NCATS Translator prototype TReK development

Vincent Emonet
Michel Dumontier

Maastricht University
Institute of Data Science

March 2020

Problems

Various tools and interfaces are required to **realize the vision of a Translator ecosystem**

Many Semantic Web technologies exist, but they remain **hard to find and deploy**

Data providers could benefit from additional guidance to **expose their structured data** with **Translator-compliant interfaces**

Transformation workflows are usually implemented per case and can be **hard to reconfigure**

Data2Services

A Command Line Interface for building and deploying standards-compliant (RDF) Knowledge Graphs with a bundle of programmatic and user interfaces.

Data2Services

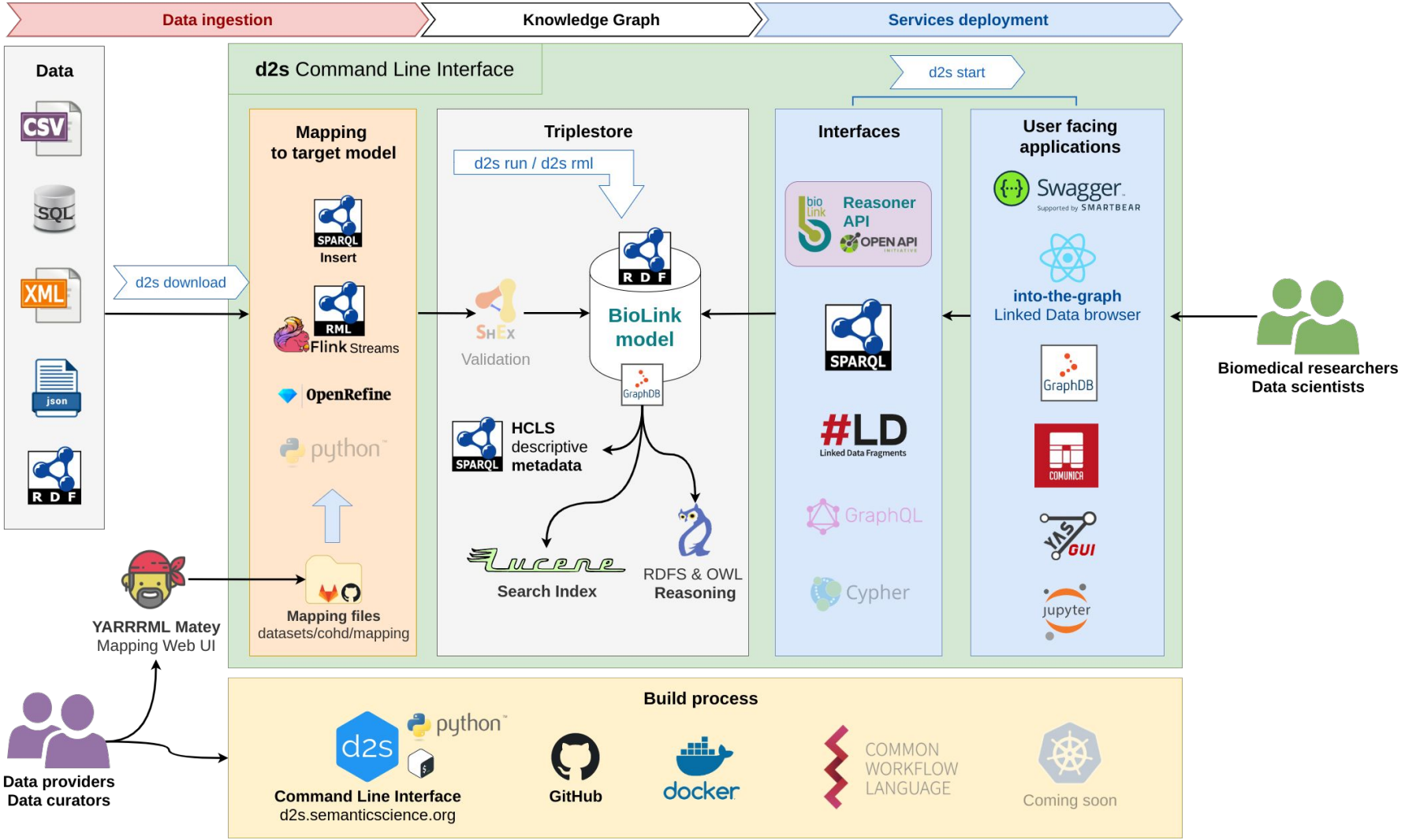
Test on local machine	Deploy on single server	Scale in cluster
Container-based deployment of **services and workflows** on a Linux or MacOS laptop 💻	**Container-based** deployment of **services and workflows** on a single Linux server	*In development*: deploy on multiple nodes in a cluster with Kubernetes or OpenShift
docker	docker	OPENSHIFT

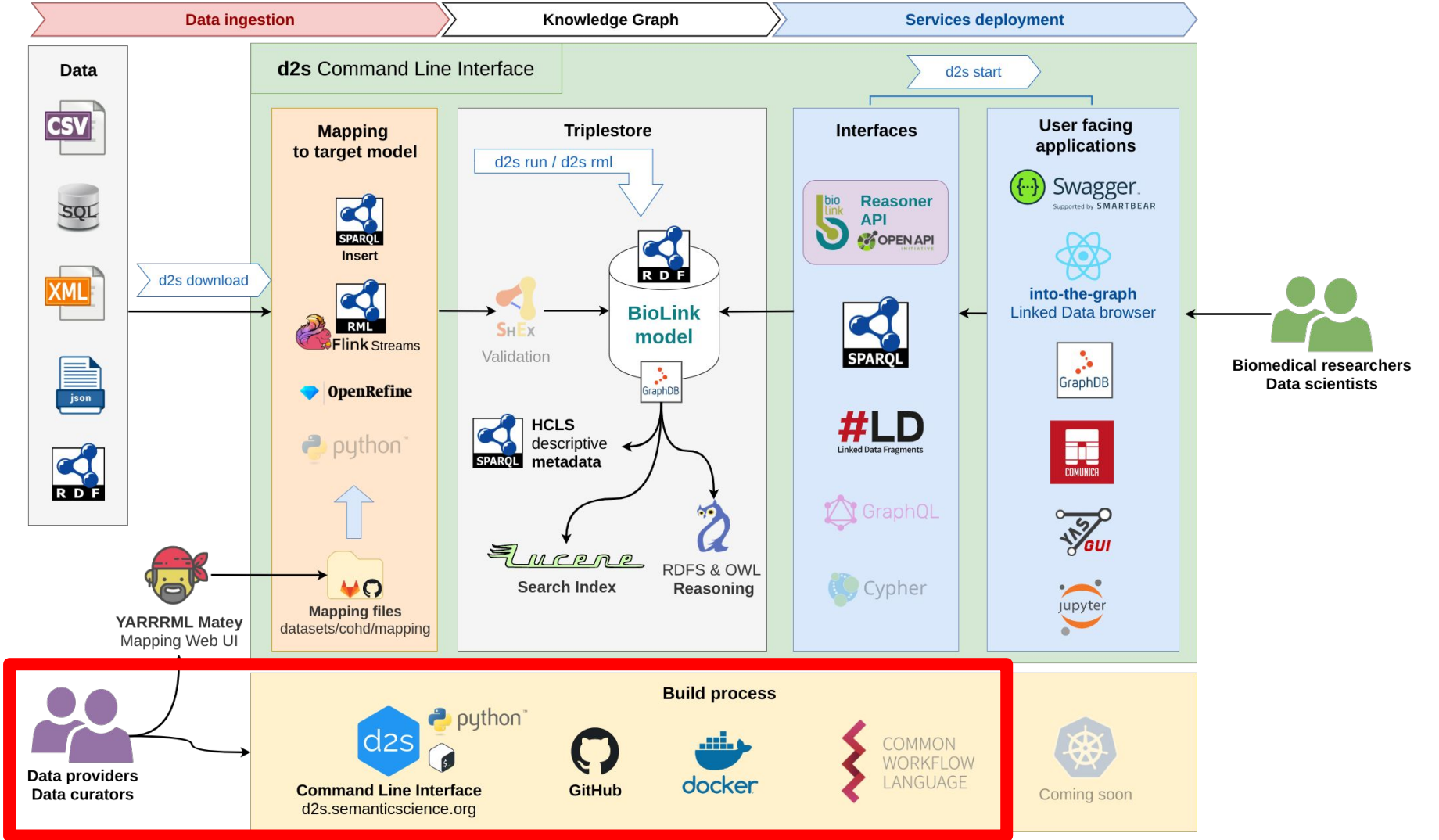
Tested on

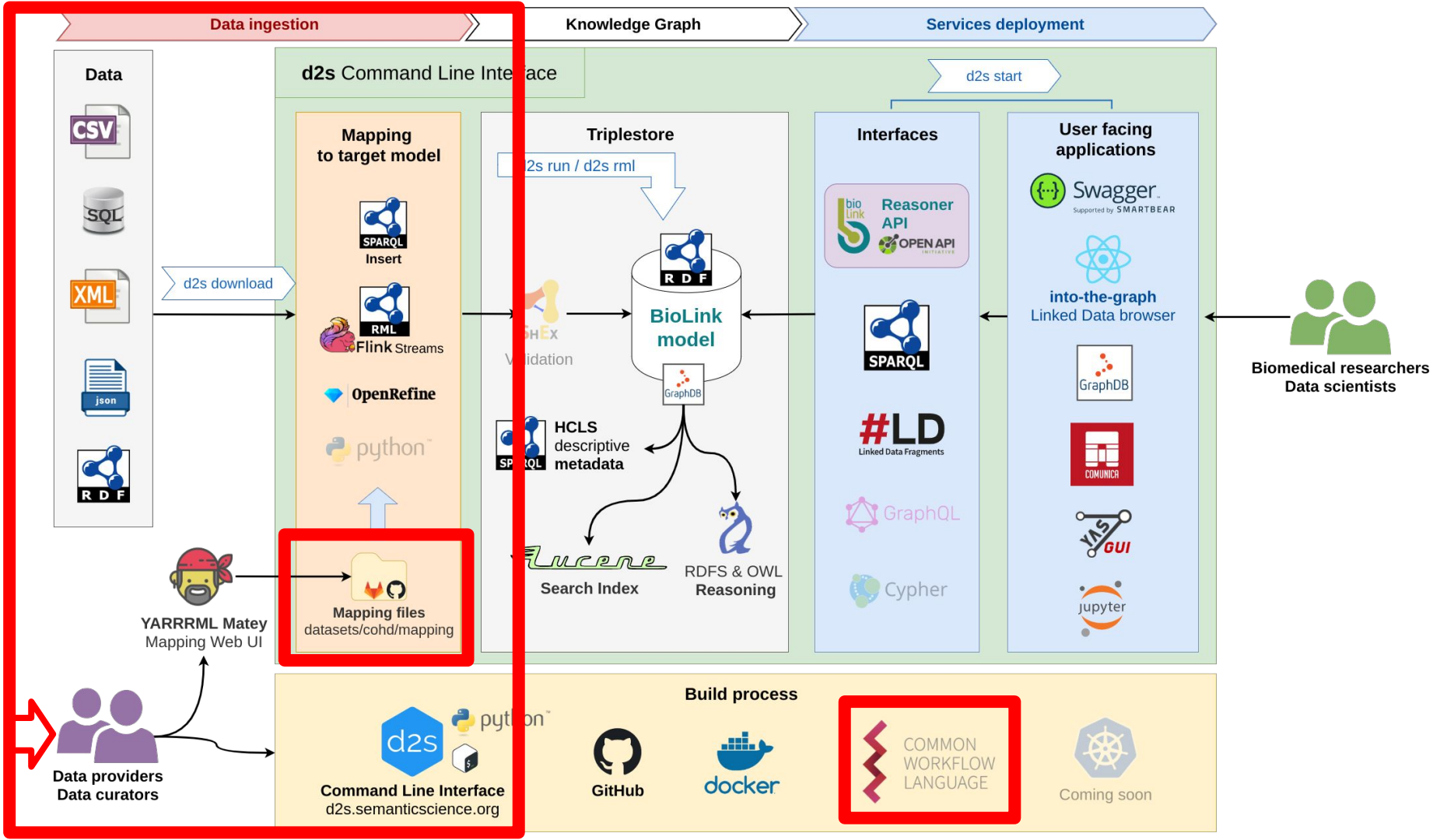
ubuntu

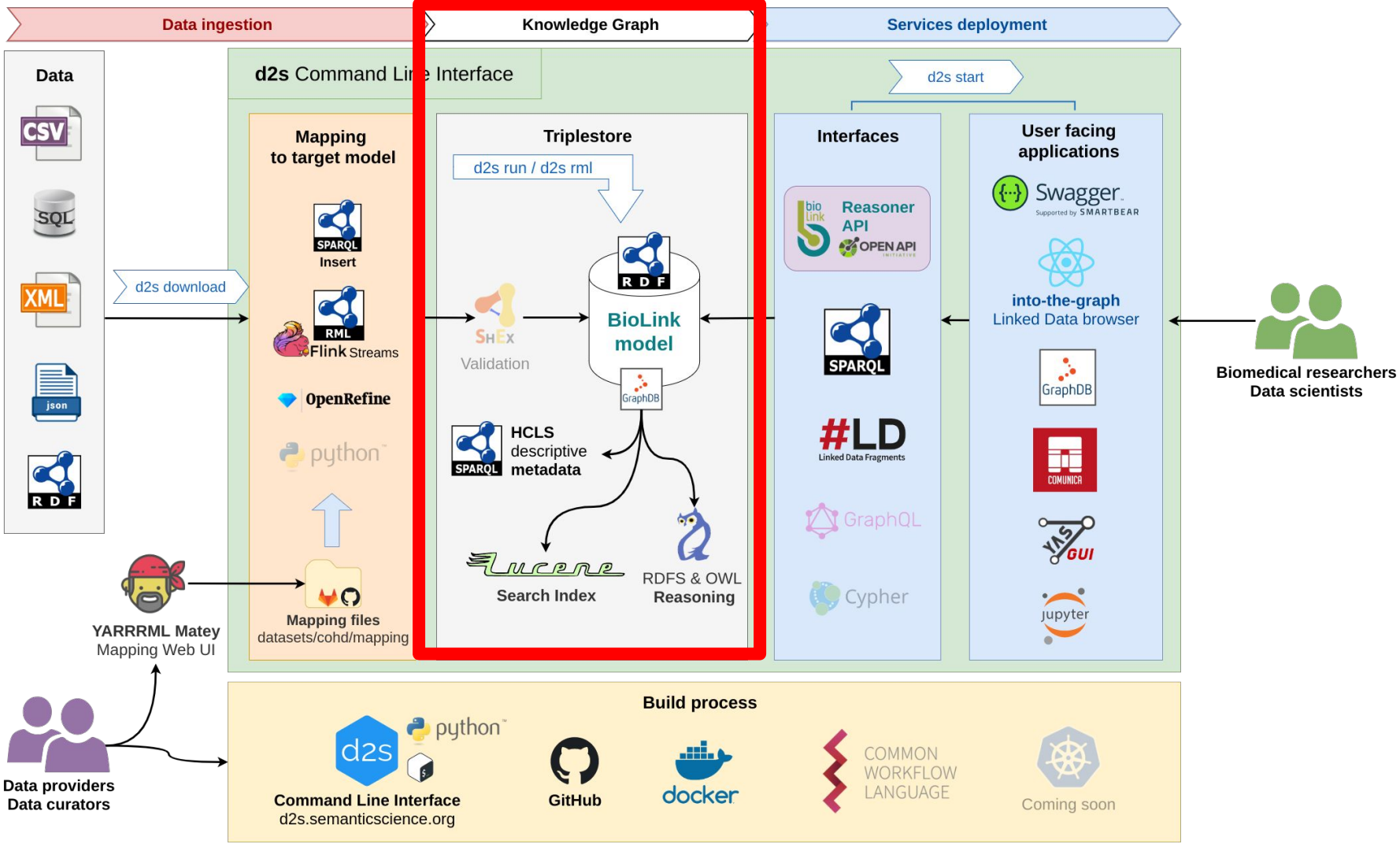
mac OS

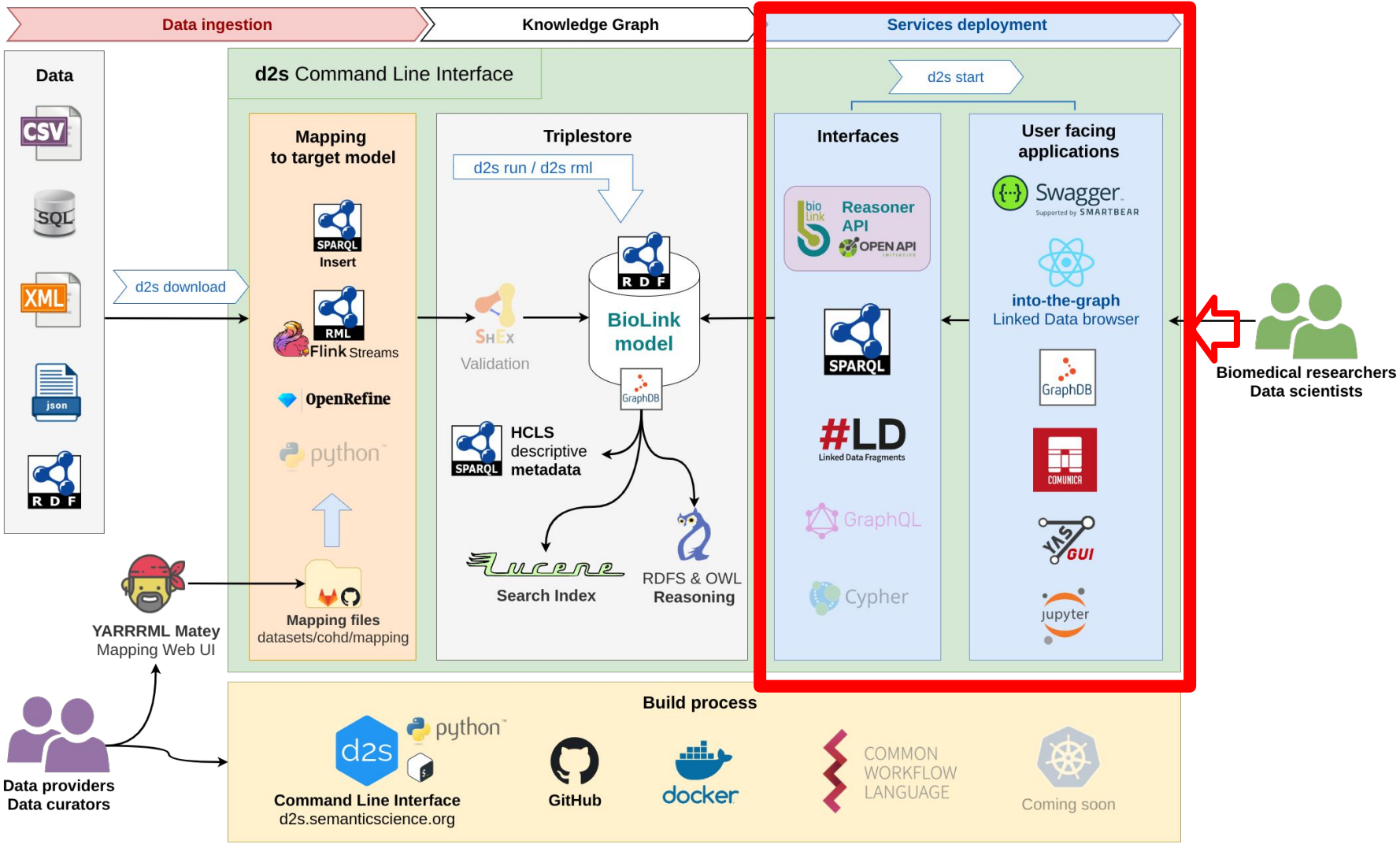
CentOS


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s download
d2s Command Line Interface
d2s start
Mapping to target model
SPARQL
Insert
RML
Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Search Index
RDFS & OWL Reasoning
Interfaces
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s start
d2s download
Mapping to target model
SPARQL
Insert
RML
Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Search Index
RDFS & OWL Reasoning
Interfaces
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s download
Mapping to target model
SPARQL
Insert
RML
Flink Streams
OpenRefine
python
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Lucene
Search Index
RDFS & OWL Reasoning
d2s start
Interfaces
bio Link
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
YARRRML Matey
Mapping Web UI
Mapping files
datasets/cohd/mapping
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s download
Mapping to target model
SPARQL Insert
RML
Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Search Index
RDFS & OWL Reasoning
d2s start
Interfaces
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YASGUI
jupyter
Biomedical researchers
Data scientists
Data providers
Data curators
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon


Data ingestion
Knowledge Graph
Services deployment
Data
CSV
SQL
XML
json
RDF
d2s Command Line Interface
d2s download
Mapping to target model
SPARQL
Insert
RML
Flink Streams
OpenRefine
python
Mapping files
datasets/cohd/mapping
YARRRML Matey
Mapping Web UI
Triplestore
d2s run / d2s rml
SHEX
Validation
RDF
BioLink model
GraphDB
HCLS descriptive metadata
Search Index
RDFS & OWL Reasoning
d2s start
Interfaces
Reasoner API
OPEN API
SPARQL
#LD
Linked Data Fragments
GraphQL
Cypher
User facing applications
Swagger
Supported by SMARTBEAR
into-the-graph
Linked Data browser
GraphDB
COMUNICA
YAS GUI
jupyter
Biomedical researchers
Data scientists
Build process
d2s
python
Command Line Interface
d2s.semanticscience.org
GitHub
docker
COMMON WORKFLOW LANGUAGE
Coming soon
Data providers
Data curators

Mapping COHD to BioLink RDF

Mappings defined in the project folder in datasets/cohd/mapping

RML mappings
for associations

Better for scalability

```
<MapAssociation> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "/mnt/workspace/input/cohd/paired_concept_counts_associations.csv";
    rml:referenceFormulation ql:CSV ];
  rr:subjectMap [
    rr:template "https://w3id.org/d2s/cohd/association/{dataset_id}_{concept_id_1}_{concept_id_2}";
    rr:class bl:Association ];
  rr:predicateObjectMap [
    rr:predicate bl:subject;
    rr:objectMap [ rr:template "http://api.ohdsi.org/WebAPI/vocabulary/concept/{concept_id_1}" ]];
```

SPARQL Insert mappings
for concepts

Map generic RDF generated from input data structure

Better to join files/tables

```
INSERT {
  GRAPH <https://w3id.org/biolink/graph/cohd> {
   ?Concept_api_uri a bl:Drug ;
      bl:id ?Concept_id ;
      bl:name ?Concept_name . }
} WHERE {
  SERVICE <http://temporary-triplestore>  {
      ?row a <https://w3id.org/d2s/concepts.tsv> ;
        d2s:Concept_id ?Concept_id ;
        d2s:Concept_name ?Concept_name .
      BIND ( iri(concat("http://api.ohdsi.org/WebAPI/vocabulary/concept/",
                        ?Concept_id)) AS ?Concept_api_uri )
```

Accessible through a Reasoner API

Built using Spring Boot framework

Servers

http://api.trek.semanticscience.org - Generated server url

Reasoner API

Query BioLink-compliant datasets using the Reasoner API

POST `/reasoner/v1/query` Execute a Reasoner API query on the BioLink-compliant triplestore.

Query the BioLink-compliant knowledge graph using the Reasoner API query specifications.

Use this example query for COHD:

```
{
  "max_results": 50,
  "message": {
    "query_graph": {
      "nodes": [
        { "id": "n00", "type": "Procedure" },
        { "id": "n01", "type": "Drug" }
      ],
      "edges": [
        { "id": "e00", "type": "Association",
          "source_id": "n00", "target_id": "n01" }
      ]
    },
    "query_options": {
       "https://w3id.org/trek/cohd/attribute/ttest_results": "1.5e+02",
       "https://w3id.org/trek/cohd/attribute/ttest_pvalue": "1.338936e-87"
    }
```

Live demo

Web UI to access the TReK Knowledge Graph through different interfaces (Reasoner API, SPARQL)

http://trek.semanticscience.org

API at http://api.trek.semanticscience.org

Documentation

https://d2s.semanticscience.org

Tool on PyPi

https://pypi.org/project/d2s

```
pip install d2s
```

Acknowledgments

NIH NCATS Translator support

For the funding over the past 3.5 years! Hackathons, weekly calls and many informative discussions to improve data interoperability and develop our solution

Columbia Open Health Data

For the successful collaboration in exposing clinical data insights

Ghent University

For the tools they developed for the RDF Mapping Language

Ontotext

For providing a GraphDB enterprise license

Alexander Malic

for his contribution to the ETL pipelines and architecture

Future developments