An Environment for Biomedical Text Mining: The LAPPS Grid

Nancy Ide • Department of Computer Science • Vassar College

Need for text mining

Number of new scientific publications is growing rapidly

New terms (genes, proteins, chemical compounds, drugs) are constantly created

Information is in textual form – unstructured data

Impossible to manage such an information overload

NLP to the rescue!

Typical Scenario

- A scientist wants to apply NLP techniques to find articles including references to certain entities (e.g., proteins, genes) and their interactions
    - Knows nothing about NLP or Computer Science
    - Unfamiliar with NLP technologies
- Searches for NLP software that might help

Typical Scenario

- Finds existing tools and frameworks that are freely available

Not to mention several commercial (i.e., pricey) options

- Questions
  - Do these things all do the same thing, or do they differ in some way?
  - Do some work better than others?
  - Are some easier to use than others?
  - How does one choose?

Confused scientist

NLP Tools for Biomedical texts

- General purpose NLP tools provide general support, not geared to BioNLP
- But there are many recently developed tools for biomedical texts; see, e.g., lists of tools at
  - http://bionlp.org
  - http://biocreative.sourceforge.net/bionlp_tools_links.html
  - http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7784950
  - http://www.nactem.ac.uk/index.php (but also lots of other things)

However…

- Many of these tools are difficult to install, configure, and use without some computational expertise
- Even more difficult to modify or adapt without computational expertise and some knowledge of NLP
- Also: which tools performing the same task perform best and/or are best suited to a given task?

Another Sneaky Underlying Problem

- Some applications are “all-in-one” black boxes
- But often desirable to compose your own application from independent modules
  - Customize certain modules
  - Test the effect of different modules on the quality of results
  - Experiment with different resources (e.g. lexicons used by NER module)
- **PROBLEM: Input and output of tools from different sources differ dramatically!!!**
  - I.e., tools are not *interoperable*
  - Often demands significant effort and expertise to adapt tools from different sources to work together (if possible at all)

What is Needed

1

Develop/ provide access to a range of freely available advanced text mining tools specially tailored to scientific publications

2

Enable scientists to easily use these tools without having to be a computer scientist or an expert in NLP

Interoperability is key!

3

Enable scientists to easily adapt existing solutions to specific domains or problems without having to be a computer scientist or an expert in NLP

Tools are Not Enough

BioNLP also needs language resources:

- Large bodies of scientific publications that can be searched and mined for information and knowledge
- Large bodies of annotated scientific publications that can be used to develop language models (e.g. via machine learning)
- Lexicons and Ontologies of biomedical terms to assist in entity recognition etc.

However…

- The same interoperability problem exists for resources!
  - Different physical formats
    - PDF, XML, plain text…
  - Different representations for annotations
    - Different physical formats
      - XML, JSON, brackets, BIO
    - Different terminologies

Enter…

The Language Applications (LAPPS) Grid

Nancy Ide, Keith Suderman

Vassar College

James Pustejovsky, Marc Verhagen

Brandeis University

Christopher Cieri, Denise DiPersio, Jonathan Wright

Linguistic Data Consortium (Penn)

Eric Nyberg, Di Wang

Carnegie Mellon University

What is the LAPPS Grid?

- US National Science Foundation-funded project
- Collaborative among Vassar College, Brandeis University, University of Pennsylvania, and Carnegie Mellon University
- Goal: Provide an infrastructure that facilitates
  - Retrieving large text collections from providers and repositories
  - Devising pipelines (workflows) of interoperable web services that automatically annotate data, provide evaluation metrics for the results, etc.
  - Saving, storing, and sharing pipelines and results for later use by yourself or others
  - Is fully open for any use

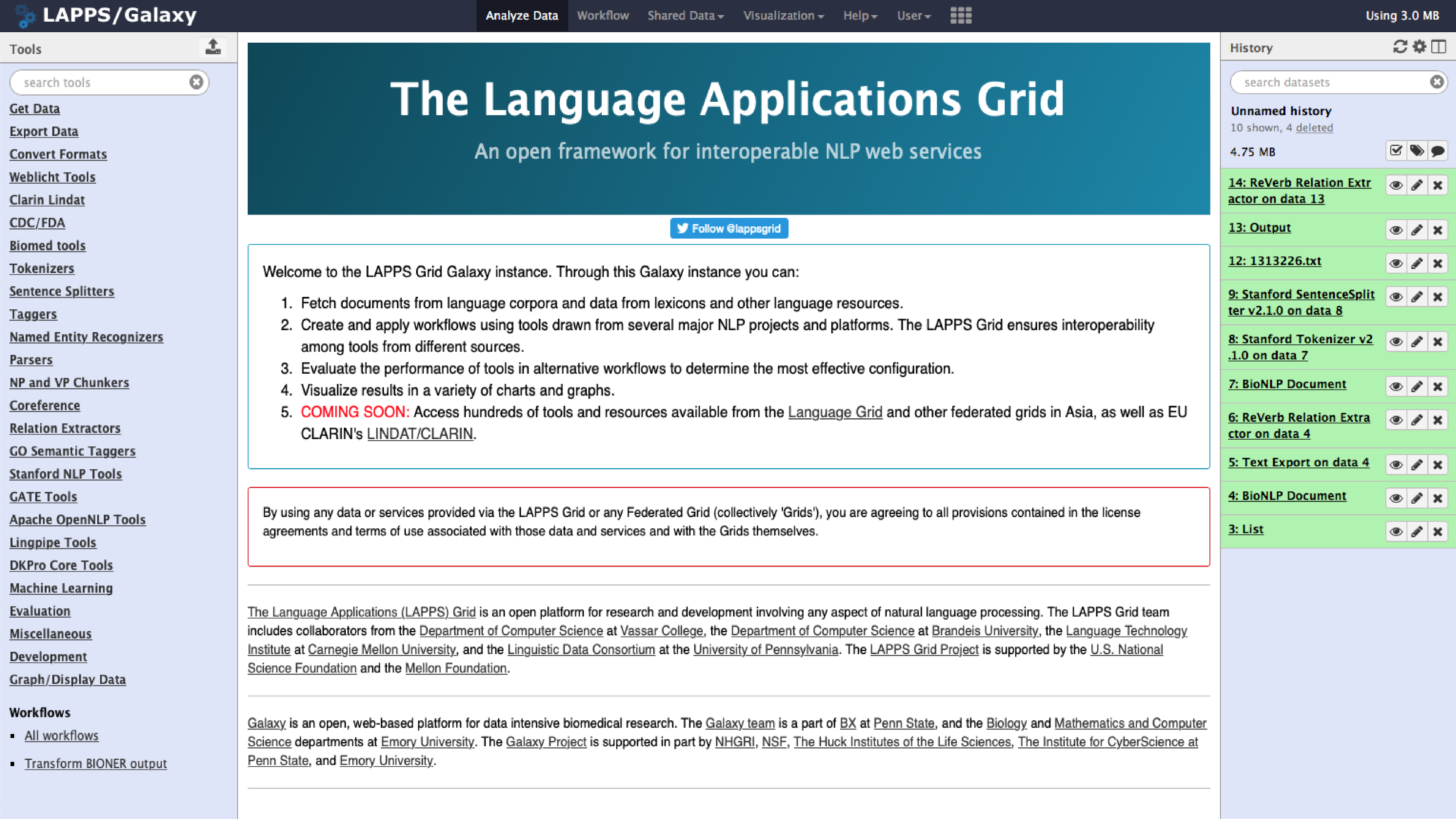
LAPPS/Galaxy Interface

The LAPPS Grid uses the GALAXY framework as a workflow engine to combine services of the Language Application Grid

Galaxy

http://galaxyproject.org

- Galaxy is an open, web-based platform designed primarily for computational genomics research
  - Accessible
    - Users without programming experience can easily specify parameters and run tools and workflows
  - Reproducible
    - Captures information so that any user can repeat and understand a complete computational analysis
  - Transparent
    - Users can share and publish analyses and workflows via the web and create interactive, web-based documents that describe a complete analysis



The Language Applications Grid

An open framework for interoperable NLP web services

Welcome to the LAPPS Grid Galaxy instance. Through this Galaxy instance you can:

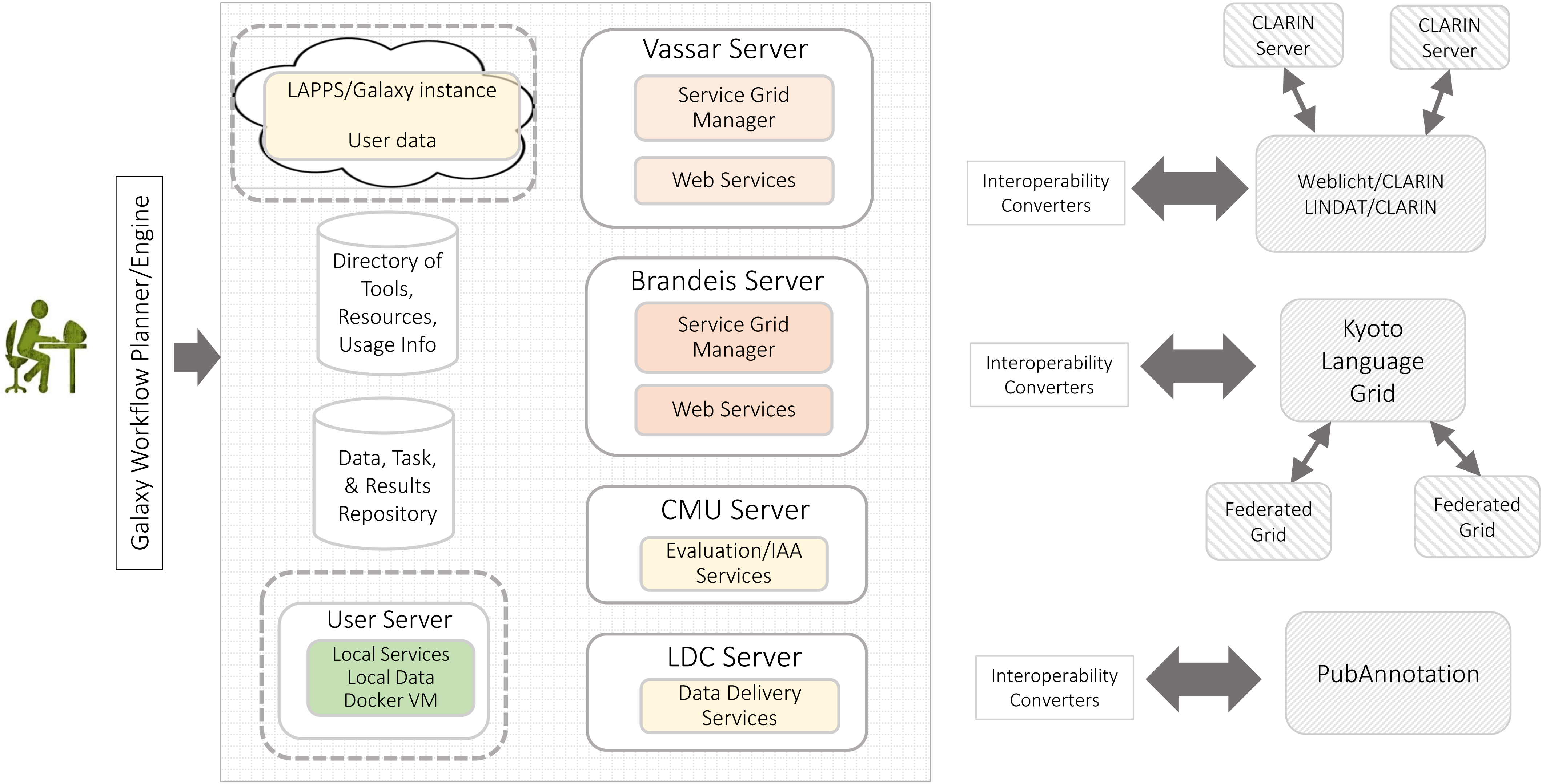
1. Fetch documents from language corpora and data from lexicons and other language resources.
2. Create and apply workflows using tools drawn from several major NLP projects and platforms. The LAPPS Grid ensures interoperability among tools from different sources.
3. Evaluate the performance of tools in alternative workflows to determine the most effective configuration.
4. Visualize results in a variety of charts and graphs.
5. COMING SOON: Access hundreds of tools and resources available from the Language Grid and other federated grids in Asia, as well as EU CLARIN's LINDAT/CLARIN.

By using any data or services provided via the LAPPS Grid or any Federated Grid (collectively 'Grids'), you are agreeing to all provisions contained in the license agreements and terms of use associated with those data and services and with the Grids themselves.

The Language Applications (LAPPS) Grid is an open platform for research and development involving any aspect of natural language processing. The LAPPS Grid team includes collaborators from the Department of Computer Science at Vassar College, the Department of Computer Science at Brandeis University, the Language Technology Institute at Carnegie Mellon University, and the Linguistic Data Consortium at the University of Pennsylvania. The LAPPS Grid Project is supported by the U.S. National Science Foundation and the Mellon Foundation.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

LAPPS Grid Overview

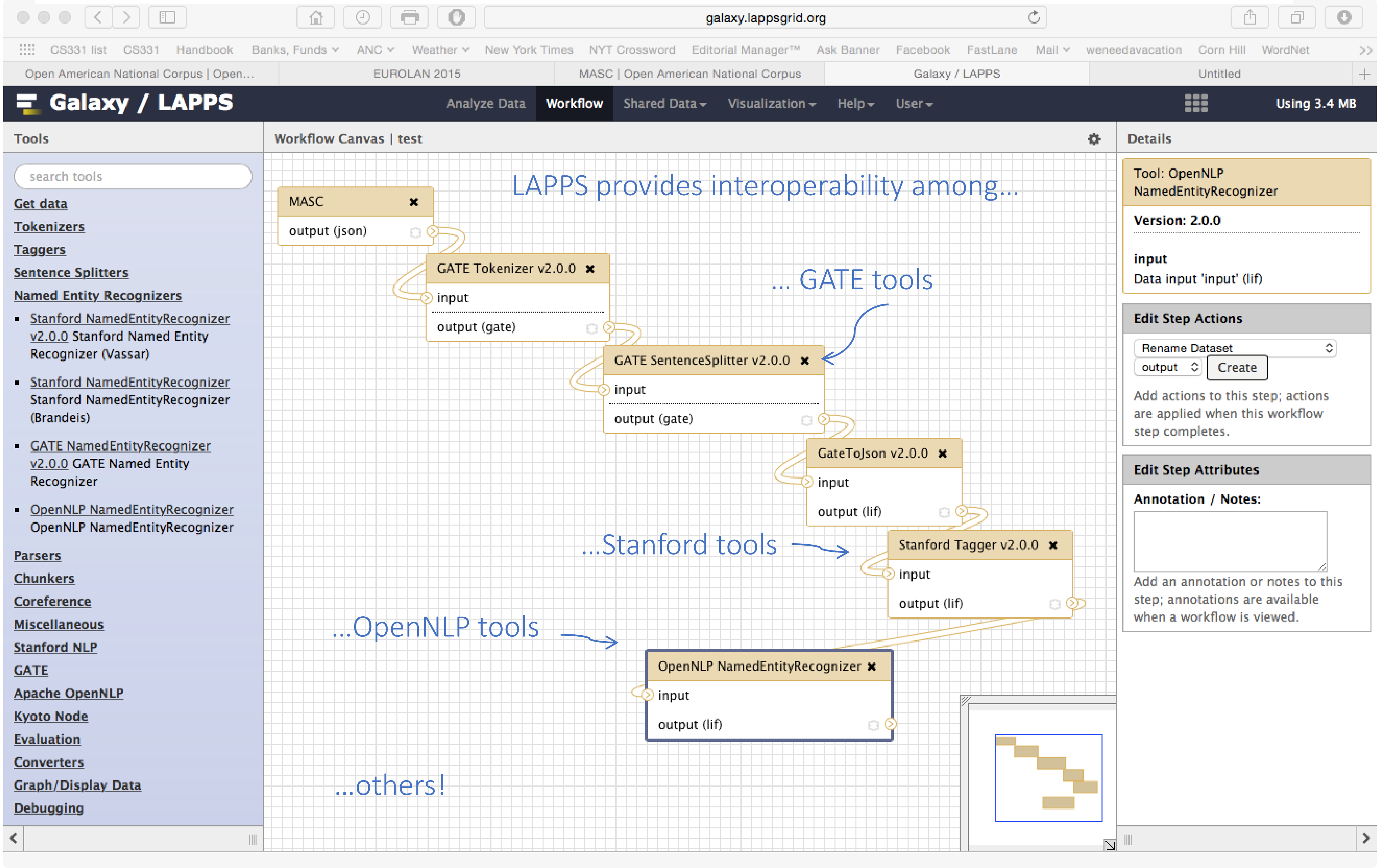


LAPPS/GALAXY

Multiple options for running a LAPPS/Galaxy instance:

1. Use the LAPPS/Galaxy web interface
   - https://galaxy.lappsgrid.org
2. Create a local Galaxy instance including:
   - All of Galaxy, or
   - The Galaxy “NLP Flavor” with only LAPPS tools
3. Create a docker image that is a self-contained vm running LAPPS/Galaxy
   - Useful when privacy required, no network connection available, etc.
4. Create a Galaxy instance in the cloud
   - Useful for large datasets, computationally intense applications
   - https://jetstream.lappsgrid.org

Workflow construction



Tools

search tools

- Get data
- Tokenizers
- Taggers
- Sentence Splitters
- Named Entity Recognizers
  - Stanford NamedEntityRecognizer v2.0.0 Stanford Named Entity Recognizer (Vassar)
  - Stanford NamedEntityRecognizer Stanford NamedEntityRecognizer (Brandeis)
  - GATE NamedEntityRecognizer v2.0.0 GATE Named Entity Recognizer
  - OpenNLP NamedEntityRecognizer OpenNLP NamedEntityRecognizer
- Parsers
- Chunkers
- Coreference
- Miscellaneous
- Stanford NLP
- GATE
- Apache OpenNLP
- Kyoto Node
- Evaluation
- Converters
- Graph/Display Data
- Debugging

Workflow Canvas | test

MASC
output (json)

GATE Tokenizer v2.0.0
input
output (gate)

GATE SentenceSplitter v2.0.0
input
output (gate)

GateToJson v2.0.0
input
output (lif)

Stanford Tagger v2.0.0
input
output (lif)

OpenNLP NamedEntityRecognizer
input
output (lif)

LAPPS provides interoperability among…

… GATE tools

…Stanford tools

…OpenNLP tools

…others!

Details

Tool: OpenNLP NamedEntityRecognizer

Version: 2.0.0

input
Data input 'input' (lif)

Edit Step Actions

Rename Dataset
output
Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Interoperability for language data

- Syntactic interoperability achieved with **common physical formats**
  - Many formats: One sentence per line, part-of-speech tag appended to word, XML, tab separated columns…
- Semantic interoperability achieved with **common definitions for labeled data**
  - E.g., labels like *noun*, *person, date* should mean the same to both systems
    - Not easy!
      - Subtle differences of opinion (e.g., should “in the future” be labeled as a DATE? Is “the White House” a LOCATION or an ORGANIZATION in a phrase like “The White House said today…”)?
      - Let alone that people do not agree on the exact definition of noun…

Obstacles

- Difficult to identify a single representation format that accommodates all kinds of language data and annotations
- Difficult to get the community to agree, adopt a single standard
- Need to accommodate legacy data and tools using other formats

Current solution

30 years of development have led to (reasonable) convergence of practice

- Key idea:
  - Instead of defining a single solution, design a universal “pivot” into and out of which other schemes can be easily mapped
  - For physical formats, requires that the pivot is a serialization of a common abstract data model (directed acyclic graph)
    - This model underlies UML, ER diagrams, RDF, JSON and JSON-LD, XML, semantic and other kinds of networks…
  - For semantics, provide a common structured set of terms to which other schemes can be mapped

Non-trivial!

How Does the LAPPS Grid Enable Interoperability?

- LAPPS Interchange Format (LIF)
  - Format that allows web services to exchange detailed information about data and its annotations
  - Syntactic interoperability
    - Handled by JSON-LD
    - Enforced by the LIF JSON schema
  - Semantic interoperability
    - Helped by using the Linked Data aspect of JSON-LD to link to the LAPPS Web Services Exchange Vocabulary

Web Service Communication in LAPPS

1

Each service in the LAPPS Grid publishes metadata:

- a discriminator (type) : tells how to interpret the payload
- a payload (typically a utf-8 string)

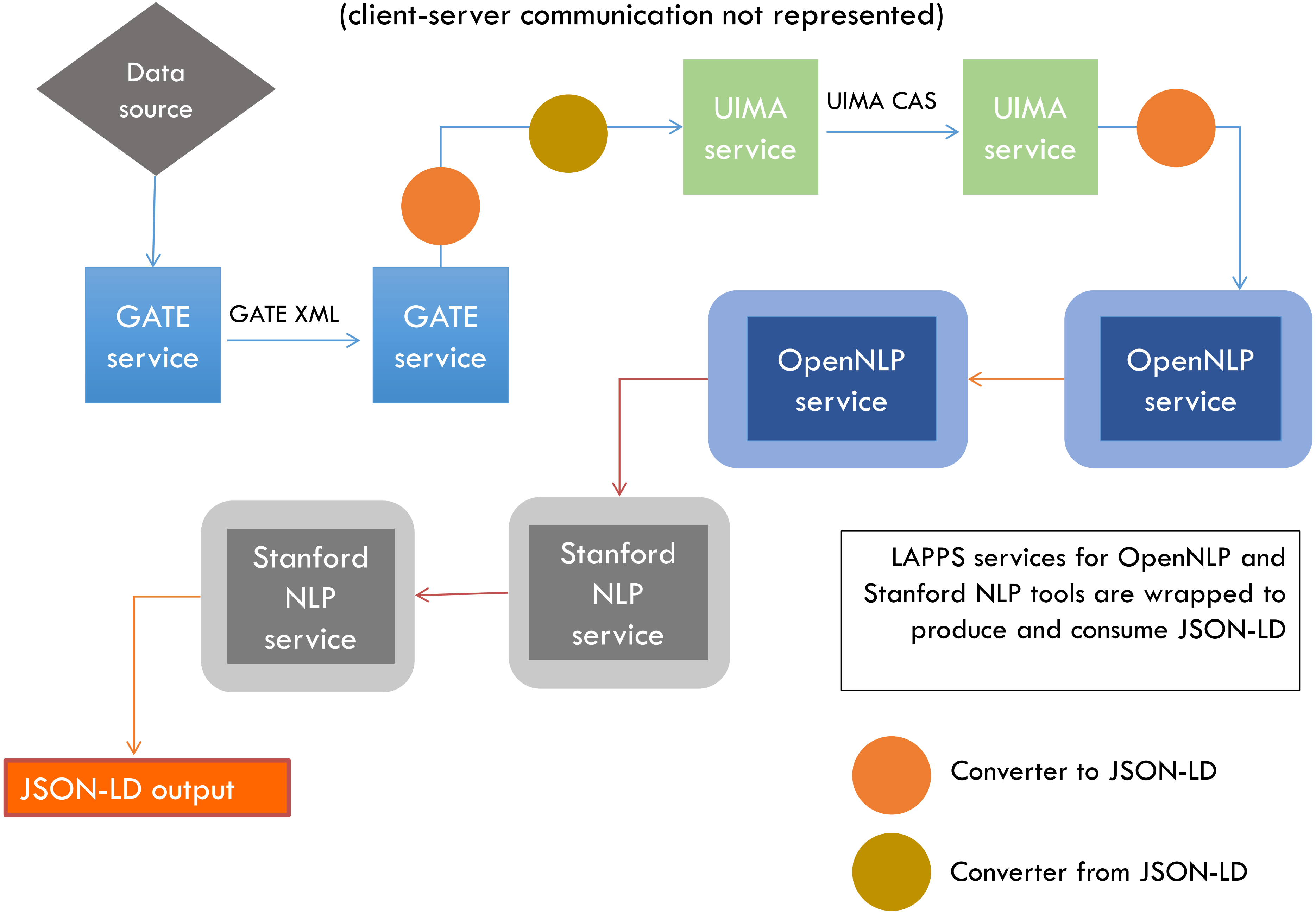
2

LAPPS uses JSON-LD as its standard format for the payload

- Converters to and from JSON-LD for services that deliver in other formats
- Some LAPPS services are wrapped to produce and consume JSON-LD

Logical flow

(client-server communication not represented)



LAPPS services for OpenNLP and Stanford NLP tools are wrapped to produce and consume JSON-LD

● Converter to JSON-LD

● Converter from JSON-LD

LAPPS Grid Web Service Exchange Vocabulary

- No accepted standard for module description or input/output interchange in the language application domain
- LAPPS Web Service Exchange Vocabulary (WS-EV)
  - Specifies a terminology for a core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data
  - Addresses a need within the community to identify a standard terminology and indicate the relations among them

Design Principles

01

Orthogonal design

- Only one entry per concept

02

Lightweight

- Easy to find on the web and reference

03

Flexible

- Use what you need, add what you need

04

(Arbitrary) decisions about what goes where

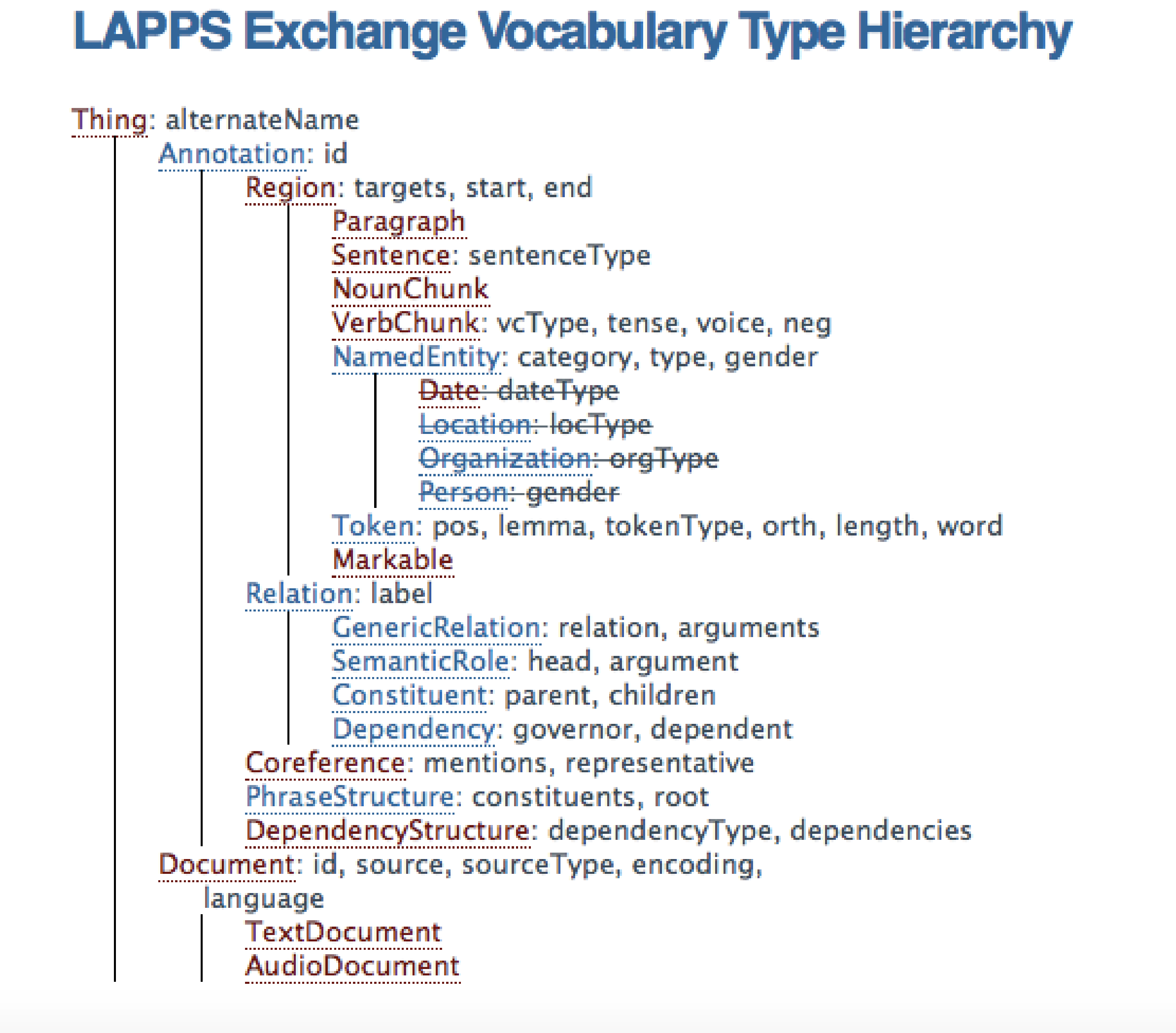
- Map to this for exchange only
- Not confined to the WS-EV terminology or organization internally

Implementation

- Bottom-up approach
  - Define objects and properties as needed to accommodate LAPPS services as they are added to the LAPPS Grid
  - Avoids *a priori* development of a comprehensive standard linguistic type system
  - “Minimalist” strategy to provide a simple core set of objects and features
  - User capacity to add/replace objects and properties to allow for dynamic typing

LAPPS WS-EV Repository

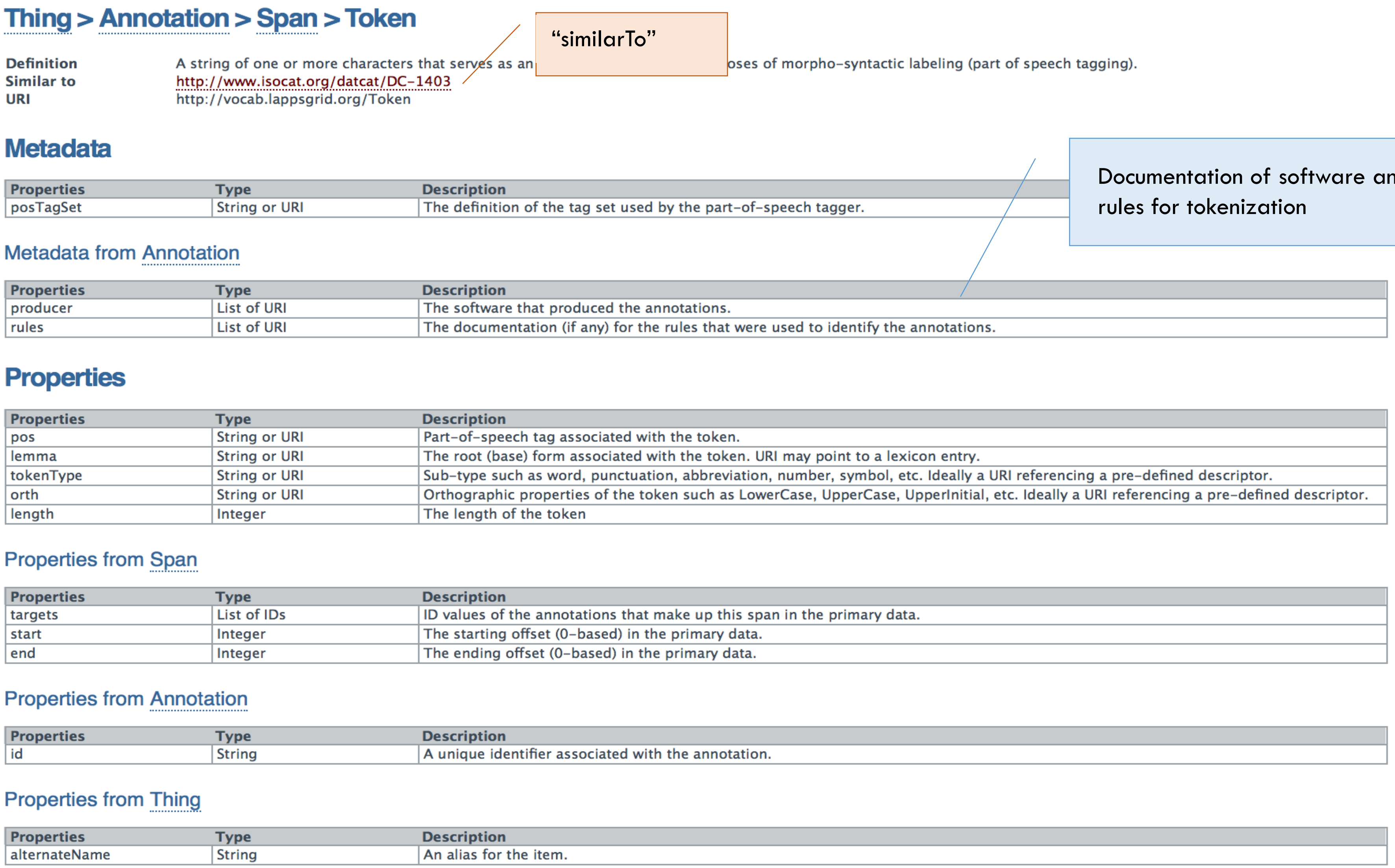
- http://vocab.lappsgrid.org
- Shallow hierarchy of elements
    - Inheritance



LAPPS Exchange Vocabulary Type Hierarchy

- Thing: alternateName
    - Annotation: id
        - Region: targets, start, end
            - Paragraph
            - Sentence: sentenceType
            - NounChunk
            - VerbChunk: vcType, tense, voice, neg
            - NamedEntity: category, type, gender
                - ~~Date: dateType~~
                - ~~Location: locType~~
                - ~~Organization: orgType~~
                - ~~Person: gender~~
            - Token: pos, lemma, tokenType, orth, length, word
            - Markable
        - Relation: label
            - GenericRelation: relation, arguments
            - SemanticRole: head, argument
            - Constituent: parent, children
            - Dependency: governor, dependent
        - Coreference: mentions, representative
        - PhraseStructure: constituents, root
        - DependencyStructure: dependencyType, dependencies
    - Document: id, source, sourceType, encoding, language
        - TextDocument
        - AudioDocument

Spec for Token



Thing > Annotation > Span > Token

Definition	A string of one or more characters that serves as an oses of morpho-syntactic labeling (part of speech tagging).
Similar to	http://www.isocat.org/datcat/DC-1403
URI	http://vocab.lappsgrid.org/Token

“similarTo”

Metadata

Properties	Type	Description
posTagSet	String or URI	The definition of the tag set used by the part-of-speech tagger.

Metadata from Annotation

Properties	Type	Description
producer	List of URI	The software that produced the annotations.
rules	List of URI	The documentation (if any) for the rules that were used to identify the annotations.

Documentation of software and rules for tokenization

Properties

Properties	Type	Description
pos	String or URI	Part-of-speech tag associated with the token.
lemma	String or URI	The root (base) form associated with the token. URI may point to a lexicon entry.
tokenType	String or URI	Sub-type such as word, punctuation, abbreviation, number, symbol, etc. Ideally a URI referencing a pre-defined descriptor.
orth	String or URI	Orthographic properties of the token such as LowerCase, UpperCase, UpperInitial, etc. Ideally a URI referencing a pre-defined descriptor.
length	Integer	The length of the token

Properties from Span

Properties	Type	Description
targets	List of IDs	ID values of the annotations that make up this span in the primary data.
start	Integer	The starting offset (0-based) in the primary data.
end	Integer	The ending offset (0-based) in the primary data.

Properties from Annotation

Properties	Type	Description
id	String	A unique identifier associated with the annotation.

Properties from Thing

Properties	Type	Description
alternateName	String	An alias for the item.

JSON-LD and the LAPPS Exchange Vocabulary

```
"@context" : "http://vocab.lappsgrid.org/",
  "metadata" : { },
  "text" : {
    "@value" : “Some of the strongest critics of our welfare system …
  ”views" : [ {
    "metadata" : {
      "contains" : {
        "Token" : {
          "producer" : "org.anc.lapps.stanford.SATokenizer:1.4.0",
          "type" : “tokenization:stanford"
        }
      }
    },
    "annotations" : [ {
      "@type" : "Token",
      "id" : "tok0",
      "start" : 18,
      "end" : 22,
      "features" : {
        ”string" : "Some"
      }
    },
...
```

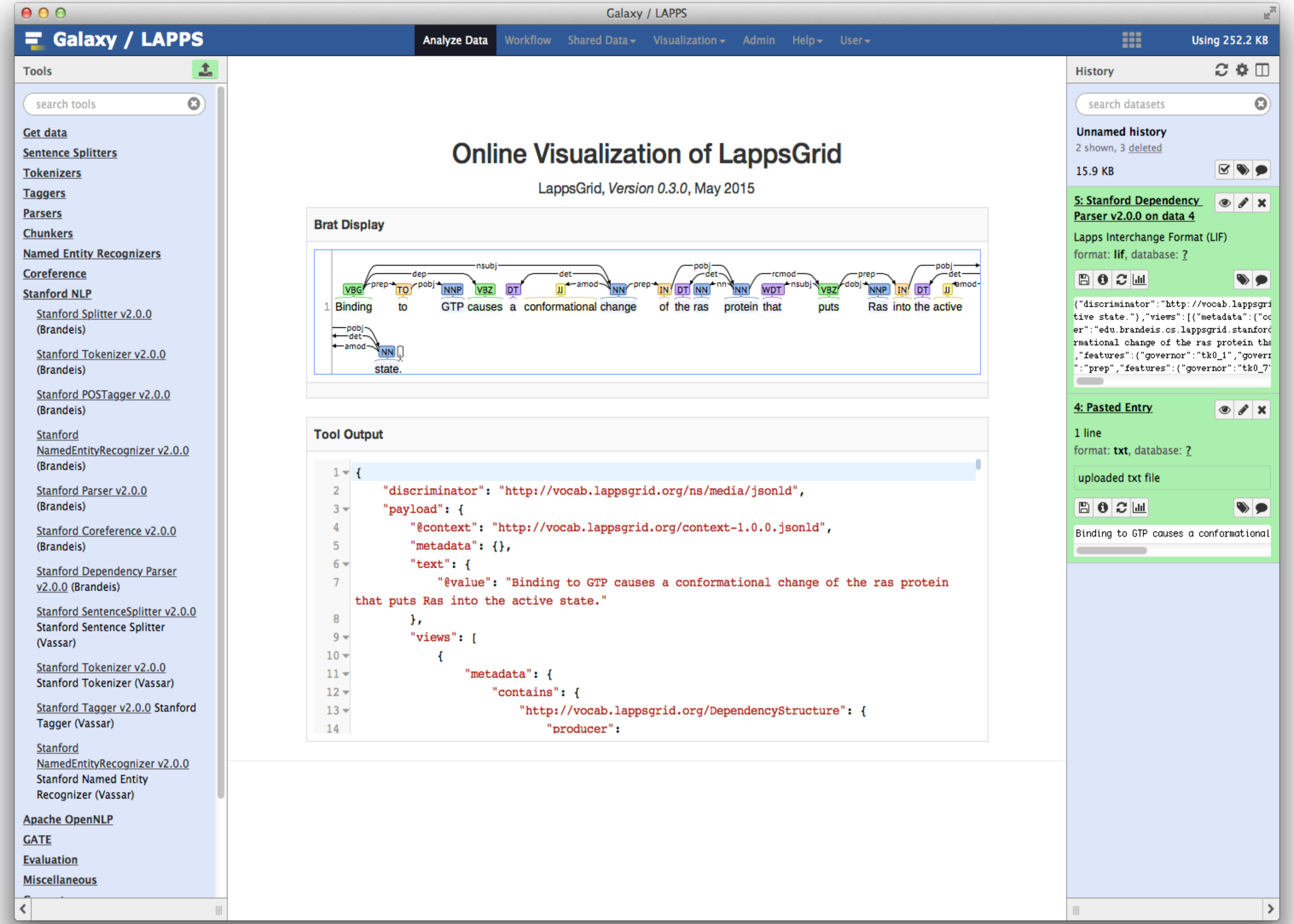
Base URI for the LAPPS Exchange Vocabulary

Metadata for the annotations

Internal LAPPS type defined at
http://vocab.lappsgrid.org.LIF

defined at http://vocab.lappsgrid.org/Token

Features defined at http://vocab.lappsgrid.org/Token#[feature-name]


Galaxy / LAPPS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 252.2 KB
Tools
search tools
Get data
Sentence Splitters
Tokenizers
Taggers
Parsers
Chunkers
Named Entity Recognizers
Coreference
Stanford NLP
Stanford Splitter v2.0.0 (Brandeis)
Stanford Tokenizer v2.0.0 (Brandeis)
Stanford POSTagger v2.0.0 (Brandeis)
Stanford NamedEntityRecognizer v2.0.0 (Brandeis)
Stanford Parser v2.0.0 (Brandeis)
Stanford Coreference v2.0.0 (Brandeis)
Stanford Dependency Parser v2.0.0 (Brandeis)
Stanford SentenceSplitter v2.0.0 Stanford Sentence Splitter (Vassar)
Stanford Tokenizer v2.0.0 Stanford Tokenizer (Vassar)
Stanford Tagger v2.0.0 Stanford Tagger (Vassar)
Stanford NamedEntityRecognizer v2.0.0 Stanford Named Entity Recognizer (Vassar)
Apache OpenNLP
GATE
Evaluation
Miscellaneous
History
search datasets
Unnamed history
2 shown, 3 deleted
15.9 KB
5: Stanford Dependency Parser v2.0.0 on data 4
Lapps Interchange Format (LIF)
format: lif, database: ?
4: Pasted Entry
1 line
format: txt, database: ?
uploaded txt file

Online Visualization of LappsGrid

LappsGrid, *Version 0.3.0*, May 2015

Brat Display

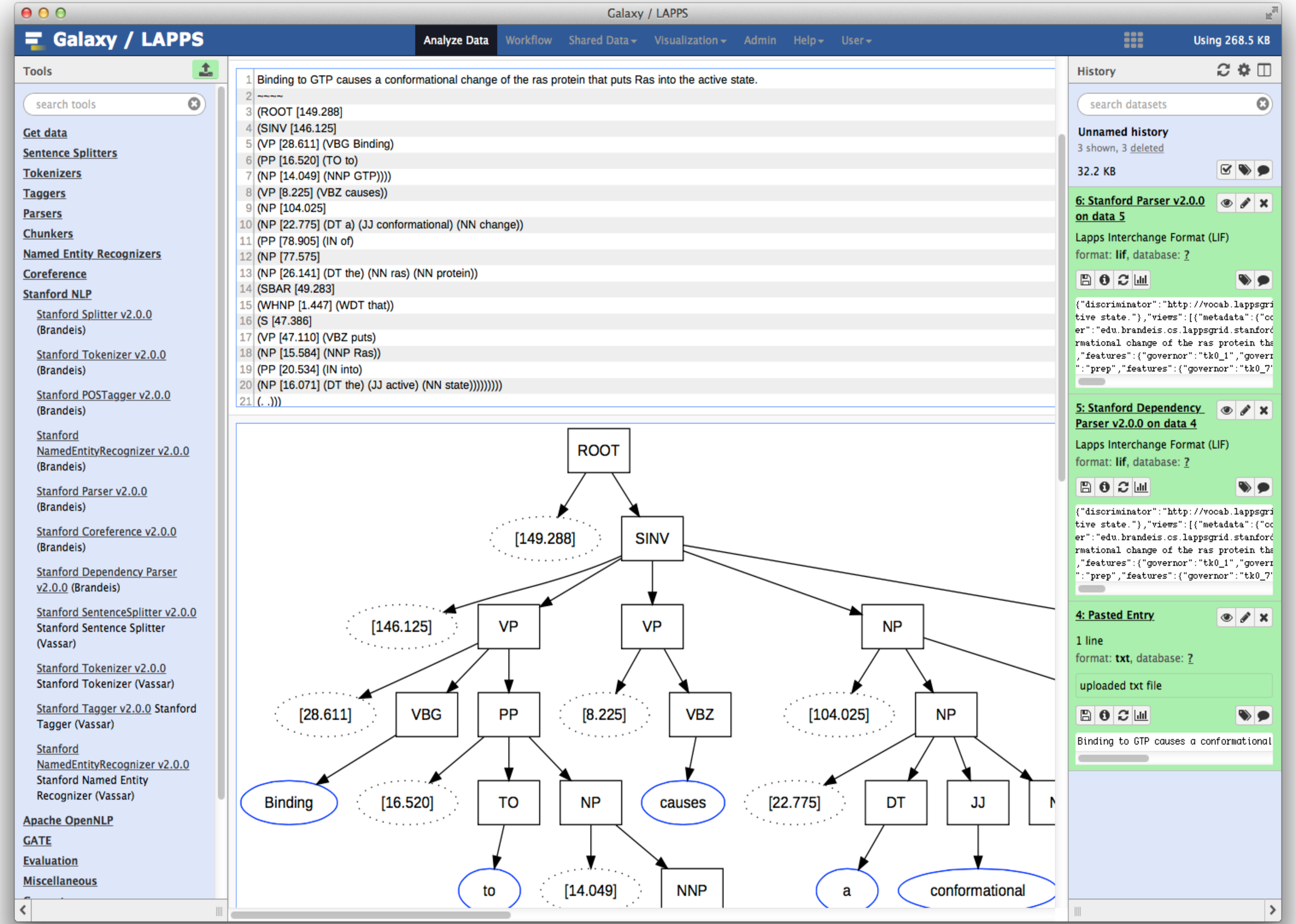
nsubj
dep
det
pobj
rcmod
prep
pobj
det
prep
pobj
amod
prep
det
nn
nsubj
dobj
det
amod
VBG TO NNP VBZ DT JJ NN IN DT NN NN WDT VBZ NNP IN DT JJ
1 Binding to GTP causes a conformational change of the ras protein that puts Ras into the active
pobj
det
amod
NN .
state.

Tool Output

```
{
    "discriminator": "http://vocab.lappsgrid.org/ns/media/jsonld",
    "payload": {
        "@context": "http://vocab.lappsgrid.org/context-1.0.0.jsonld",
        "metadata": {},
        "text": {
            "@value": "Binding to GTP causes a conformational change of the ras protein
that puts Ras into the active state."
        },
        "views": [
            {
                "metadata": {
                    "contains": {
                        "http://vocab.lappsgrid.org/DependencyStructure": {
                            "producer":
```

```
{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","govern
":"prep","features":{"governor":"tk0_7"
```

```
Binding to GTP causes a conformational
```


Galaxy / LAPPS
Galaxy / LAPPS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 268.5 KB
Tools
search tools
Get data
Sentence Splitters
Tokenizers
Taggers
Parsers
Chunkers
Named Entity Recognizers
Coreference
Stanford NLP
Stanford Splitter v2.0.0 (Brandeis)
Stanford Tokenizer v2.0.0 (Brandeis)
Stanford POSTagger v2.0.0 (Brandeis)
Stanford NamedEntityRecognizer v2.0.0 (Brandeis)
Stanford Parser v2.0.0 (Brandeis)
Stanford Coreference v2.0.0 (Brandeis)
Stanford Dependency Parser v2.0.0 (Brandeis)
Stanford SentenceSplitter v2.0.0 Stanford Sentence Splitter (Vassar)
Stanford Tokenizer v2.0.0 Stanford Tokenizer (Vassar)
Stanford Tagger v2.0.0 Stanford Tagger (Vassar)
Stanford NamedEntityRecognizer v2.0.0 Stanford Named Entity Recognizer (Vassar)
Apache OpenNLP
GATE
Evaluation
Miscellaneous

```
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
~~~~
(ROOT [149.288]
(SINV [146.125]
(VP [28.611] (VBG Binding)
(PP [16.520] (TO to)
(NP [14.049] (NNP GTP))))
(VP [8.225] (VBZ causes))
(NP [104.025]
(NP [22.775] (DT a) (JJ conformational) (NN change))
(PP [78.905] (IN of)
(NP [77.575]
(NP [26.141] (DT the) (NN ras) (NN protein))
(SBAR [49.283]
(WHNP [1.447] (WDT that))
(S [47.386]
(VP [47.110] (VBZ puts)
(NP [15.584] (NNP Ras))
(PP [20.534] (IN into)
(NP [16.071] (DT the) (JJ active) (NN state))))))))
(. .)))
```

6: Stanford Parser v2.0.0 on data 5

Lapps Interchange Format (LIF)
format: lif, database: ?

{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","govern
":"prep","features":{"governor":"tk0_7"

5: Stanford Dependency Parser v2.0.0 on data 4

Lapps Interchange Format (LIF)
format: lif, database: ?

{"discriminator":"http://vocab.lappsgri
tive state."},"views":[{"metadata":{"co
er":"edu.brandeis.cs.lappsgrid.stanford
rmational change of the ras protein tha
,"features":{"governor":"tk0_1","govern
":"prep","features":{"governor":"tk0_7"

4: Pasted Entry

1 line
format: txt, database: ?

uploaded txt file

Binding to GTP causes a conformational

Evaluation in the LAPPS Grid

- CMU has implemented services for state-of-the-art Open Advancement techniques
  - Used in the development of IBM's Watson to achieve steady performance gains over the four years of its development
- Provides an unprecedented tool for NLP development
  - Could take the field to a new level of productivity
- Enables rapid identification of
  - frequent error categories within modules and documents
  - which module(s) and error type(s) have the greatest impact on overall performance

Open Advancement in a Nutshell

01

Evaluate multiple possible solutions (tool configurations) for a given problem

- Determine the optimal solution available using given components, resources, and evaluation data

02

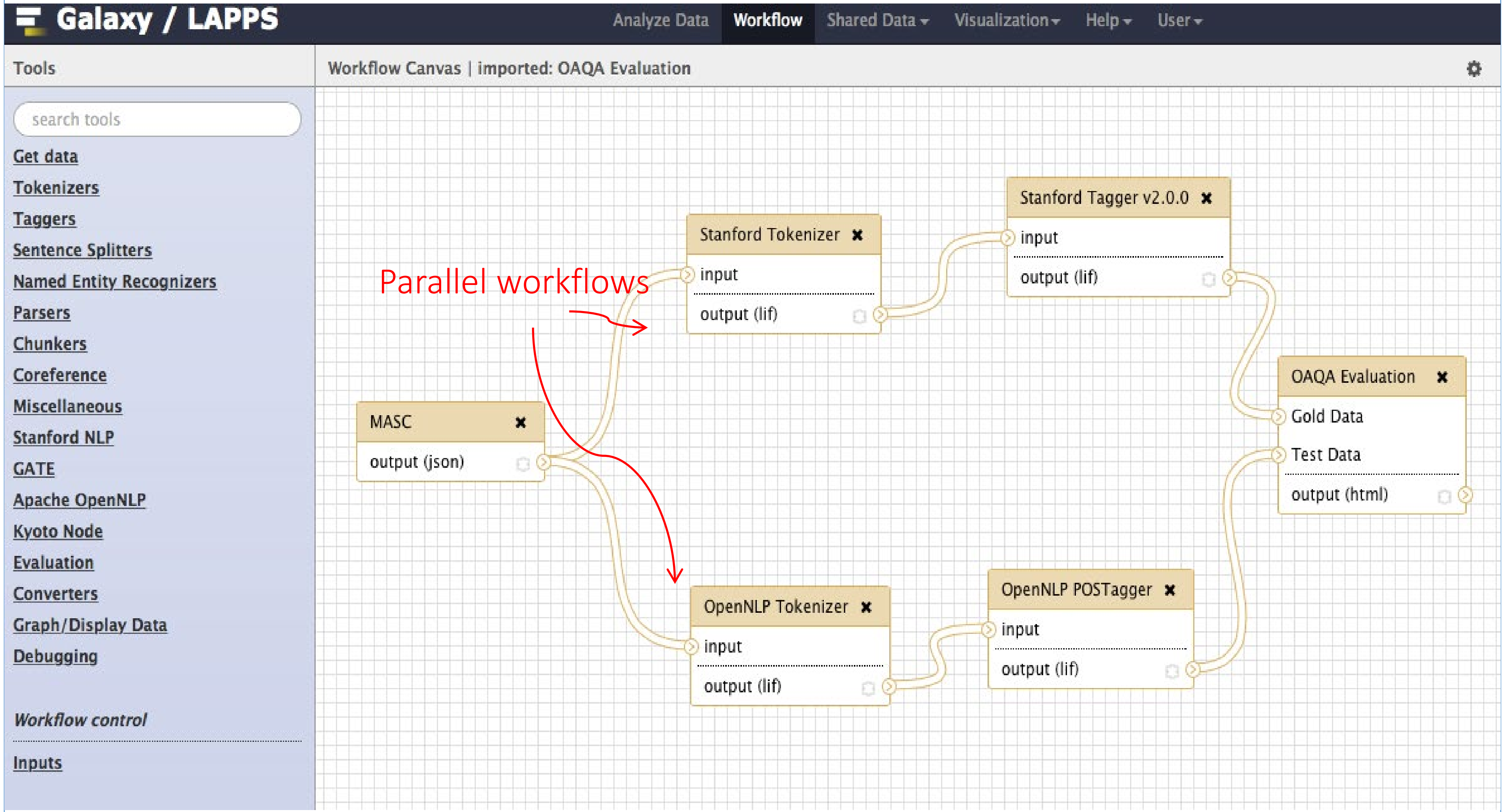
Output of the optimal solution subjected to error analysis

- Identify the most frequent errors with the highest impact
- Consider possible enhancements
  - Aim to achieve the largest possible reduction in error rate by addressing the most frequent error types

03

Evaluate performance of new configurations

- Determine if a significant improvement has been achieved in comparison with prior baselines



Parallel workflows

BioNLP-oriented Tools in the LAPPS Grid

Penn BioTokenizer

Biomedical NER

- Annotates proteins, DNA, RNA, cellLines, cellTYpes

Gene annotator

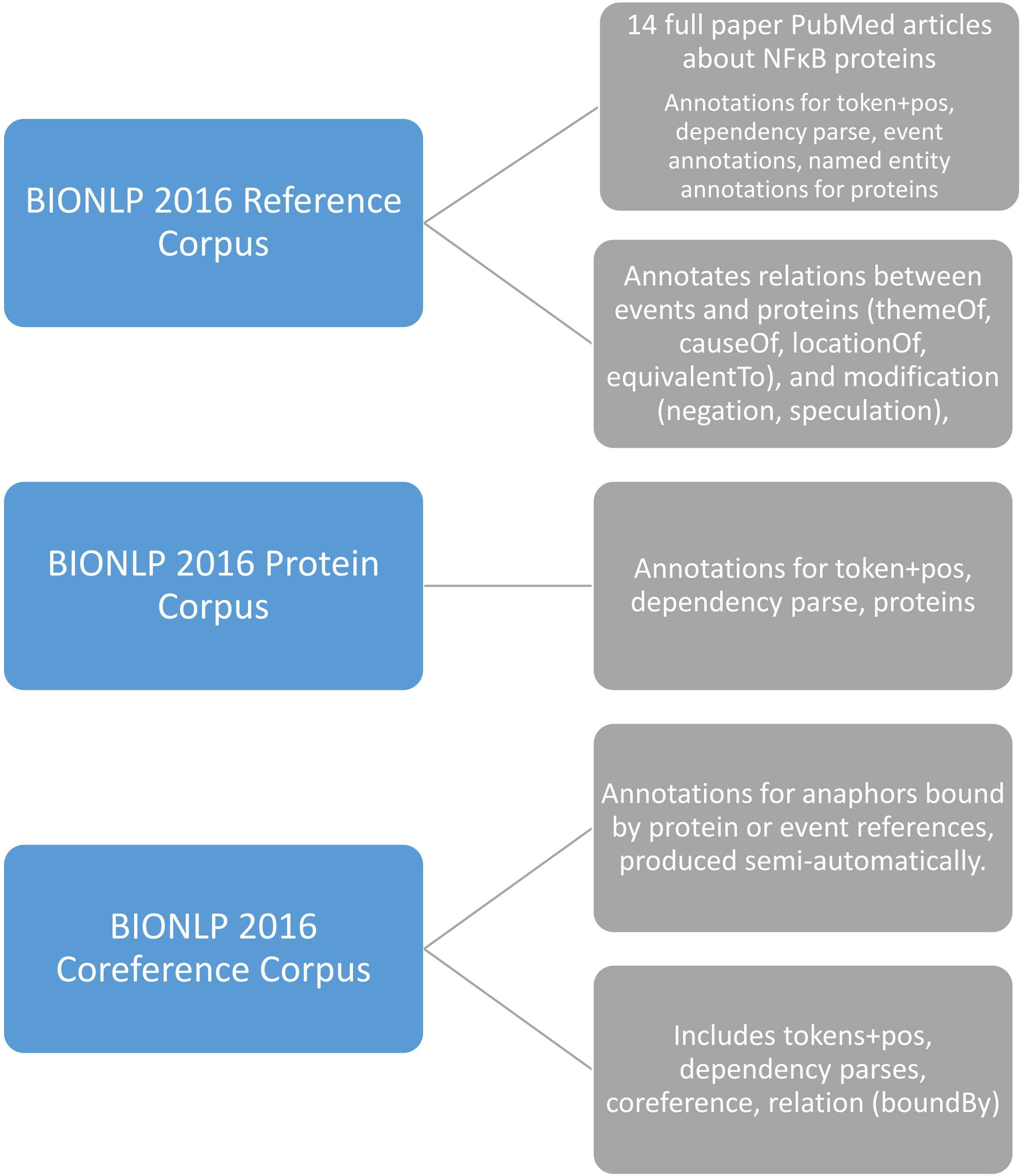
CDC/FDC CTakes

UCREL Semantic Tagger

Other LAPPS Grid Tools Useful for BioNLP

- TimeML Events
- LingPipe Dictionary-based NER
- Several different NER modules, tokenizers, parsers, chunkers, etc.
- HeidelTime
- Evaluation tools (Open Advancement)

Gold Standard Biomedical Data in the LAPPS Grid



BIONLP 2016 Reference Corpus

14 full paper PubMed articles about NFκB proteins

Annotations for token+pos, dependency parse, event annotations, named entity annotations for proteins

Annotates relations between events and proteins (themeOf, causeOf, locationOf, equivalentTo), and modification (negation, speculation),

BIONLP 2016 Protein Corpus

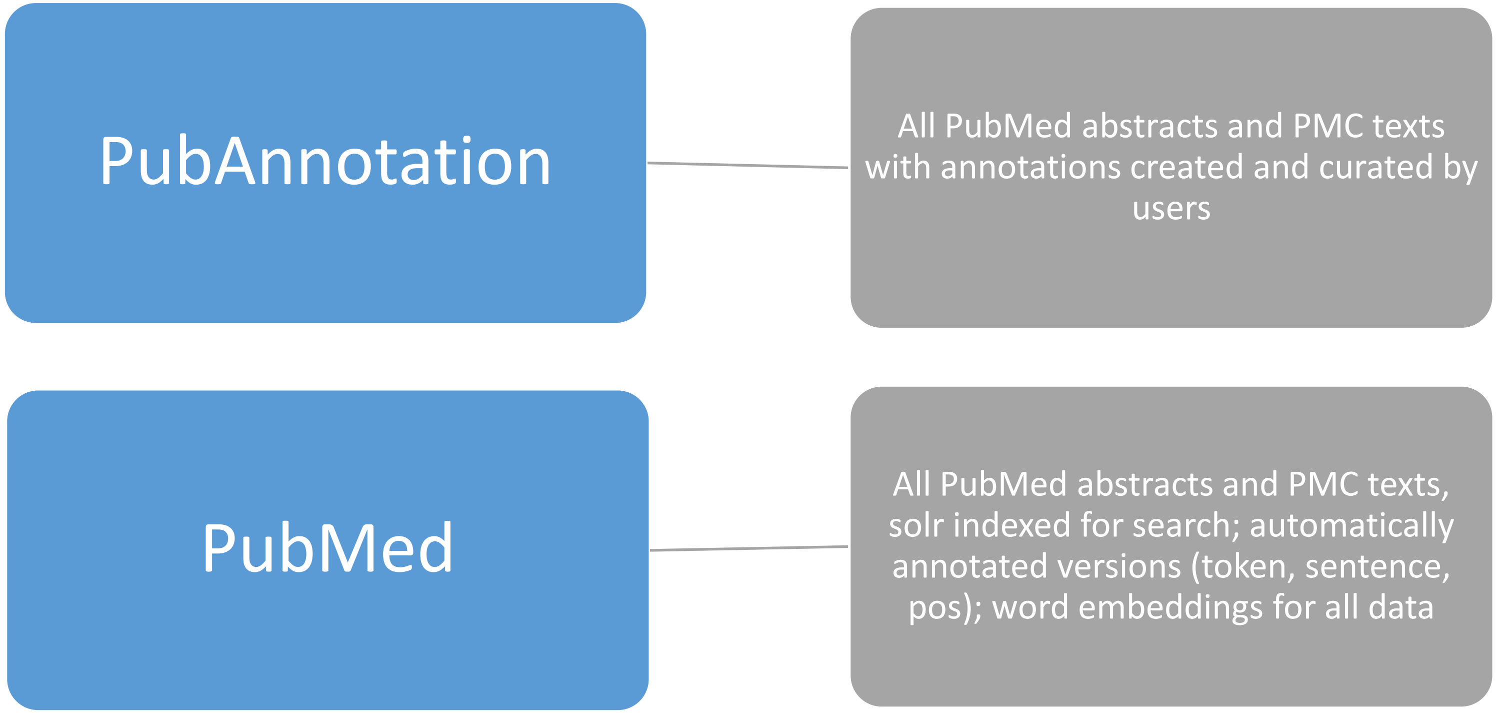
Annotations for token+pos, dependency parse, proteins

BIONLP 2016 Coreference Corpus

Annotations for anaphors bound by protein or event references, produced semi-automatically.

Includes tokens+pos, dependency parses, coreference, relation (boundBy)

Access to Biomedical data from the LAPPS Grid



PubAnnotation

All PubMed abstracts and PMC texts with annotations created and curated by users

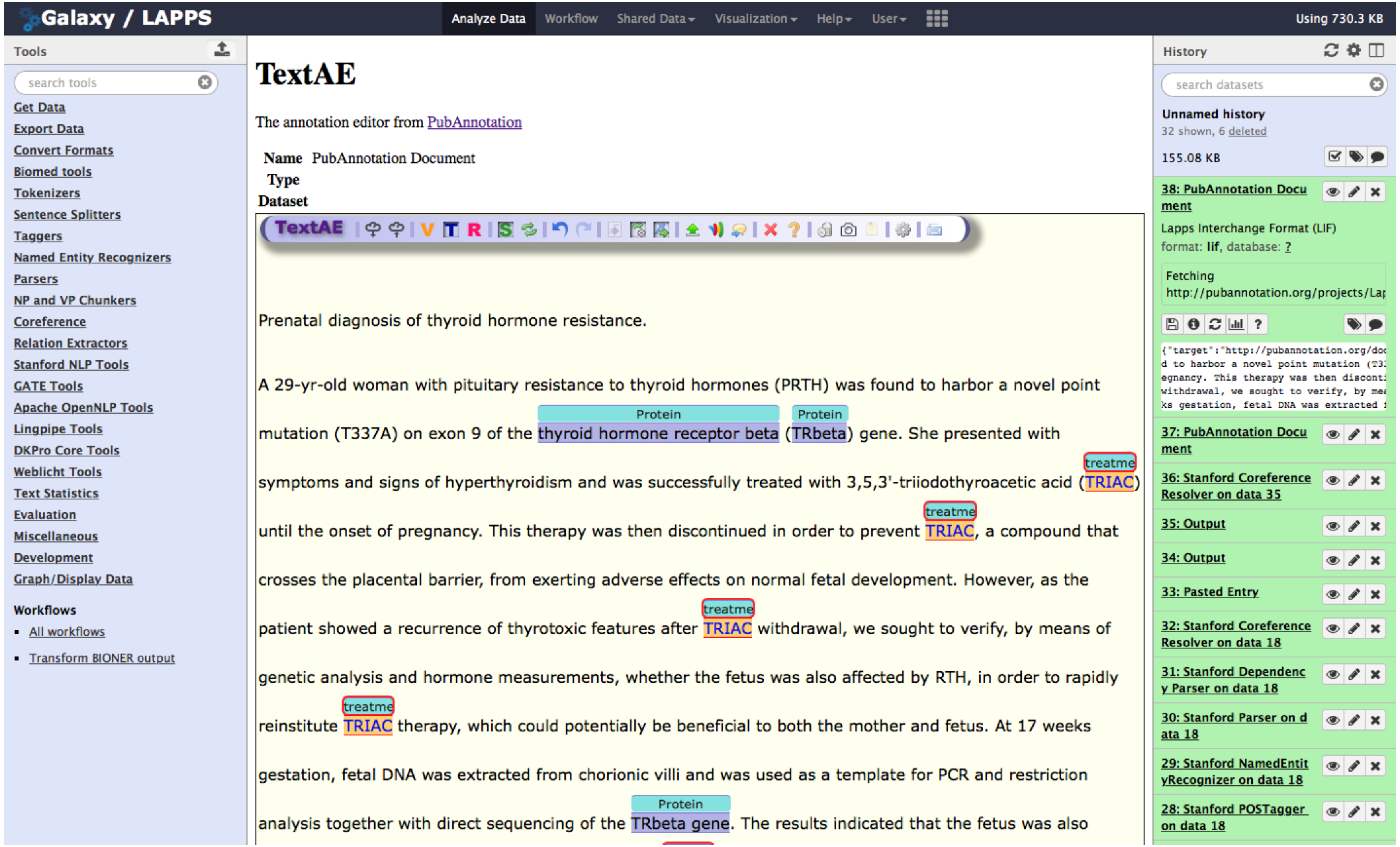
PubMed

All PubMed abstracts and PMC texts, solr indexed for search; automatically annotated versions (token, sentence, pos); word embeddings for all data

PubAnnotation

- A **repository of annotations applied to biomedical publications**, all of which are aligned to the canonical text in either PubMed or PubMed Central
  - All PubAnnotation **annotations are thus linked to each other through the canonical texts**
- PubAnnotation includes **TextAE**, a powerful and easy-to-use Javascript **app for text annotation and visualization**

TextAE Visualization and Editing in the LAPPS Grid



TextAE

The annotation editor from PubAnnotation

Name PubAnnotation Document
Type
Dataset

Prenatal diagnosis of thyroid hormone resistance.

A 29-yr-old woman with pituitary resistance to thyroid hormones (PRTH) was found to harbor a novel point mutation (T337A) on exon 9 of the thyroid hormone receptor beta (TRbeta) gene. She presented with symptoms and signs of hyperthyroidism and was successfully treated with 3,5,3'-triiodothyroacetic acid (TRIAC) until the onset of pregnancy. This therapy was then discontinued in order to prevent TRIAC, a compound that crosses the placental barrier, from exerting adverse effects on normal fetal development. However, as the patient showed a recurrence of thyrotoxic features after TRIAC withdrawal, we sought to verify, by means of genetic analysis and hormone measurements, whether the fetus was also affected by RTH, in order to rapidly reinstitute TRIAC therapy, which could potentially be beneficial to both the mother and fetus. At 17 weeks gestation, fetal DNA was extracted from chorionic villi and was used as a template for PCR and restriction analysis together with direct sequencing of the TRbeta gene. The results indicated that the fetus was also

Unnamed history
32 shown, 6 deleted
155.08 KB

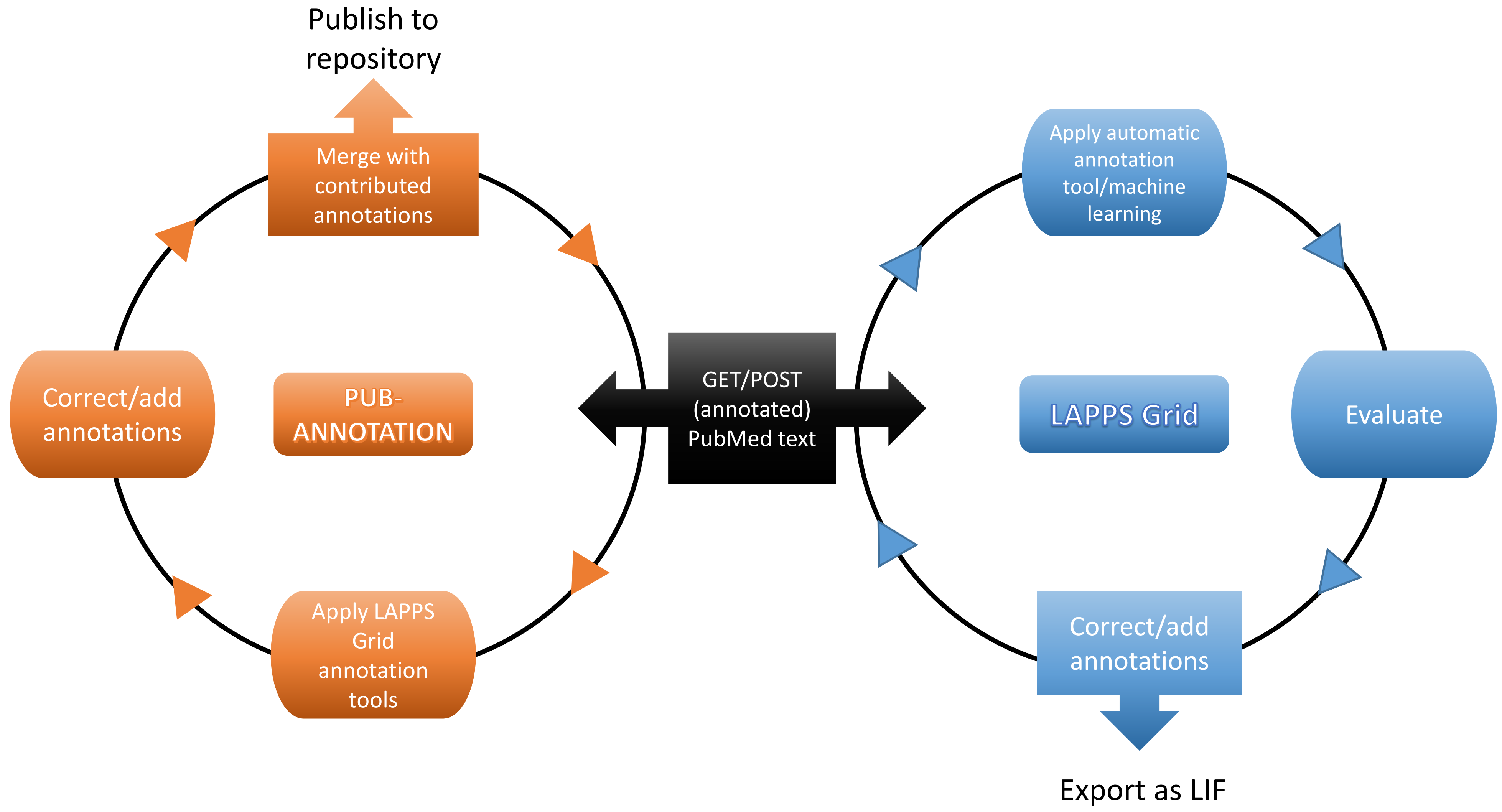
38: PubAnnotation Document
Lapps Interchange Format (LIF)
format: **lif**, database: ?

Fetching http://pubannotation.org/projects/La

```
{"target":"http://pubannotation.org/do
d to harbor a novel point mutation (T3
egnancy. This therapy was then disconti
withdrawal, we sought to verify, by mea
ks gestation, fetal DNA was extracted 
```

37: PubAnnotation Document
36: Stanford Coreference Resolver on data 35
35: Output
34: Output
33: Pasted Entry
32: Stanford Coreference Resolver on data 18
31: Stanford Dependency Parser on data 18
30: Stanford Parser on data 18
29: Stanford NamedEntityRecognizer on data 18
28: Stanford POSTagger on data 18

Interaction between PubAnnotation and LAPPS Grid



Publish to repository
Merge with contributed annotations
Correct/add annotations
PUB-ANNOTATION
Apply LAPPS Grid annotation tools
GET/POST (annotated) PubMed text
Apply automatic annotation tool/machine learning
LAPPS Grid
Evaluate
Correct/add annotations
Export as LIF

Current Activities

- **NSF ABI grant**
  - Collaboration between Vassar College and **Galaxy Principal Investigators** to
    - Develop tools, ready-made workflows, etc. for mining biomedical publications
    - Provide seamless integration of text mining capabilities and the vast array of tools provided in Galaxy
- Collaboration with the US government **Centers for Disease Control** and **Food and Drug Administration** to adapt the LAPPS Grid for summarization and mining of clinical reports

Current Activities

- **NSF EAGER grant** (Vassar, Brandeis, Tufts, Penn State) to develop and implement **methods for domain adaptation** to accommodate specific areas of scientific text mining research
- Collaboration with **PubAnnotation** to fully integrate the two platforms to enable **iterative development of language models** via machine learning
- Nascent collaboration with **University of Wisconsin's "Geo Deep-Dive" project**, access to millions of scientific publications (many copyrighted) using their extensive HPC facilities

LAPPS Grid is a Work in Progress

- Recent shift to scientific text mining
- Establishing an increasing number of fruitful collaborations
- Seeking contributions of software, data, resources, ideas

Thank you