata available to Translator
- **Milestone 5:** **Demonstrate enhanced query capabilities made possible by our work**

Metadata Provider *prototype*

Database: NCBI BioSample Extract (4,346 samples from Homo sapiens)

Example 9 (Systemic lupus erythematosus)

I need to find information about biological samples in the setting of systemic lupus erythematosus.

Original BioSample Metadata

disease=systemic lupus erythematosus

Search on BioSample's website

Samples	Projects	Centers	Tissues	Cell Types	Cell Lines
518	4	4	2	9	1

Processed Metadata

disease=systemic lupus erythematosus

Samples	Projects	Centers	Tissues	Cell Types	Cell Lines
770	11	11	3	11	1

Next Steps

- **Collaboration**
  - Elucidate use cases to help select future metadata and data resources
  - Help to extend BioLink model to handle more clinical and translational data
  - Assist Translator community in use of our techniques
- **Stanford-Based Research**
  - Extend software to process additional metadata and data resources
  - Expand our architecture to include
    - More extensive semantic and lexical capabilities
    - Transformation of *queries* using standard ontologies

Standards Compliance

- Published all code and documents on **GitHub**
- Documented API in the **SmartAPI Registry**
- Adopted **BioLink model** where possible
- Used **BioPortal API** to access standard biomedical ontologies
- Pending: Conform with emerging Reasoner Standard API

Metadata Provider – Public Links

- User Interface: http://kp.metadatacenter.org
- API: http://api.kp.metadatacenter.org
- SmartAPI: http://smart-api.info/ui/4692da88e681a6b23e1ea9ed2152bd85
- Jupyter Notebook describing Annotation Pipeline: https://github.com/metadatacenter/metadata-provider/blob/master/metadata-provider-annotator/translator-demo.ipynb
- Jupyter Notebook demonstrating integration with BioThings Explorer: https://bit.ly/2TPUZNB
- GitHub repo: https://github.com/metadatacenter/metadata-provider

Metadata Provider