OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

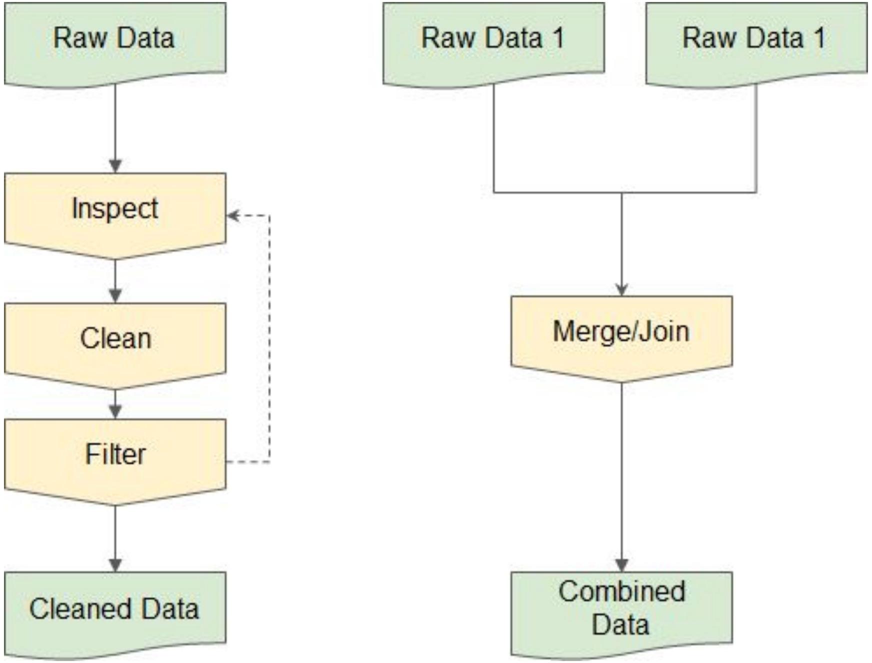
DataCure

Data curation and creation of pre-reasoned datasets and searching

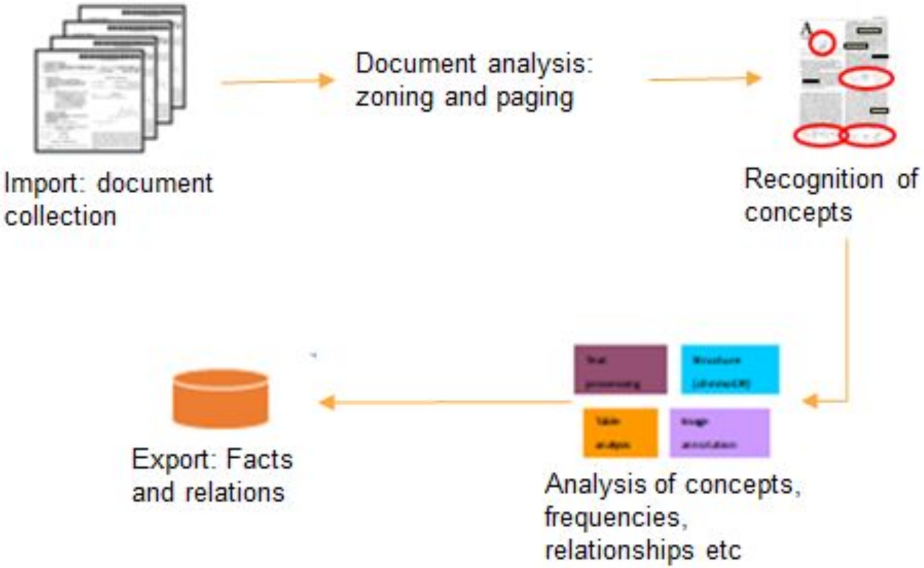
Noffisat Oki, Tim Dudgeon, Marc Jacobs, Danyel Jennen, Thomas Exner

OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment
Project Number 731075

Case Study objective



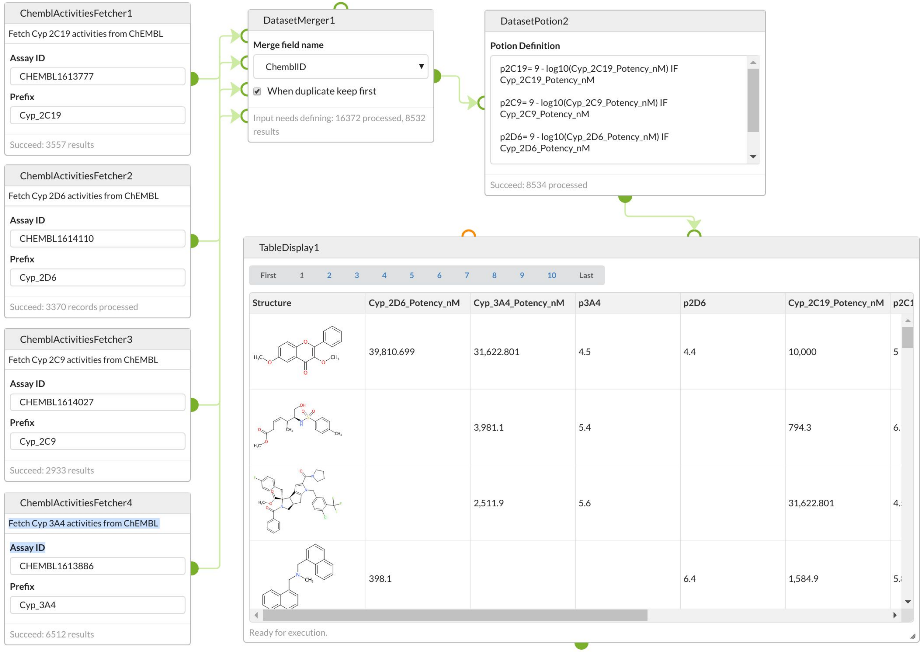
Data curation and merging


Import: document collection
Document analysis: zoning and paging
Recognition of concepts
Analysis of concepts, frequencies, relationships etc
Export: Facts and relations

Text mining

CypP450 data curation with Squonk

- Merge multiple datasets from ChEMBL into single set
- Uses ChEMBL identifiers to identify common structures
- Generates a dataset that can be used for machine learning
- See on GitHub



ChemblActivitiesFetcher1

Fetch Cyp 2C19 activities from ChEMBL

Assay ID

CHEMBL1613777

Prefix

Cyp_2C19

Succeed: 3557 results

ChemblActivitiesFetcher2

Fetch Cyp 2D6 activities from ChEMBL

Assay ID

CHEMBL1614110

Prefix

Cyp_2D6

Succeed: 3370 records processed

ChemblActivitiesFetcher3

Fetch Cyp 2C9 activities from ChEMBL

Assay ID

CHEMBL1614027

Prefix

Cyp_2C9

Succeed: 2933 results

ChemblActivitiesFetcher4

Fetch Cyp 3A4 activities from ChEMBL

Assay ID

CHEMBL1613886

Prefix

Cyp_3A4

Succeed: 6512 results

DatasetMerger1

Merge field name

ChemblID ▾

☑ When duplicate keep first

Input needs defining: 16372 processed, 8532 results

DatasetPotion2

Potion Definition

p2C19= 9 - log10(Cyp_2C19_Potency_nM) IF Cyp_2C19_Potency_nM

p2C9= 9 - log10(Cyp_2C9_Potency_nM) IF Cyp_2C9_Potency_nM

p2D6= 9 - log10(Cyp_2D6_Potency_nM) IF Cyp_2D6_Potency_nM

Succeed: 8534 processed

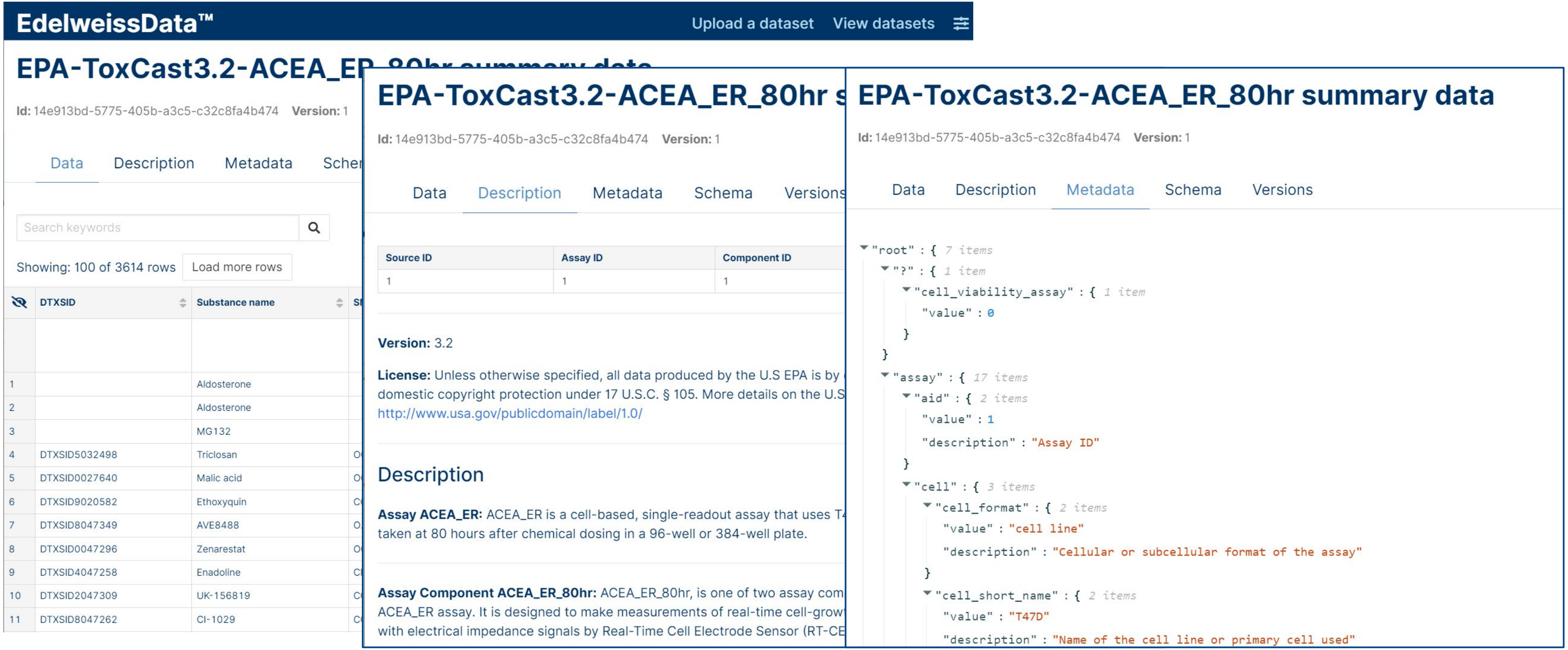
TableDisplay1

First 1 2 3 4 5 6 7 8 9 10 Last

Structure	Cyp_2D6_Potency_nM	Cyp_3A4_Potency_nM	p3A4	p2D6	Cyp_2C19_Potency_nM	p2C1
	39,810.699	31,622.801	4.5	4.4	10,000	5
		3,981.1	5.4		794.3	6.
		2,511.9	5.6		31,622.801	4.
	398.1			6.4	1,584.9	5.

Ready for execution.

Data merging via data APIs



EPA-ToxCast3.2-ACEA_ER_80hr summary data

Id: 14e913bd-5775-405b-a3c5-c32c8fa4b474 **Version:** 1

Showing: 100 of 3614 rows

	DTXSID	Substance name	SI
1		Aldosterone	
2		Aldosterone	
3		MG132	
4	DTXSID5032498	Triclosan	O
5	DTXSID0027640	Malic acid	O
6	DTXSID9020582	Ethoxyquin	C
7	DTXSID8047349	AVE8488	O
8	DTXSID0047296	Zenarestat	O
9	DTXSID4047258	Enadoline	C
10	DTXSID2047309	UK-156819	C
11	DTXSID8047262	CI-1029	C

EPA-ToxCast3.2-ACEA_ER_80hr s

Id: 14e913bd-5775-405b-a3c5-c32c8fa4b474 **Version:** 1

Source ID	Assay ID	Component ID
1	1	1

Version: 3.2

License: Unless otherwise specified, all data produced by the U.S EPA is by domestic copyright protection under 17 U.S.C. § 105. More details on the U.S http://www.usa.gov/publicdomain/label/1.0/

Description

Assay ACEA_ER: ACEA_ER is a cell-based, single-readout assay that uses T4 taken at 80 hours after chemical dosing in a 96-well or 384-well plate.

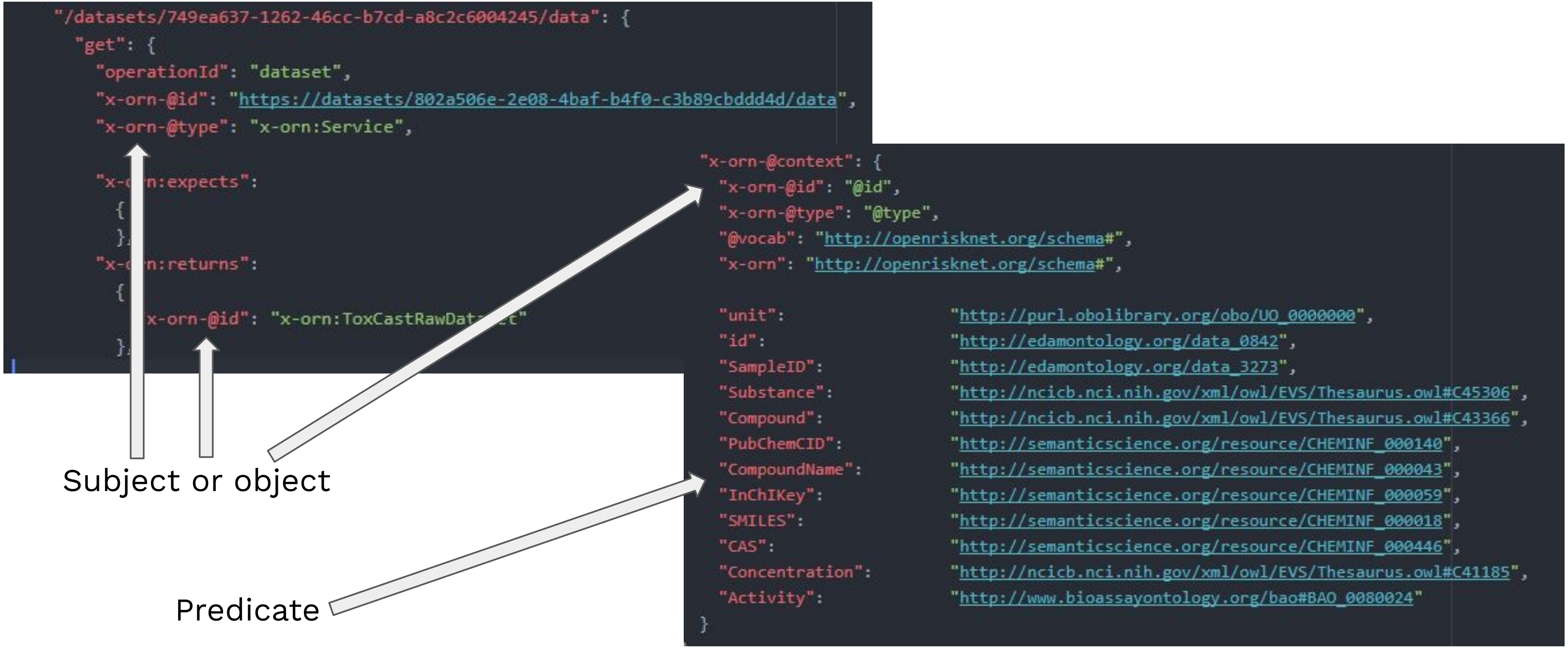
Assay Component ACEA_ER_80hr: ACEA_ER_80hr, is one of two assay com ACEA_ER assay. It is designed to make measurements of real-time cell-grow with electrical impedance signals by Real-Time Cell Electrode Sensor (RT-CE

EPA-ToxCast3.2-ACEA_ER_80hr summary data

Id: 14e913bd-5775-405b-a3c5-c32c8fa4b474 **Version:** 1

```
▼"root" : { 7 items
  ▼"?" : { 1 item
    ▼"cell_viability_assay" : { 1 item
       "value" : 0
    }
  }
  ▼"assay" : { 17 items
    ▼"aid" : { 2 items
       "value" : 1
       "description" : "Assay ID"
    }
    ▼"cell" : { 3 items
      ▼"cell_format" : { 2 items
         "value" : "cell line"
         "description" : "Cellular or subcellular format of the assay"
      }
      ▼"cell_short_name" : { 2 items
         "value" : "T47D"
         "description" : "Name of the cell line or primary cell used"
```

OpenAPI + JSON-LD



```
  "/datasets/749ea637-1262-46cc-b7cd-a8c2c6004245/data": {
    "get": {
      "operationId": "dataset",
      "x-orn-@id": "https://datasets/802a506e-2e08-4baf-b4f0-c3b89cbddd4d/data",
      "x-orn-@type": "x-orn:Service",

      "x-orn:expects":
        {
        },
      "x-orn:returns":
        {
          "x-orn-@id": "x-orn:ToxCastRawDataset"
        },
```

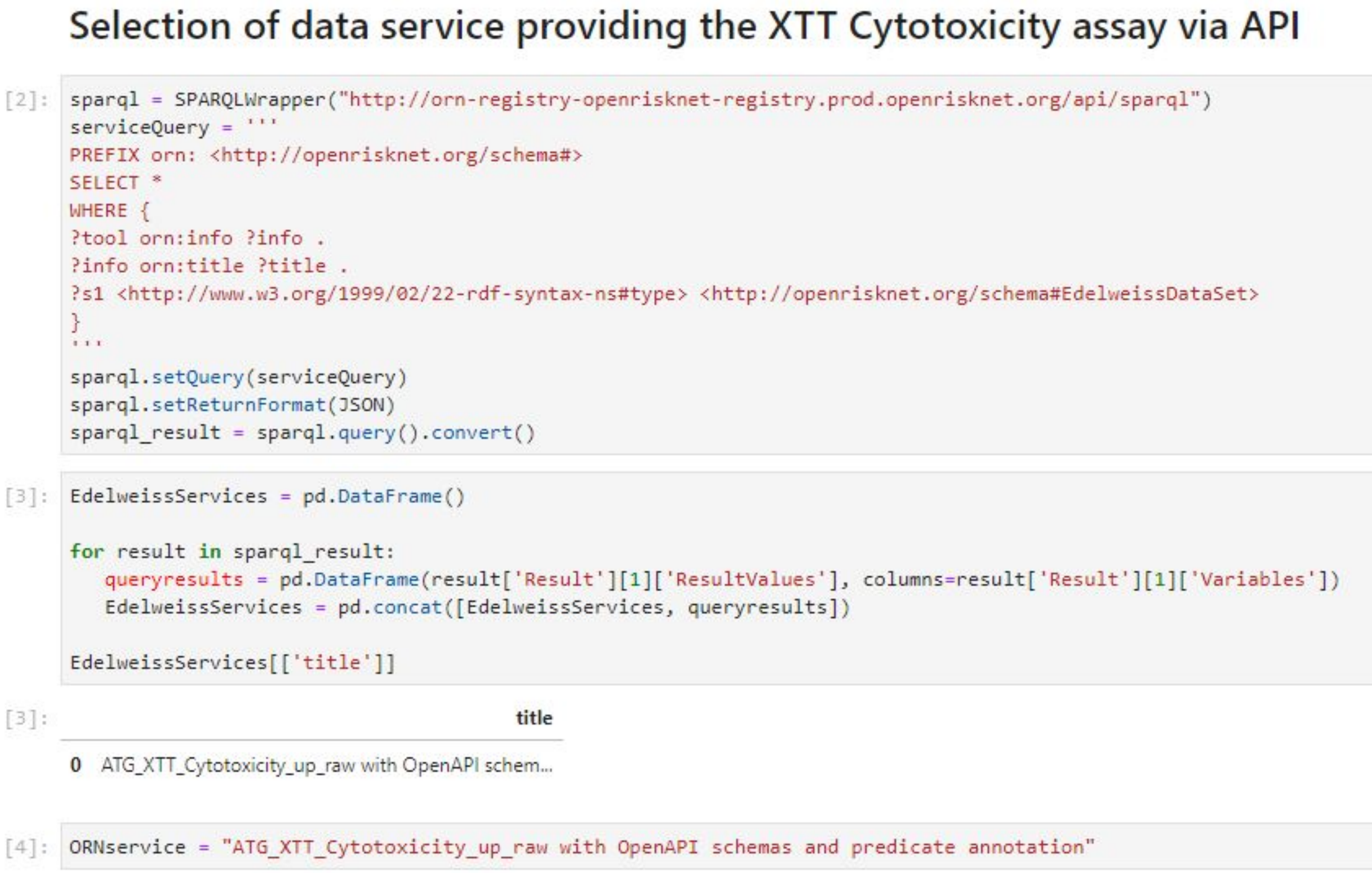
```
"x-orn-@context": {
  "x-orn-@id": "@id",
  "x-orn-@type": "@type",
  "@vocab": "http://openrisknet.org/schema#",
  "x-orn": "http://openrisknet.org/schema#",

  "unit":                    "http://purl.obolibrary.org/obo/UO_0000000",
  "id":                      "http://edamontology.org/data_0842",
  "SampleID":                "http://edamontology.org/data_3273",
  "Substance":               "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C45306",
  "Compound":                "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43366",
  "PubChemCID":              "http://semanticscience.org/resource/CHEMINF_000140",
  "CompoundName":            "http://semanticscience.org/resource/CHEMINF_000043",
  "InChIKey":                "http://semanticscience.org/resource/CHEMINF_000059",
  "SMILES":                  "http://semanticscience.org/resource/CHEMINF_000018",
  "CAS":                     "http://semanticscience.org/resource/CHEMINF_000446",
  "Concentration":           "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C41185",
  "Activity":                "http://www.bioassayontology.org/bao#BAO_0080024"
}
```

Subject or object

Predicate

Finding datasets



Selection of data service providing the XTT Cytotoxicity assay via API

[2]:

```
sparql = SPARQLWrapper("http://orn-registry-openrisknet-registry.prod.openrisknet.org/api/sparql")
serviceQuery = '''
PREFIX orn: <http://openrisknet.org/schema#>
SELECT *
WHERE {
?tool orn:info ?info .
?info orn:title ?title .
?s1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://openrisknet.org/schema#EdelweissDataSet>
}
'''
sparql.setQuery(serviceQuery)
sparql.setReturnFormat(JSON)
sparql_result = sparql.query().convert()
```

[3]:

```
EdelweissServices = pd.DataFrame()

for result in sparql_result:
   queryresults = pd.DataFrame(result['Result'][1]['ResultValues'], columns=result['Result'][1]['Variables'])
   EdelweissServices = pd.concat([EdelweissServices, queryresults])

EdelweissServices[['title']]
```

[3]:

	title
0	ATG_XTT_Cytotoxicity_up_raw with OpenAPI schem...

[4]:

```
ORNservice = "ATG_XTT_Cytotoxicity_up_raw with OpenAPI schemas and predicate annotation"
```

Text mining

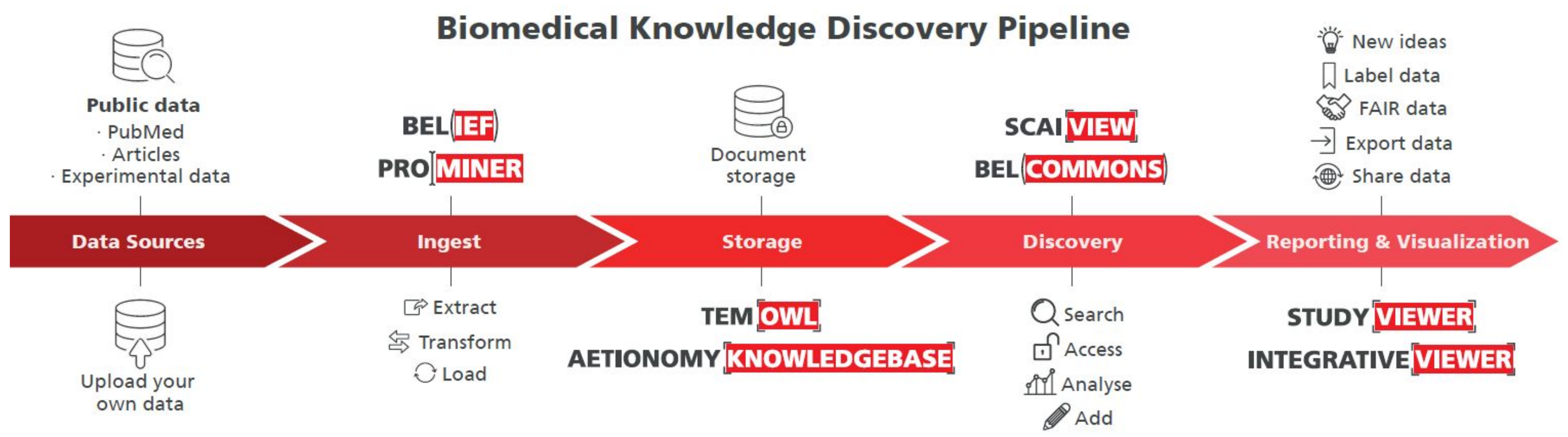
API

K8s hosted:

- https://api.scaiview.com/swagger-ui.html
- https://sam1.api.scaiview.com/swagger-ui.html

Workflows

jupyter https://biohub.scai.fraunhofer.de



Biomedical Knowledge Discovery Pipeline

Data Sources → Ingest → Storage → Discovery → Reporting & Visualization

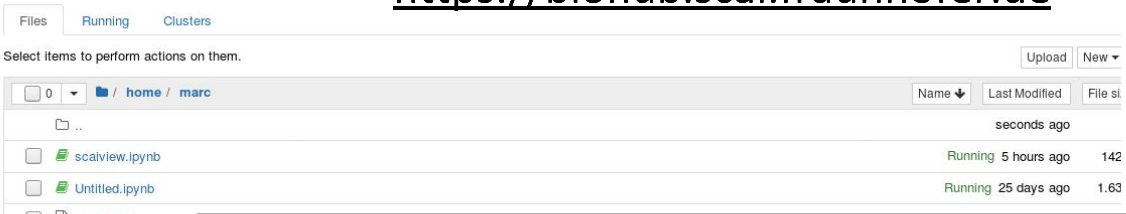
Data Sources: Public data · PubMed · Articles · Experimental data; Upload your own data

Ingest: BEL IEF, PRO MINER; Extract, Transform, Load

Storage: Document storage; TEM OWL, AETIONOMY KNOWLEDGEBASE

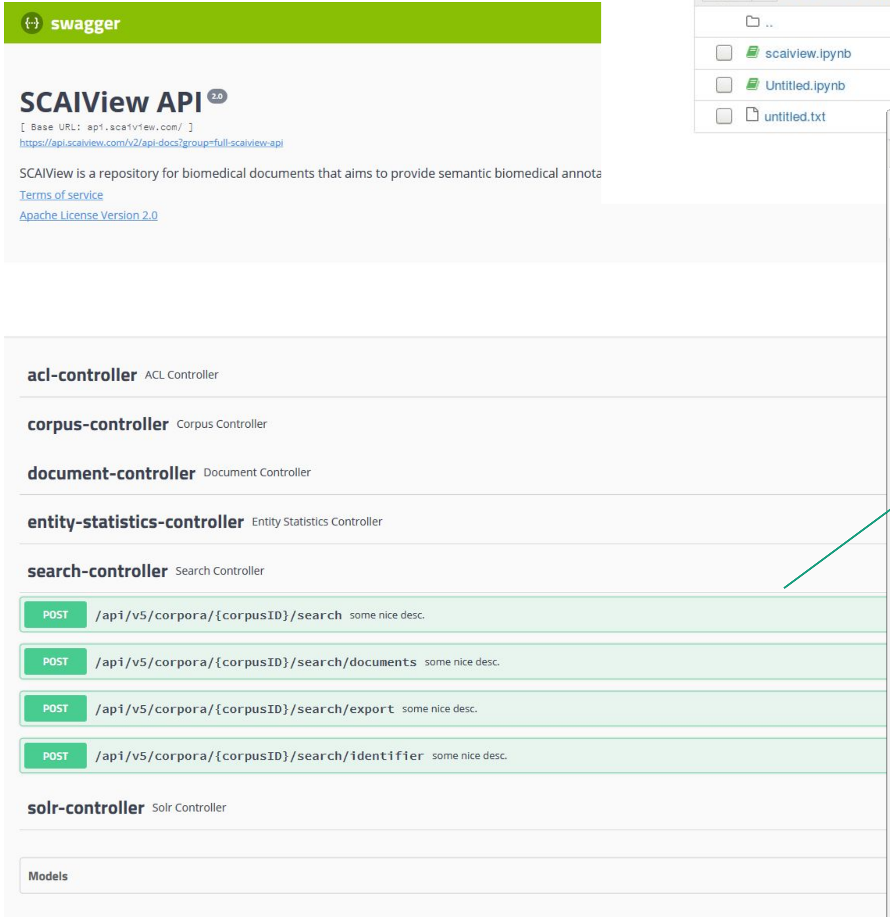
Discovery: SCAI VIEW, BEL COMMONS; Search, Access, Analyse, Add

Reporting & Visualization: New ideas, Label data, FAIR data, Export data, Share data; STUDY VIEWER, INTEGRATIVE VIEWER https://biohub.scai.fraunhofer.de



Select items to perform actions on them.

	Name	Last Modified	File si
□ 0	/ home / marc		
	..	seconds ago	
□	scaiview.ipynb	Running 5 hours ago	142
□	Untitled.ipynb	Running 25 days ago	1.63
□	untitled.txt		



SCAIView API 2.0

[Base URL: api.scaiview.com/]

https://api.scaiview.com/v2/api-docs?group=full-scaiview-api

SCAIView is a repository for biomedical documents that aims to provide semantic biomedical annota

Terms of service

Apache License Version 2.0

acl-controller ACL Controller

corpus-controller Corpus Controller

document-controller Document Controller

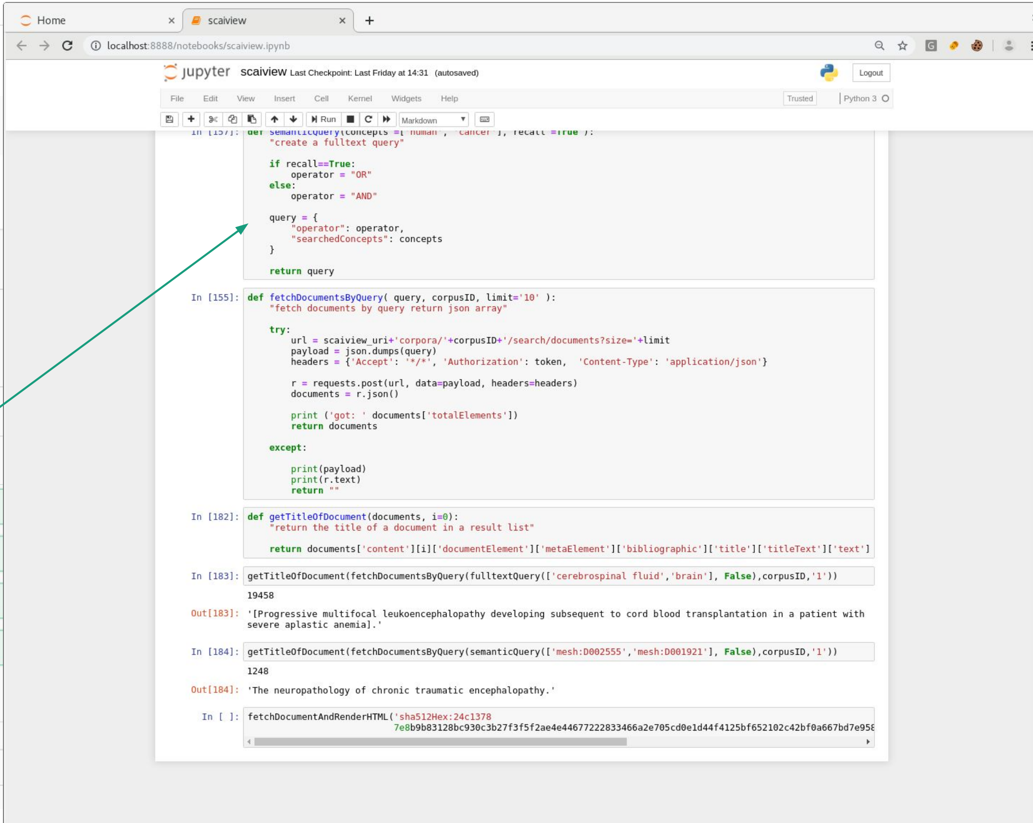
entity-statistics-controller Entity Statistics Controller

search-controller Search Controller

- POST `/api/v5/corpora/{corpusID}/search` some nice desc.
- POST `/api/v5/corpora/{corpusID}/search/documents` some nice desc.
- POST `/api/v5/corpora/{corpusID}/search/export` some nice desc.
- POST `/api/v5/corpora/{corpusID}/search/identifier` some nice desc.

solr-controller Solr Controller

Models



scaiview Last Checkpoint: Last Friday at 14:31 (autosaved)

```
In [157]: def semanticQuery(concepts =[ 'human', 'cancer' ], recall =True ):
              "create a fulltext query"

              if recall==True:
                  operator = "OR"
              else:
                  operator = "AND"

              query = {
                  "operator": operator,
                  "searchedConcepts": concepts
              }

              return query

In [155]: def fetchDocumentsByQuery( query, corpusID, limit='10' ):
              "fetch documents by query return json array"

              try:
                  url = scaiview_uri+'corpora/'+corpusID+'/search/documents?size='+limit
                  payload = json.dumps(query)
                  headers = {'Accept': '*/*', 'Authorization': token,  'Content-Type': 'application/json'}

                  r = requests.post(url, data=payload, headers=headers)
                  documents = r.json()

                  print ('got: ' documents['totalElements'])
                  return documents

              except:

                  print(payload)
                  print(r.text)
                  return ""

In [182]: def getTitleOfDocument(documents, i=0):
              "return the title of a document in a result list"

              return documents['content'][i]['documentElement']['metaElement']['bibliographic']['title']['titleText']['text']

In [183]: getTitleOfDocument(fetchDocumentsByQuery(fulltextQuery(['cerebrospinal fluid','brain'], False),corpusID,'1'))

          19458

Out[183]: '[Progressive multifocal leukoencephalopathy developing subsequent to cord blood transplantation in a patient with
          severe aplastic anemia].'

In [184]: getTitleOfDocument(fetchDocumentsByQuery(semanticQuery(['mesh:D002555','mesh:D001921'], False),corpusID,'1'))

          1248

Out[184]: 'The neuropathology of chronic traumatic encephalopathy.'

In [ ]:   fetchDocumentAndRenderHTML('sha512Hex:24c1378
                                      7e8b9b83128bc930c3b27f3f5f2ae4e44677222833466a2e705cd0e1d44f4125bf652102c42bf0a667bd7e958
```

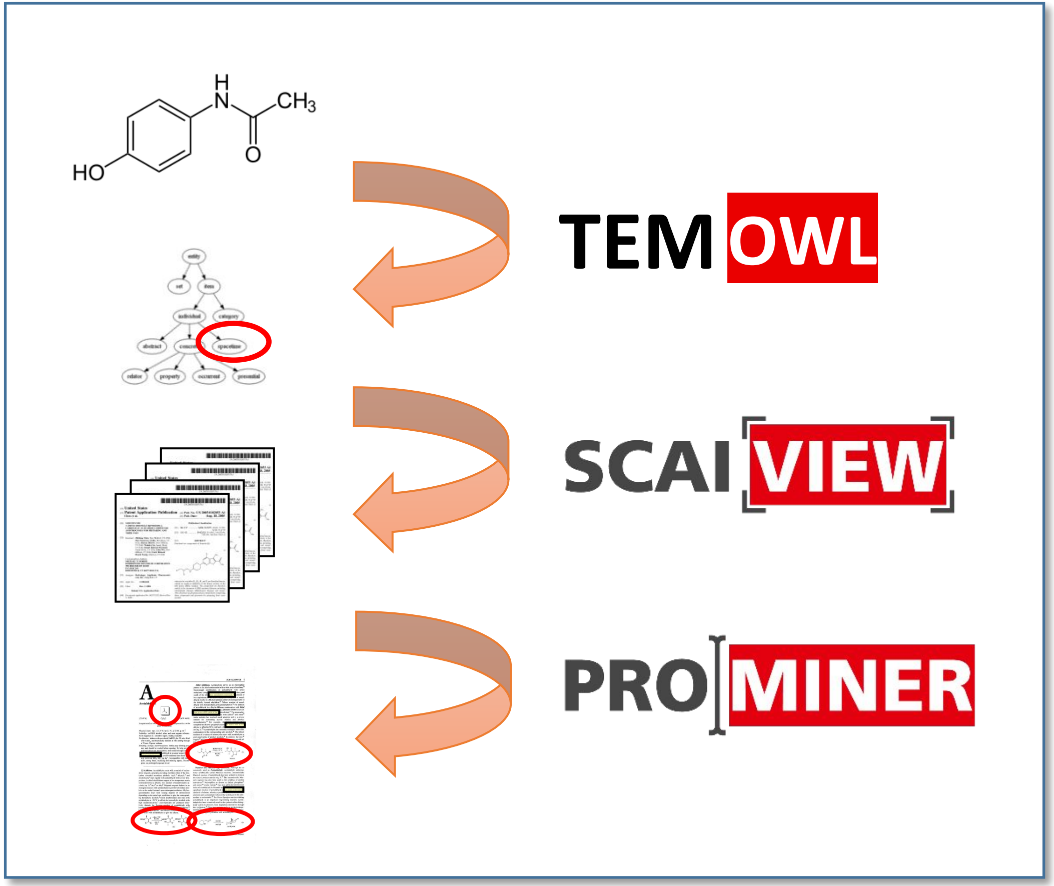
OpenRiskNet example workflow

Task:

- Identify the concept of acetaminophen (definition, identifiers, synonyms)
- Find all relevant documents in the context of acetaminophen and carcinogenity
- What are the most relevant statements

Technology:

- Semantic index of PubMed/PMC (> 20 terminologies)
- Solr index + OLS index + UIMA pipeline


HO
H
N
CH3
O
TEM OWL
SCAI VIEW
PRO MINER