GazOntology

A Tool for Building GATE Gazetteer Lists from Ontologies

Daniel R. Schlegel, Rose Fontana, Adrian Naaktgeboren

Department of Computer Science

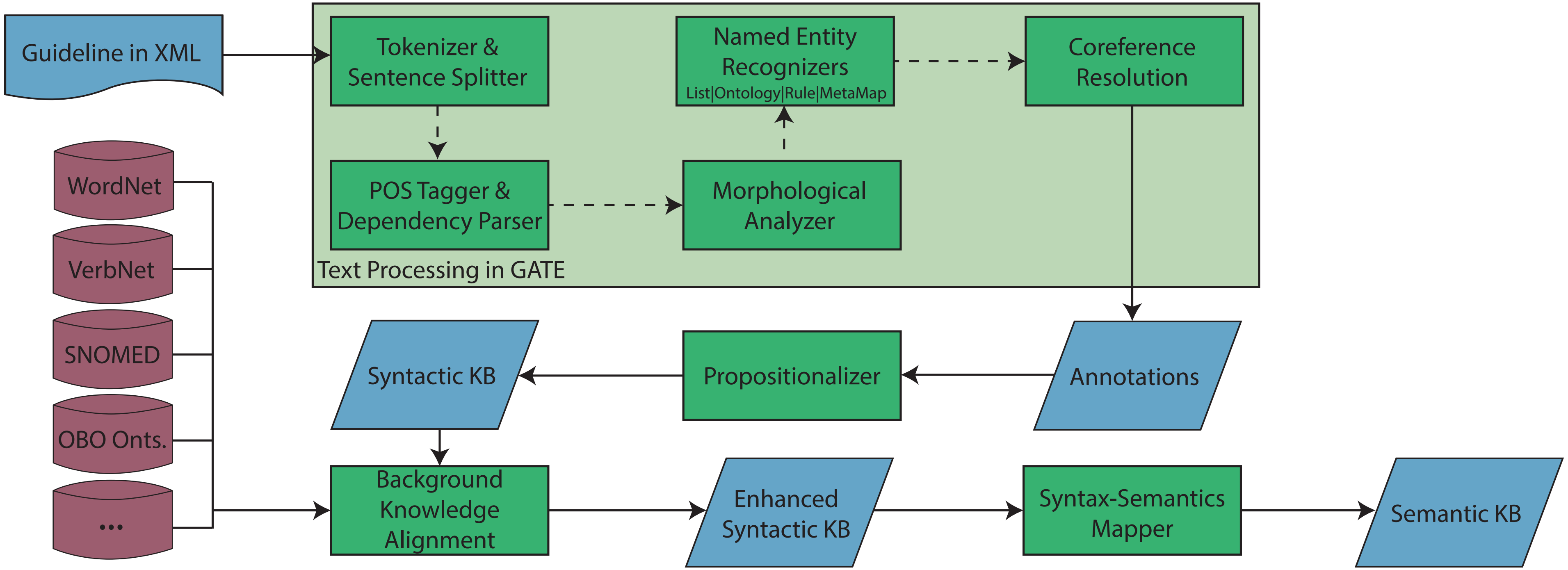
OSWEGO
STATE UNIVERSITY OF NEW YORK

Outline

- Motivation: Our Application and Philosophy
- GATE and ‘Gazetteer’ Lists
- Two Examples
- Our Application, Revisited

Clinical Tractor

- A tool for natural language understanding of clinical practice guidelines
  - Under development
  - Goal: represent the semantics of guidelines for the purpose of automatically generating computer interpretable guidelines.



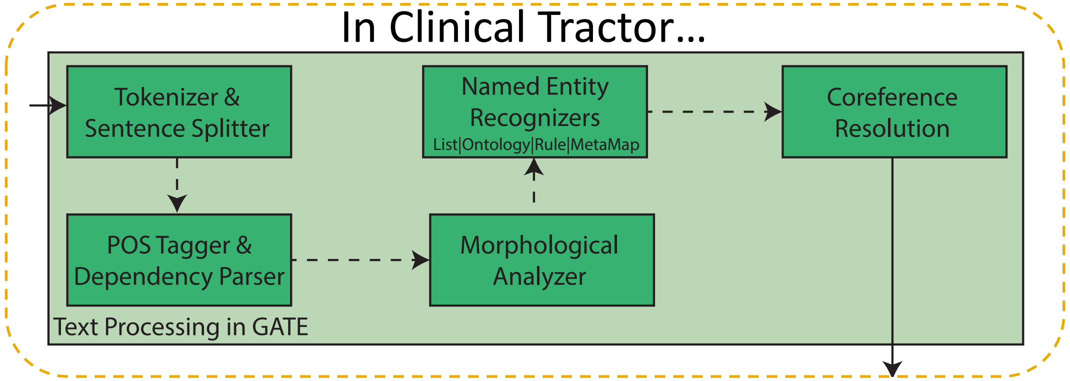
Our Natural Language Understanding Strategy

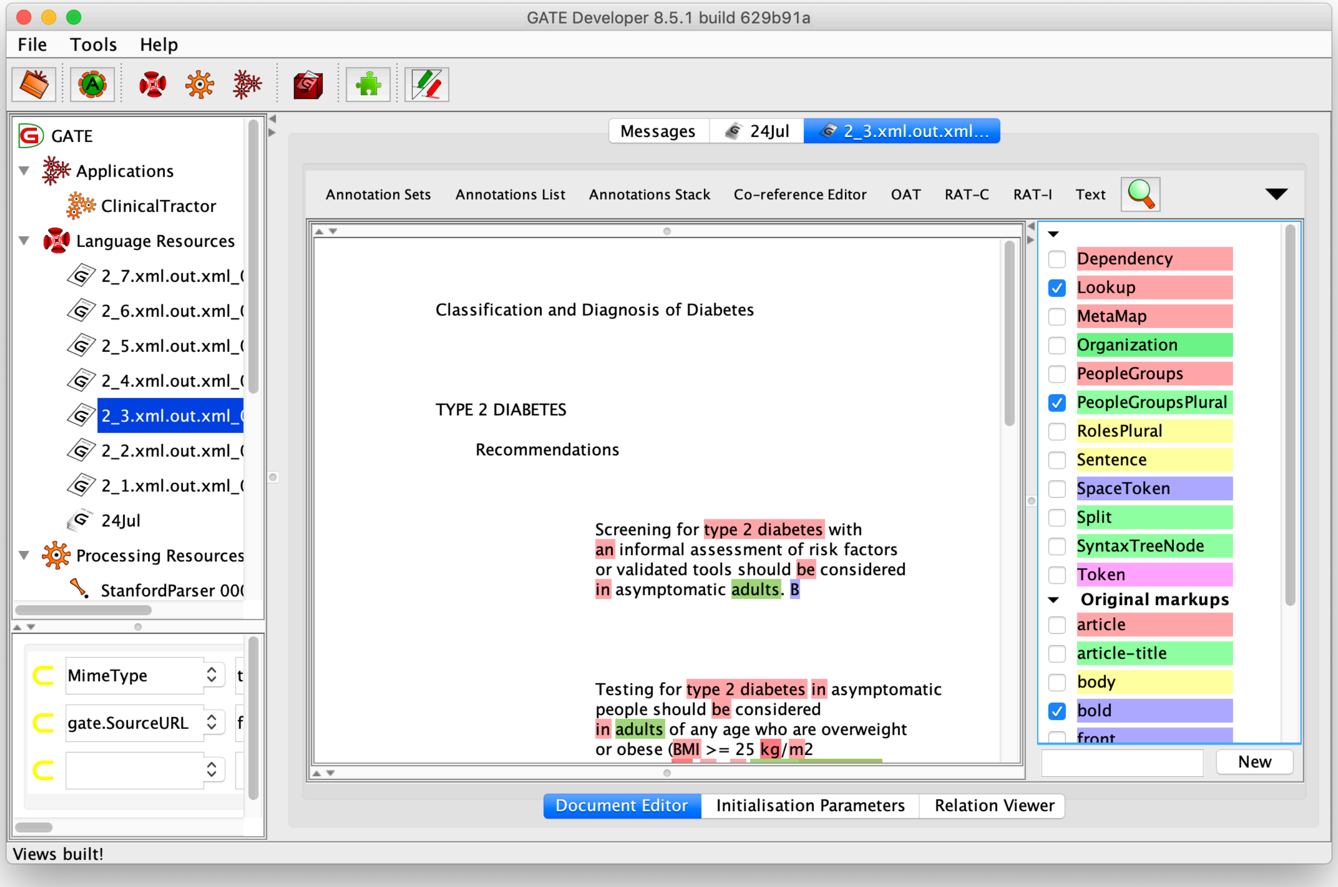
- Use off-the-shelf tools and data for the standard NLP pipeline
  - Make customizations as needed for the domain
- When mapping syntax to semantics:
  - Be as general as possible
  - Avoid specific cases wherever possible
  - Make judicious use of background knowledge to maintain generality

We use GATE to coordinate the standard NLP pipeline. We want it to provide us “hooks” into background knowledge we can use later.

GATE: General Architecture for Text Engineering

- Packages several tools for NLP
  - Notably for us today, named entity recognition tools
- Easy to use interface
  - But can run files in batch
- Extendable to add additional *processing resources*


In Clinical Tractor…



Classification and Diagnosis of Diabetes

TYPE 2 DIABETES

Recommendations

Screening for type 2 diabetes with an informal assessment of risk factors or validated tools should be considered in asymptomatic adults. B

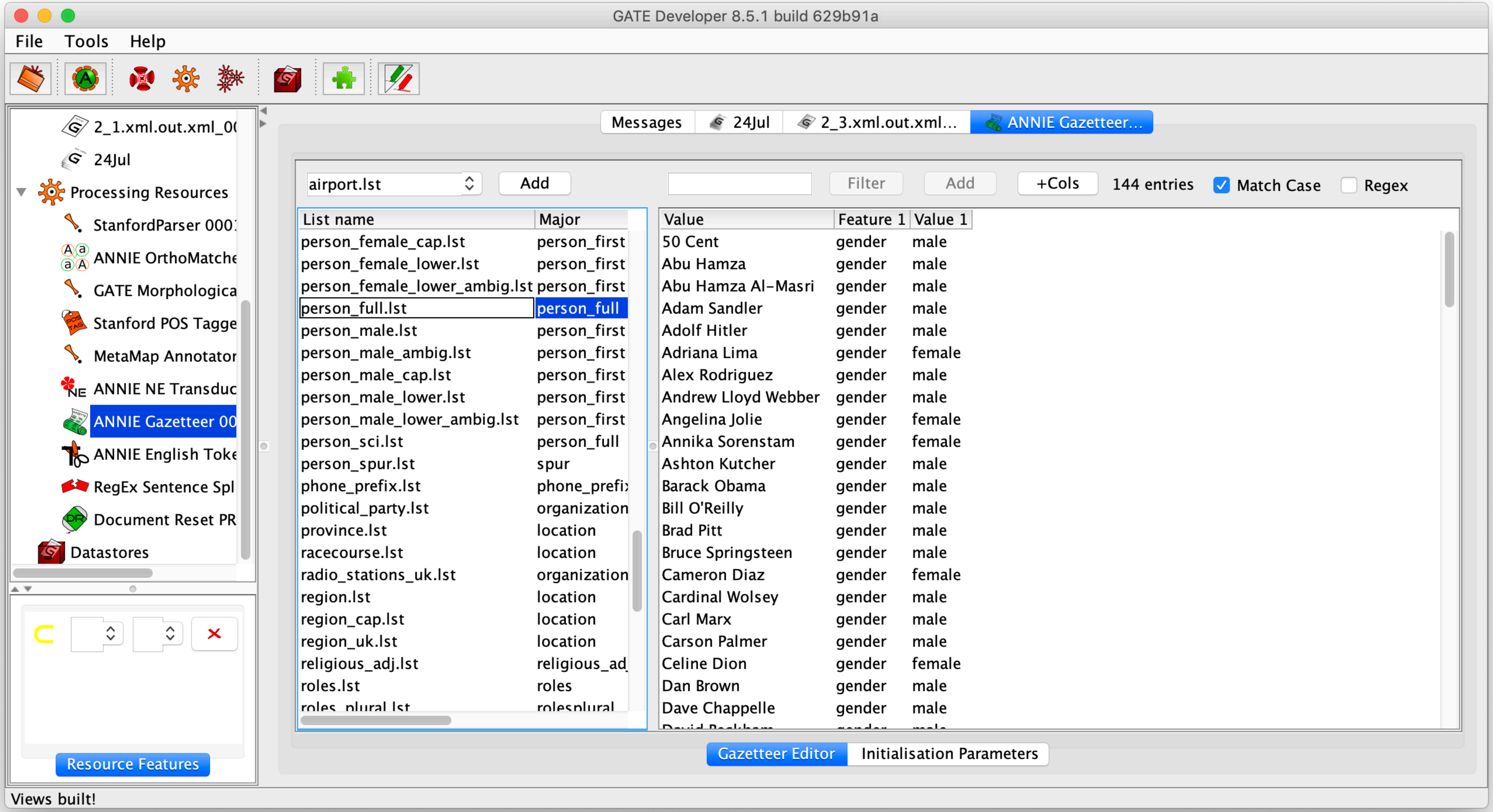
Testing for type 2 diabetes in asymptomatic people should be considered in adults of any age who are overweight or obese (BMI >= 25 kg/m2

□ Dependency
☑ Lookup
□ MetaMap
□ Organization
□ PeopleGroups
☑ PeopleGroupsPlural
□ RolesPlural
□ Sentence
□ SpaceToken
□ Split
□ SyntaxTreeNode
□ Token

Original markups

□ article
□ article-title
□ body
☑ bold
□ front

Gazetteer Lists



airport.lst

144 entries ☑ Match Case □ Regex

List name	Major
person_female_cap.lst	person_first
person_female_lower.lst	person_first
person_female_lower_ambig.lst	person_first
person_full.lst	person_full
person_male.lst	person_first
person_male_ambig.lst	person_first
person_male_cap.lst	person_first
person_male_lower.lst	person_first
person_male_lower_ambig.lst	person_first
person_sci.lst	person_full
person_spur.lst	spur
phone_prefix.lst	phone_prefi:
political_party.lst	organization
province.lst	location
racecourse.lst	location
radio_stations_uk.lst	organization
region.lst	location
region_cap.lst	location
region_uk.lst	location
religious_adj.lst	religious_ad
roles.lst	roles
roles_plural.lst	rolesplural

Value	Feature 1	Value 1
50 Cent	gender	male
Abu Hamza	gender	male
Abu Hamza Al-Masri	gender	male
Adam Sandler	gender	male
Adolf Hitler	gender	male
Adriana Lima	gender	female
Alex Rodriguez	gender	male
Andrew Lloyd Webber	gender	male
Angelina Jolie	gender	female
Annika Sorenstam	gender	female
Ashton Kutcher	gender	male
Barack Obama	gender	male
Bill O'Reilly	gender	male
Brad Pitt	gender	male
Bruce Springsteen	gender	male
Cameron Diaz	gender	female
Cardinal Wolsey	gender	male
Carl Marx	gender	male
Carson Palmer	gender	male
Celine Dion	gender	female
Dan Brown	gender	male
Dave Chappelle	gender	male

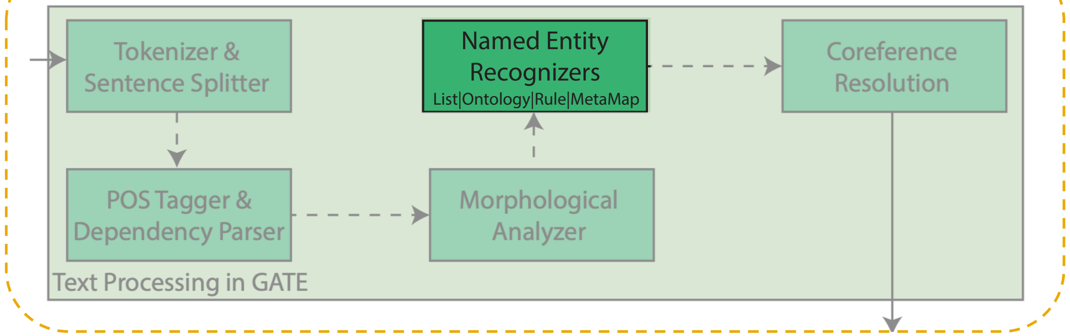
Views built!

GazOntology

- Benefit of biomedical domain: lots of ontologies!
  - How can we leverage them in GATE?

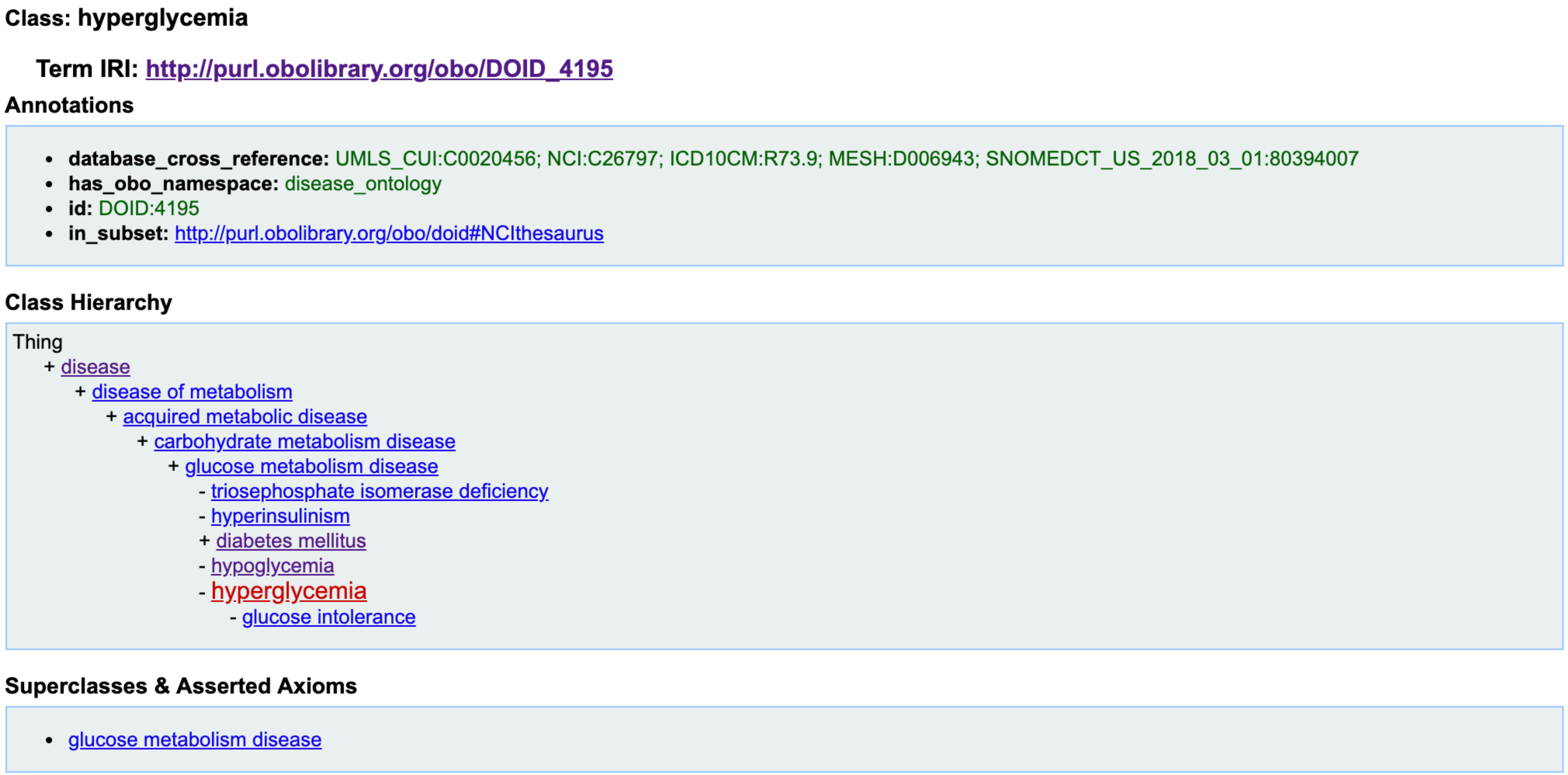
- Build Gazetteer lists from ontology terms
  - Use labels and synonyms
  - Retain mapping to IRI

In Clinical Tractor…



A First Example…

“Blood glucose rather than A1C should be used to diagnose the acute onset of type 1 diabetes in individuals with symptoms of hyperglycemia.” –Standards of Care in Diabetes 2017



Class: hyperglycemia

Term IRI: http://purl.obolibrary.org/obo/DOID_4195

Annotations

- **database_cross_reference:** UMLS_CUI:C0020456; NCI:C26797; ICD10CM:R73.9; MESH:D006943; SNOMEDCT_US_2018_03_01:80394007
- **has_obo_namespace:** disease_ontology
- **id:** DOID:4195
- **in_subset:** http://purl.obolibrary.org/obo/doid#NCIthesaurus

Class Hierarchy

```
Thing
  + disease
    + disease of metabolism
      + acquired metabolic disease
        + carbohydrate metabolism disease
          + glucose metabolism disease
            - triosephosphate isomerase deficiency
            - hyperinsulinism
            + diabetes mellitus
            - hypoglycemia
            - hyperglycemia
              - glucose intolerance
```

Superclasses & Asserted Axioms

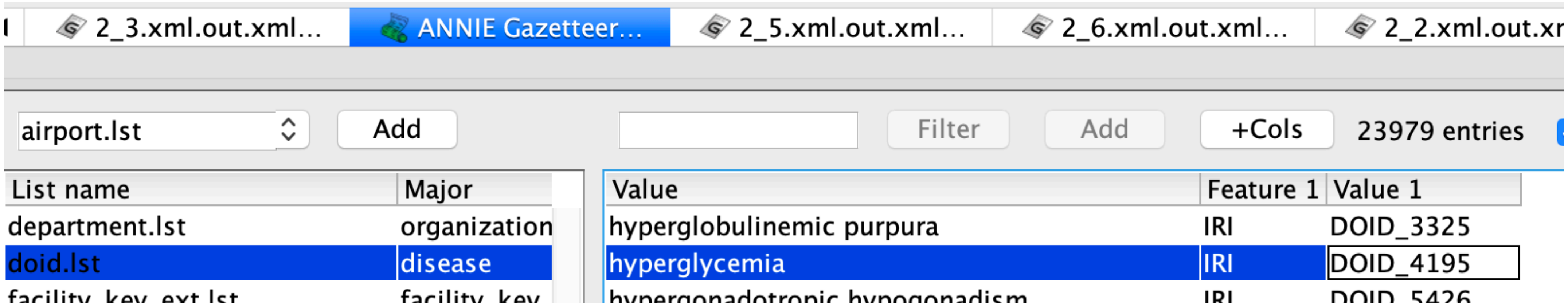
- glucose metabolism disease

Using GazOntology

- Download the tool and the ontology of interest
- Run the tool

```
java -jar GazOntology-1.0.1.jar -f doid.owl -o doid.lst
```

- Add the Gazetteer list to GATE



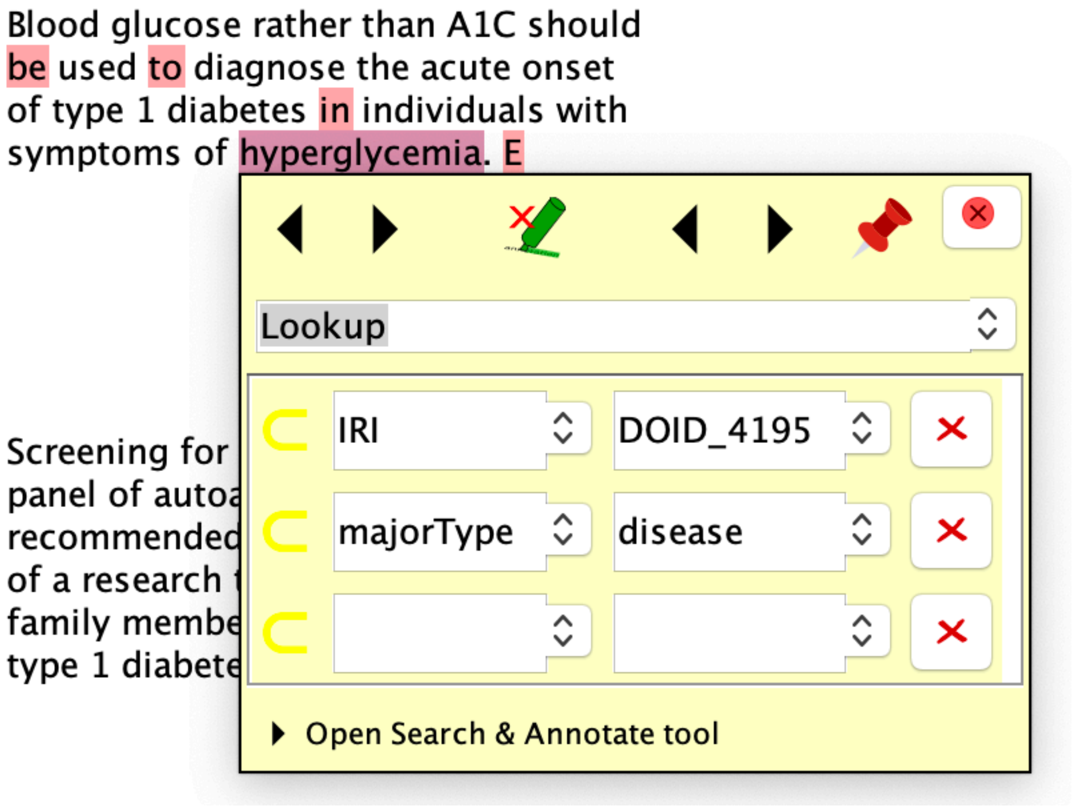
airport.lst

23979 entries

List name	Major
department.lst	organization
doid.lst	disease
facility_key_ext.lst	facility_key

Value	Feature 1	Value 1
hyperglobulinemic purpura	IRI	DOID_3325
hyperglycemia	IRI	DOID_4195
hypergonadotropic hypogonadism	IRI	DOID_5426

After Running GATE…



Blood glucose rather than A1C should be used to diagnose the acute onset of type 1 diabetes in individuals with symptoms of hyperglycemia. E

Lookup

IRI	DOID_4195
majorType	disease

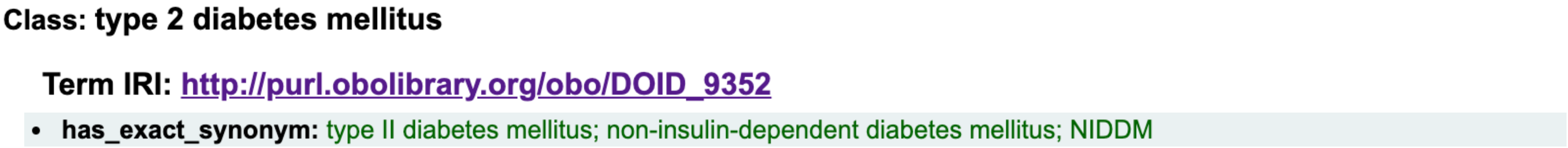
▸ Open Search & Annotate tool

Screening for
panel of autoa
recommended
of a research
family membe
type 1 diabete

One Advantage: Add Synonyms Easily

- Adding missing synonyms is simple in the GATE interface
  - Add them in the ontology Gazetter list directly, or make a new list for additions
  - Likely easier to update ontology versions this way than editing the ontology directly

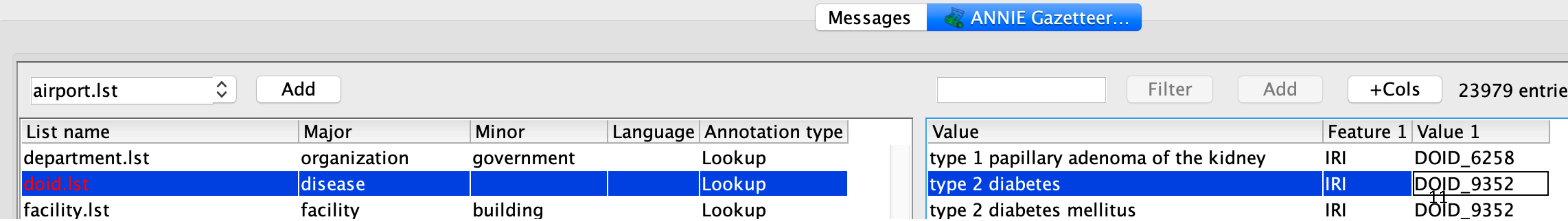
”Screening for type 2 diabetes with an informal assessment of risk factors or validated tools should be considered in asymptomatic adults."



Class: type 2 diabetes mellitus

Term IRI: http://purl.obolibrary.org/obo/DOID_9352

- **has_exact_synonym:** type II diabetes mellitus; non-insulin-dependent diabetes mellitus; NIDDM

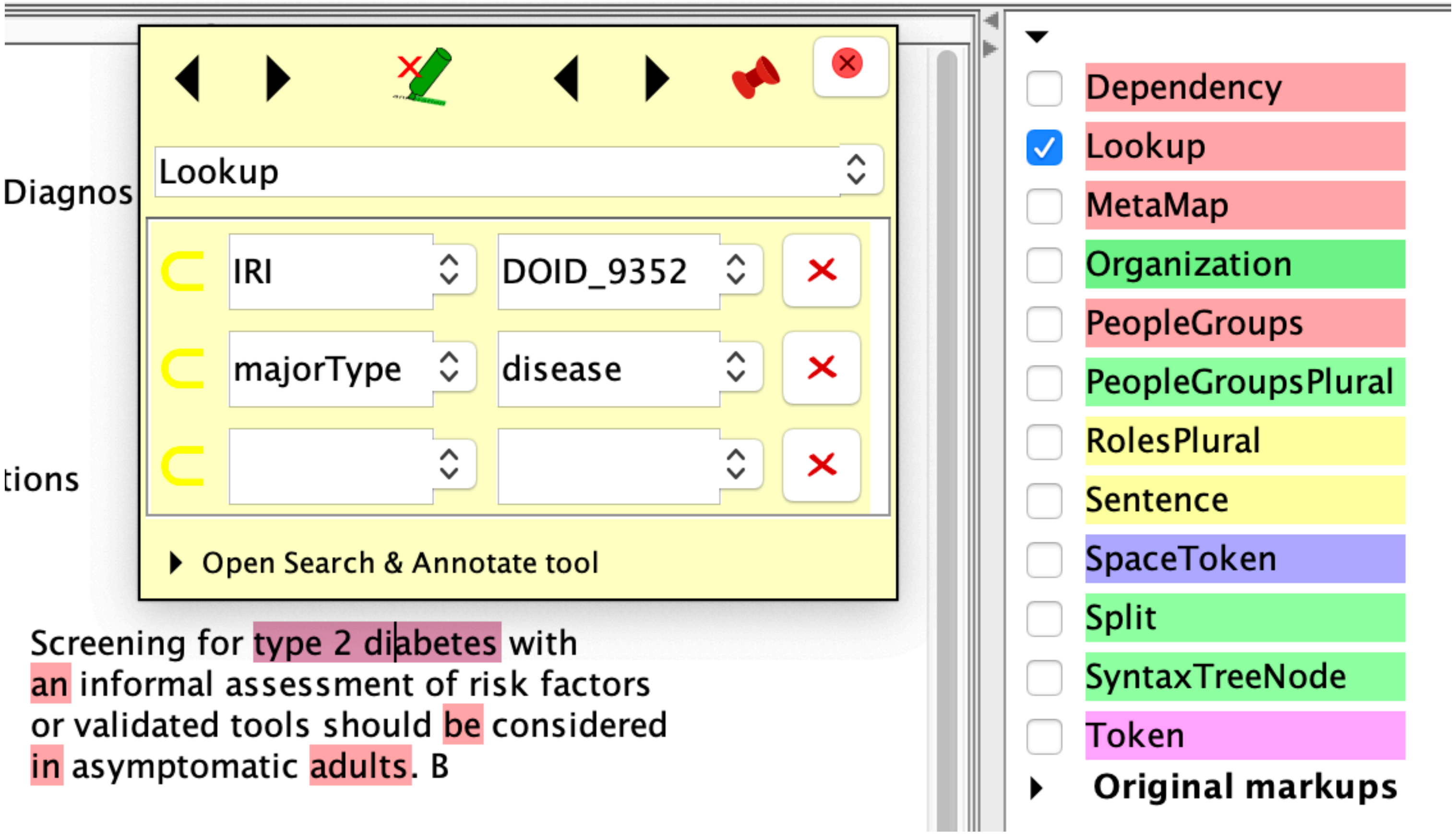


23979 entrie

List name	Major	Minor	Language	Annotation type
department.lst	organization	government		Lookup
doid.lst	disease			Lookup
facility.lst	facility	building		Lookup

Value	Feature 1	Value 1
type 1 papillary adenoma of the kidney	IRI	DOID_6258
type 2 diabetes	IRI	DOID_9352
type 2 diabetes mellitus	IRI	DOID_9352

Easy to Test: Just Re-run the Gazetteer



Diagnos

tions

Lookup

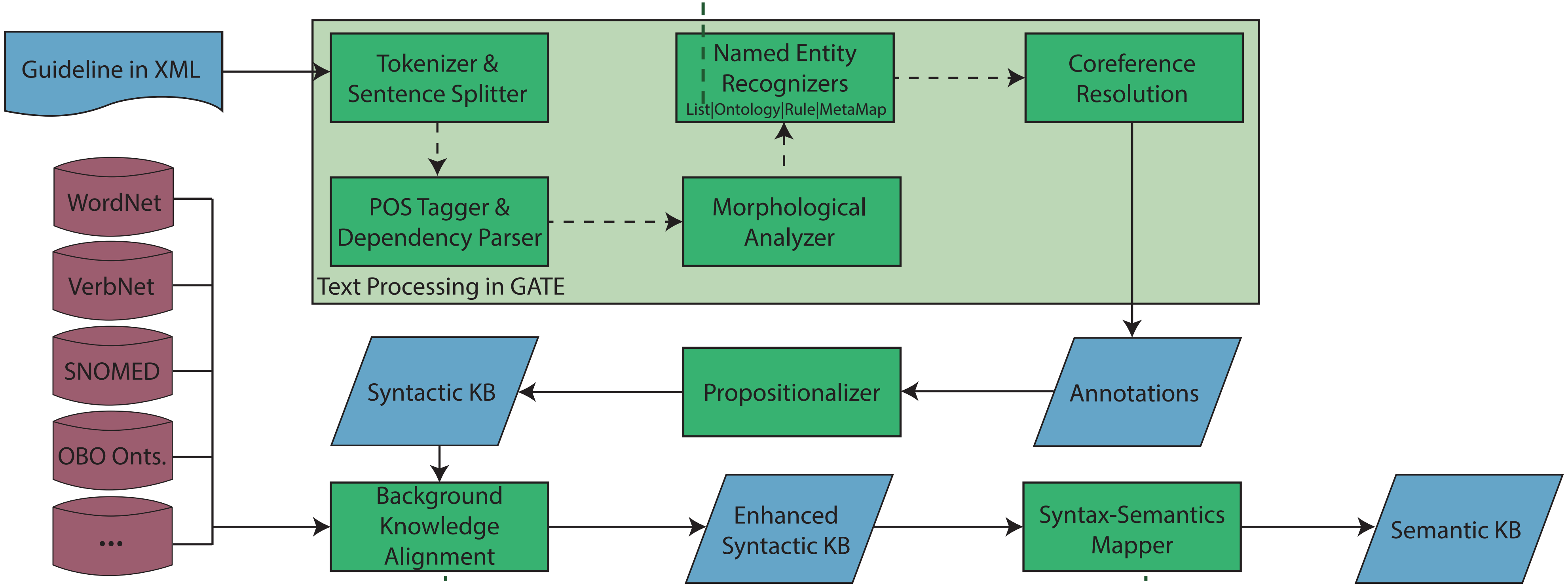
IRI	DOID_9352
majorType	disease

Screening for type 2 diabetes with an informal assessment of risk factors or validated tools should be considered in asymptomatic adults. B

- ☐ Dependency
- ☑ Lookup
- ☐ MetaMap
- ☐ Organization
- ☐ PeopleGroups
- ☐ PeopleGroupsPlural
- ☐ RolesPlural
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ SyntaxTreeNode
- ☐ Token

Original markups

Use in Clinical Tractor



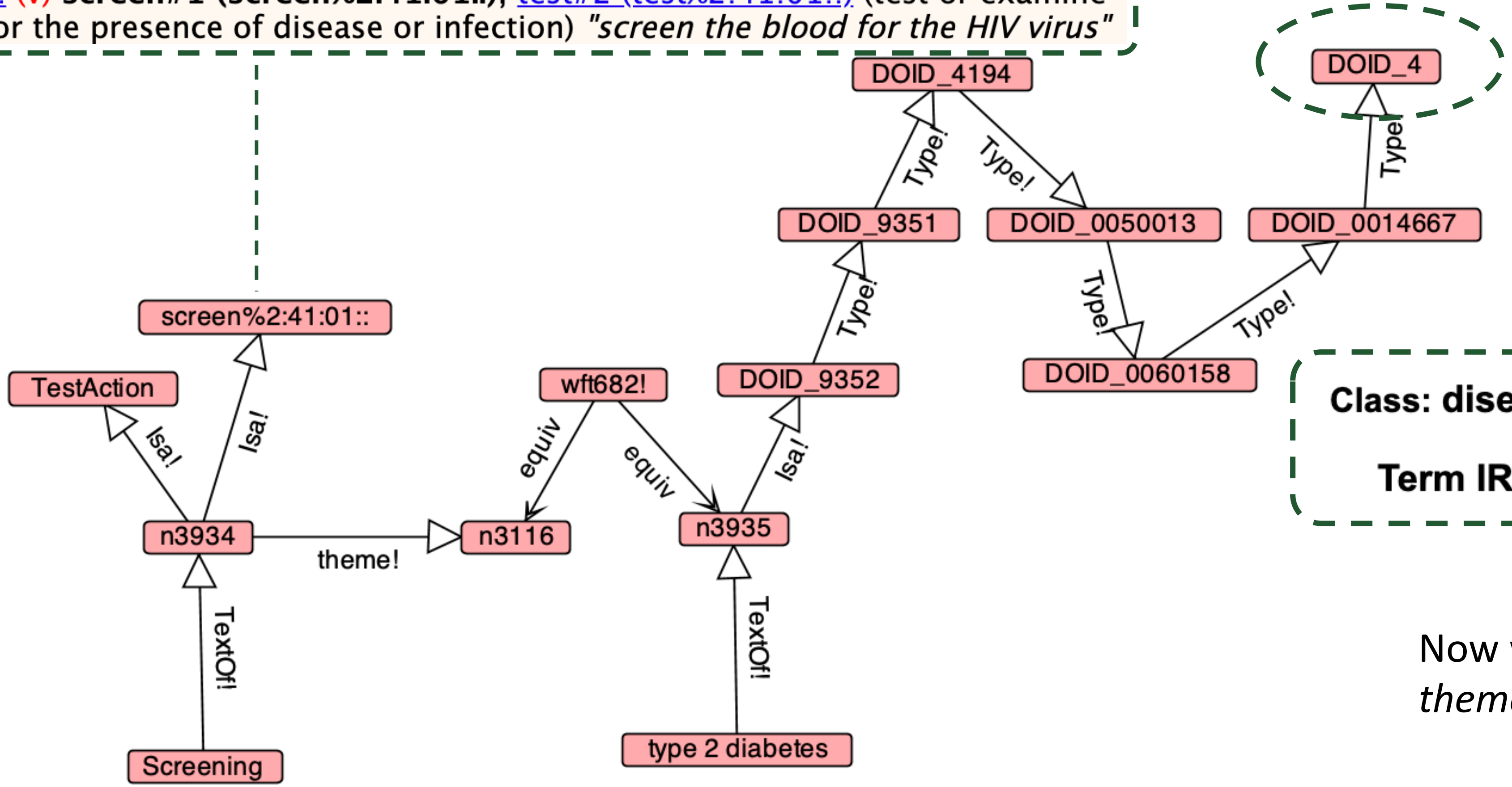
Our new lists live here

Using the IRI from GazOntology, import relevant portions of the ontology into the knowledge base

Use higher level terms for inference

In Clinical Tractor…

S: (v) **screen#1 (screen%2:41:01::)**, test#2 (test%2:41:01::) (test or examine for the presence of disease or infection) *"screen the blood for the HIV virus"*



Class: disease

Term IRI: http://purl.obolibrary.org/obo/DOID_4

Now we can write rules which rely on the *theme* of a *Testing* action being a *disease*.

"Screening for type 2 diabetes …"

More about GazOntology

- Open source: https://github.com/oswegonlu/GazOntology
- Several options:
  - Include / exclude broad or related synonyms
  - Include only a subhierarchy of the ontology
  - Include only a single level of the ontology
- We're open to adding features! Just submit a bug report!

Thank You!

GazOntology is available at:

https://github.com/oswegonlu/GazOntology

This work was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R15LM013030. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.