Base Platform for Knowledge Graphs with Free Software

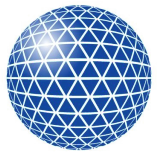
Simon Bin, Claus Stadler, Norman Radtke, Kurt Junghanns,
Sabine Gründer-Fahrer, Michael Martin

D2R2'23 @ ESWC 2023 **May 28th, 2023**

Supported by

Federal Ministry
for Economic Affairs
and Climate Action

GA No. 01MK21007A



InfAI®
Institute for Applied Informatics

Project Vision

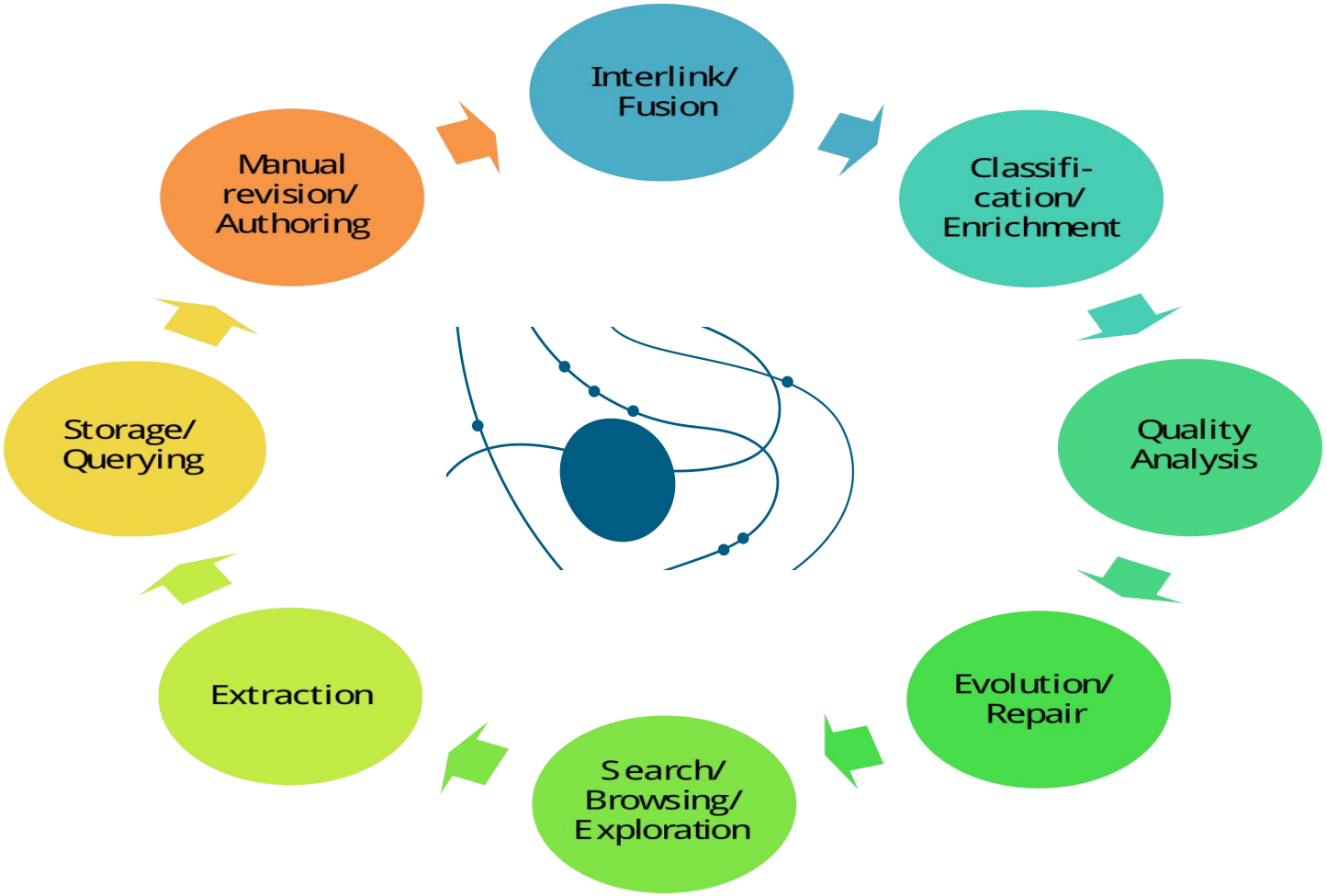
“Cognitive Economy Intelligence Platform for resilience of economic ecosystems”

COYPU

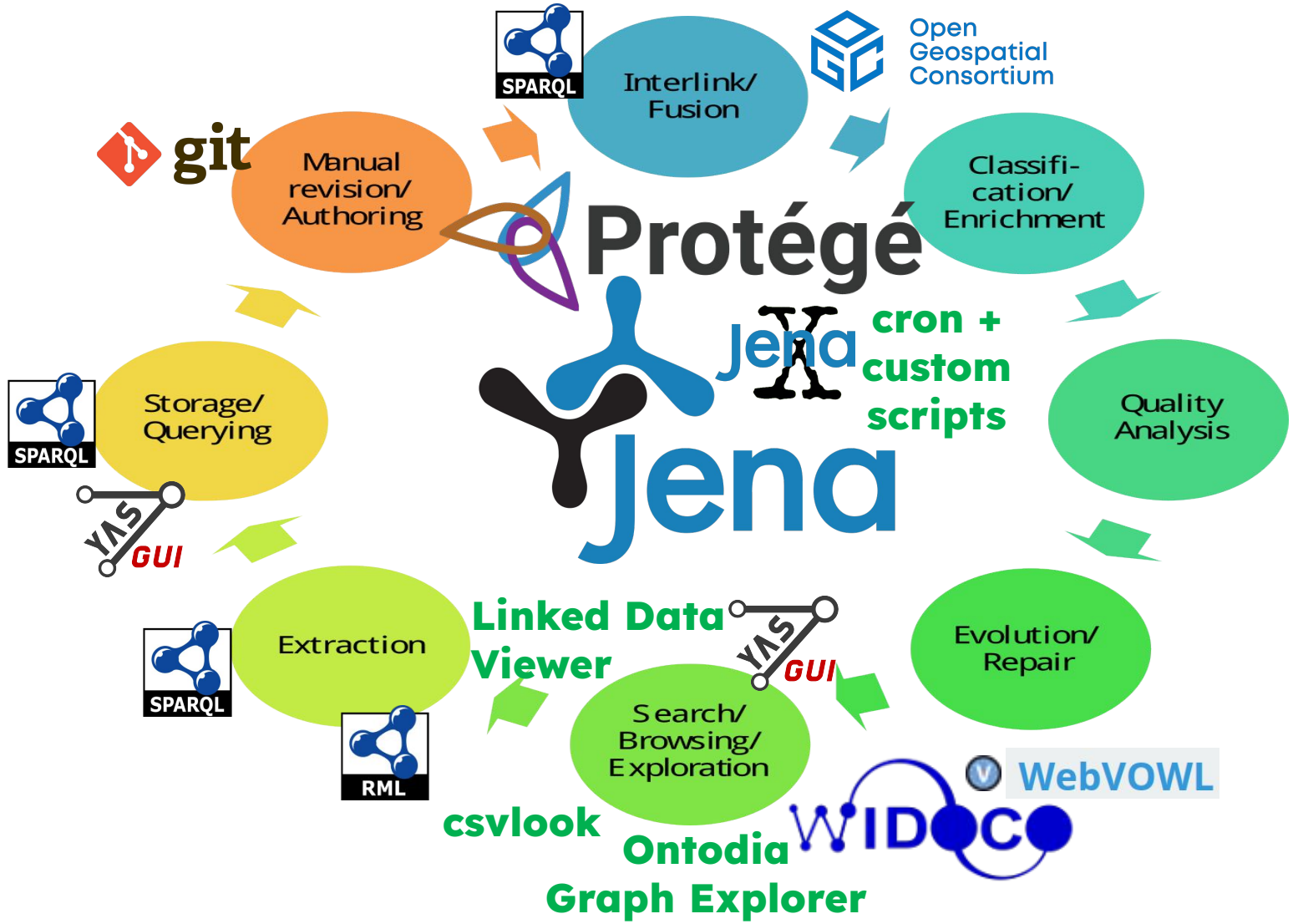
- Intelligent Platform for the integration, structuring, networking, analysis and evaluation of heterogeneous data
- cognitive modelling of data and flexibly configurable analysis tools
- qualitative, up-to-the-minute insights into facts, trends, interdependencies and forecasts for individual value networks
- Analyses for different regions, industries and company sizes as well as on a macroeconomic level
- Information transparency for companies and markets, especially in the SME sector

⇒ **Resilience of existing value chains**

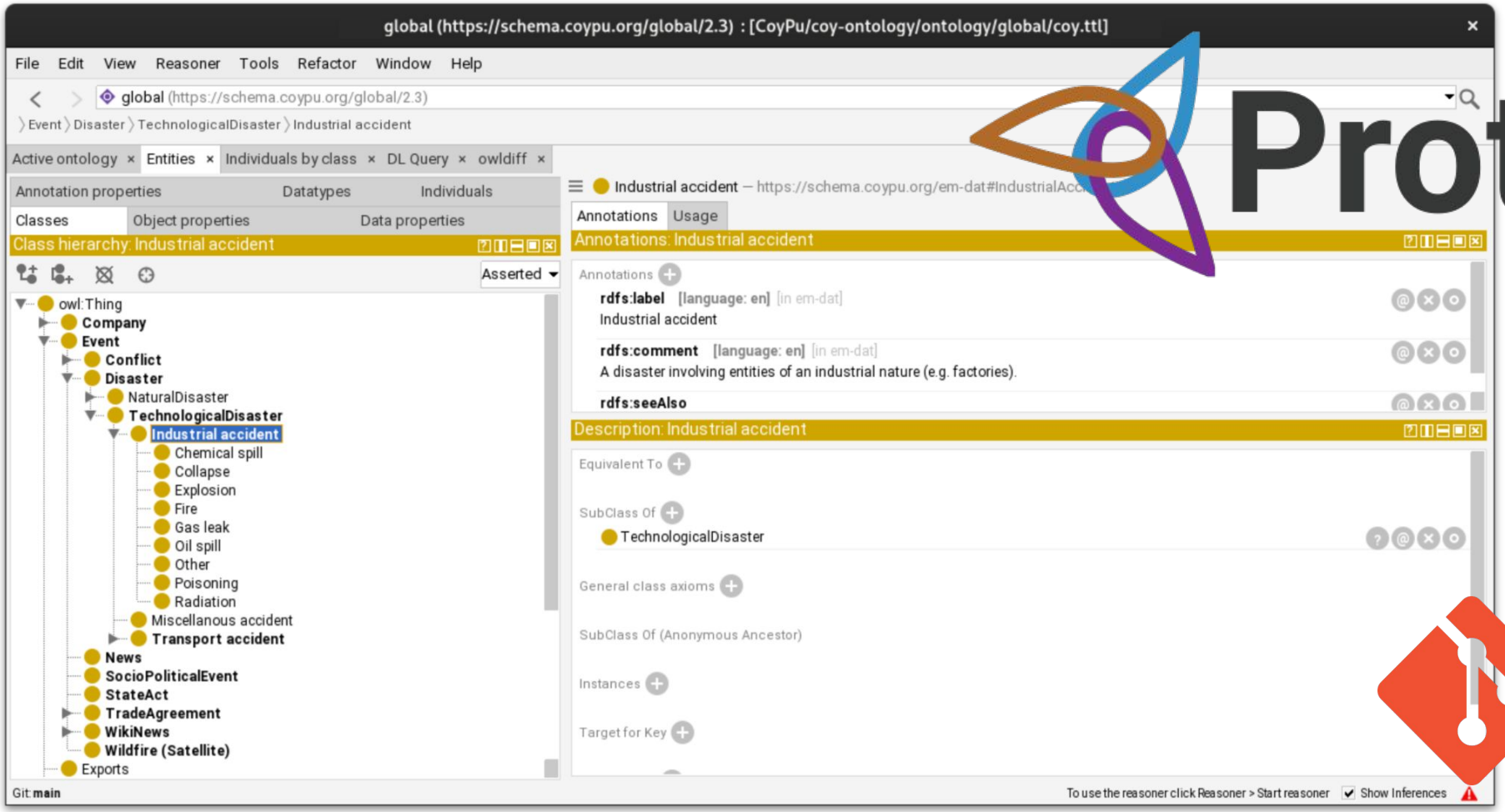
Linked Data Lifecycle



Linked Data Lifecycle



Ontology creation



Protégé

git

Ontology documentation

WIDOCO

owl documentation environment

WebVOWL

Ontology Specification Draft

Customer[C] back to ToC or Class ToC

IRI: https://schema.coypu.org/global#Customer

An organization that sources something needed such as a product or service from another organization.

Term status

CoyPu graph: Not currently in use

has super-classes

Company [C]

identity militias[C] back to ToC or Class ToC

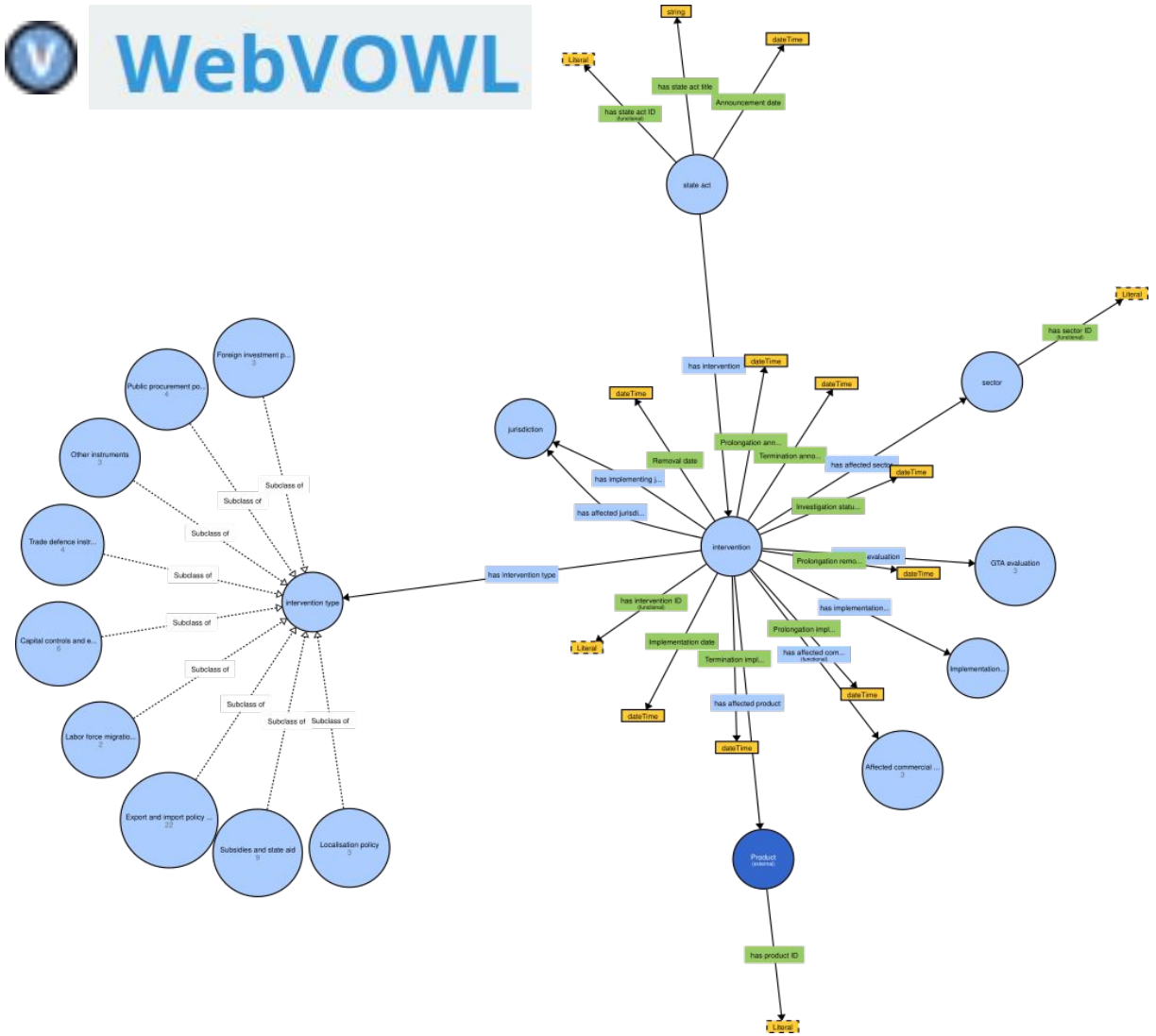
IRI: https://schema.coypu.org/global#IdentityMilitias

Term status

CoyPu graph: In use, distinct subjects: 7977

Term rationale

CoyPu graph: ---USED BUT NOT IN ONTOLOGY---



Data catalogue

COYPU

Data Sources Overview

- CoyPu technical docs
- Data Sources main repository

```
:airports
    dct:license             <https://metadata.coypu.org/licenses/CC-BY_v4> ;
    dcat:landingPage        <https://datacatalog.worldbank.org/search/dataset/0038117/Global-Airports> ;
    t:graph                 <https://data.coypu.org/infrastructure/airports/> ;
    t:licenseNote           "CC-BY-4.0 by Worldbank" ;
    t:licensePage           <https://datacatalog.worldbank.org/search/dataset/0038117/Global-Airports> ;
    t:theme                 "Global airports" ;
    t:type                  "Dump"
.
```

Graphs & Data

This is a list of Data Sources loaded into `data.coypu.org` on Fuseki.

Static

Dataset	Named Graph	Domain
AIRPORTS	https://data.coypu.org/infrastructure/airports/	Global airports
CITIES	https://data.coypu.org/cities/	Global cities
CLIMATETRACE	https://data.coypu.org/infrastructure/climatetrace/	Natural resource extraction and processing
CORRUPTION	https://data.coypu.org/country-statistic/corruption/	Corruption perception index per country
COUNTRIES	https://data.coypu.org/countries/	Counties and regions
DALICC	http://dalicc.net/licenselibrary/	License graph from DaliCC framework

(SPARQL +)

csvlook

(+ Markdown renderer)

Licences in the data catalogue

Talk by **Kurt Junghanns**: *Efficient Use of DALICC in Data Processing Pipelines with Fuzzy License Information*.

Today 17:00–17:20 in Orpheas

RML Mapping Language

```
<map_stops_0> a rr:TriplesMap ;
```

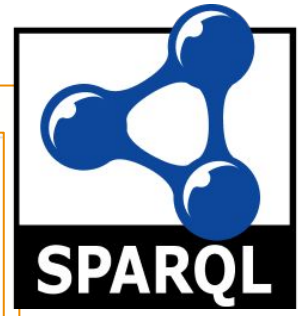
```
rml:logicalSource [ a rml:LogicalSource ;
  rml:referenceFormulation ql:CSV ;
  rml:source "STOPS.csv"
] ;
```

```
rr:subjectMap [ a rr:SubjectMap ;
  rr:template "http://example.org/stops/{stop_id}"
] ;
```

```
rr:predicateObjectMap  [ a rr:PredicateObjectMap ;
  rr:predicateMap  [ a rr:PredicateMap ;
    rr:constant  wgs84:long
  ] ;
  rr:objectMap [ a rr:ObjectMap ;
    rml:reference  "stop_lon" ;
      rr:datatype    xsd:double ;
      rr:termType    rr:Literal
  ] ;
] ;
```

```
rr:predicateObjectMap  [ a rr:PredicateObjectMap ;
  rr:predicateMap  [ a rr:PredicateMap ;
    rr:constant  <http://vocab.gtfs.org/terms#parentStation>
  ] ;
  rr:objectMap [ a rr:ObjectMap ;
    rr:joinCondition [ rr:child   "parent_station" ;
                       rr:parent  "stop_id" ] ;
    rr:parentTriplesMap  <map_stops_0>
  ] ;
] .
```

Mapping data using Tarql and Javascript



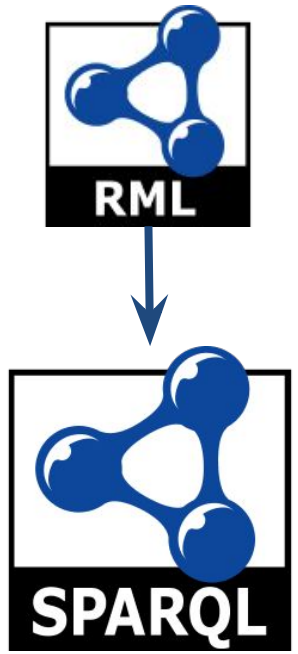

```
    ?actor_1
      rdfs:label ?actor1_label ;
      rdf:type ?actor_type ;
      .

      ?actor_2
      rdfs:label ?actor2_label ;
      rdf:type ?actor2_type ;
      .

}
WHERE {
  # JS function to parse event date in format 'dd MMMM yyyy' to ISO format
  #BIND(json:js("function(x) {return new Date(x).toISOString();}", ?event_date) AS ?date)
  BIND(xsd:date(STRDT(REPLACE(js:toIsoDateString(?event_date), "Z", ""), xsd:dateTime)) AS ?date)

  BIND(IRI(CONCAT(str(data:), ?event_id_cnty)) AS ?event)

  # event subtype
  BIND(IRI(CONCAT(str(acled:), REPLACE(js:toTitleCase(REPLACE(?sub_event_type, "-", " ")), " ", ""))) AS ?type)

  # geo lon/lat
  BIND(STRDT(?latitude, xsd:double) AS ?lat)
  BIND(STRDT(?longitude, xsd:double) AS ?lon)
```

```
function toIsoDateString(str) {
  return new Date(str).toISOString();
}

function toTitleCase(str) {
  return str.toLowerCase().replace(/\b\w/g, s => s.toUpperCase());
}
```

Executing RML Mappings as parallel SPARQL queries

Talk by **Claus Stadler**: *Scaling RML and SPARQL-based Knowledge Graph Construction with Apache Spark*.

Today 17:20–18:00 in Hermes



Mapping data using JSONL + GeoSPARQL

Open Geospatial Consortium

```
CONSTRUCT {
  ?country
    geo:hasGeometry ?geometry_node .
  ?geometry_node a geo:Geometry ;
    geo:asWKT ?geometry .
}
WHERE
{
  {
    SELECT
    ?country
    ?geometry_node
    ?geometry
    {
      <env:INPUT> url:textLines ?text .
      BIND(STRDT(?text, xsd:json) AS ?item)

      BIND(json:path(?item, "$.properties") AS ?properties)

      BIND(json:path(?properties, "$.ISO_CODE") AS ?iso3)
      BIND(URI(concat("https://data.coypu.org/country/", ?iso3)) AS ?country)
      BIND(URI(concat(str(?country), "/geometry/boundary")) AS ?geometry_node)

      BIND(json:path(?item, "$.geometry") AS ?geometry_)
      BIND(spatialF:transformDatatype(STRDT(str(?geometry_), geo:geoJSONLiteral), geo:wktLiteral) AS ?geometry)
```

Executing RML Mappings as parallel SPARQL queries

Talk by **Lars Michaelis**: *WikiEvents - A Novel Resource for NLP Downstream Tasks*.

Tomorrow 16:55–17:15 in Polymnia & Kalia

“Pipeline”

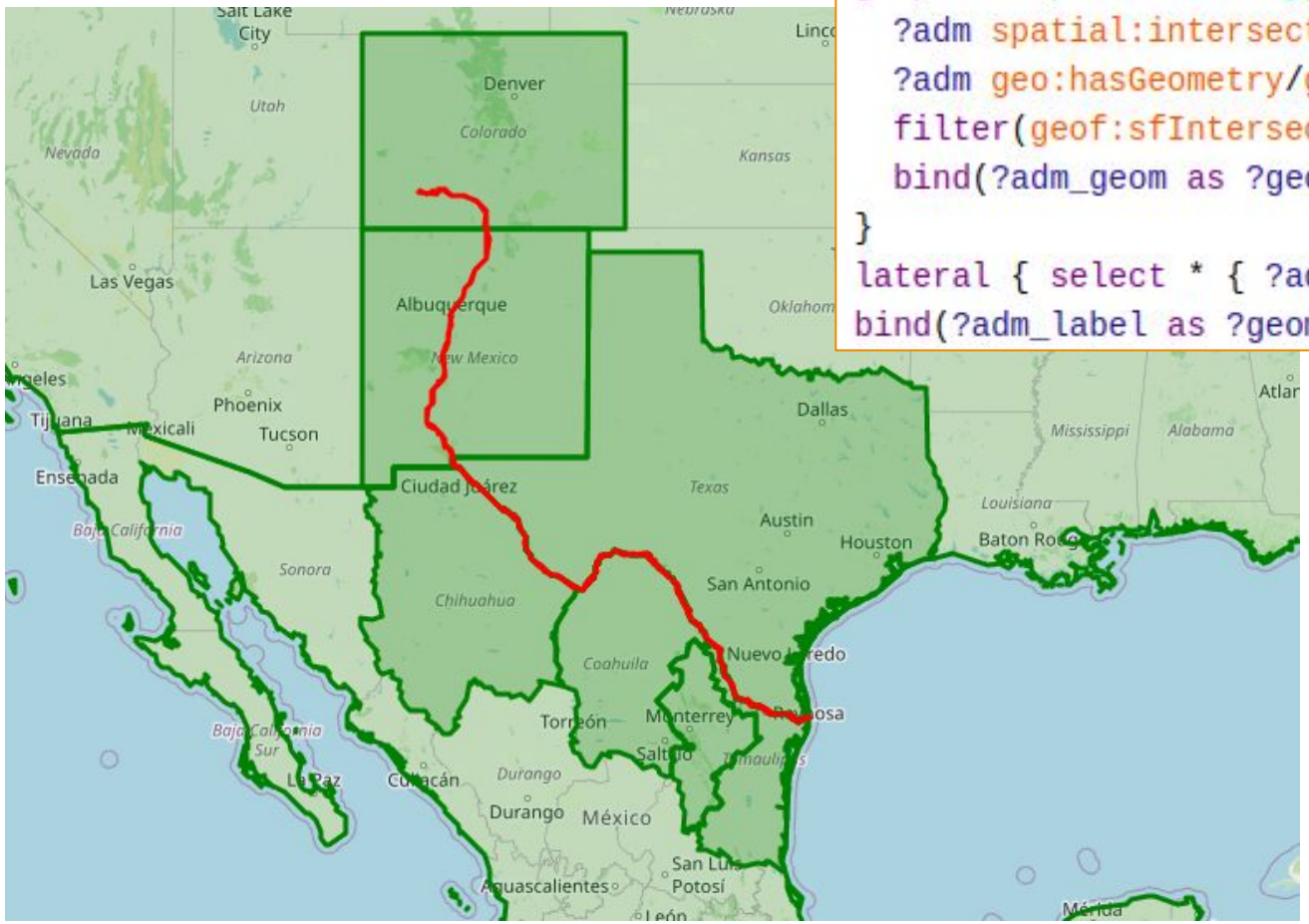
sbin@sems-kg-1 :: fg [~]

```
0 22 1-11,13-31 * 0,1,2,3,4,5,6 /data/coypu/tools/maintenance/live_reload_all_daily_and_recreate_tmp.sh
crontab [+]                                                                    1,1           All
cd


# run commodities fetcher&transformer
python /data/coypu/tools/commodities/main.py
python /data/coypu/tools/disasters/main.py
python /data/coypu/tools/wildfires/main.py

# extract daily wiki events
/data/coypu/tools/wikievents_loader/wikievents_loader_daily 2>&1 | tee /data/coypu/tools/log/wikievents_
loader_daily.log

"$script_root/live_reload_all_daily_from_disk.sh"



/data/coypu/tools/maintenance/live_reload_all_daily_and_recreate_tmp.sh [+]          28,0-1          Bot
```

cron + custom scripts

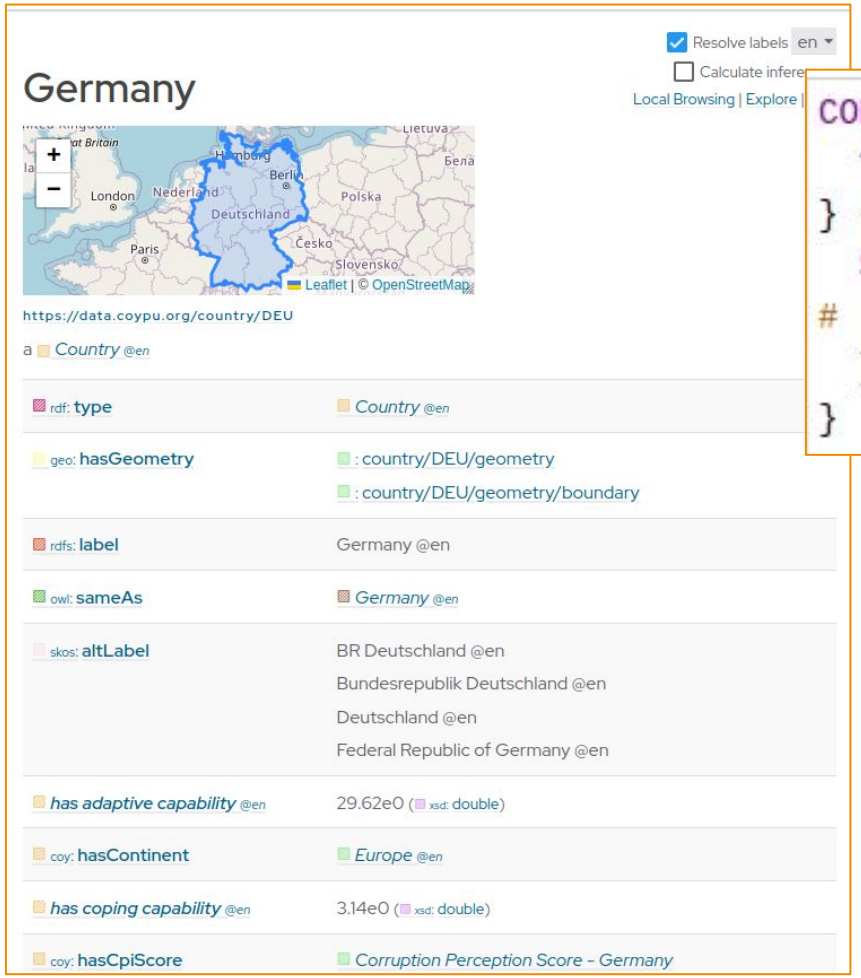
Triple Store + SPARQL Endpoint

Combining geodata in the graph




```
graph <https://data.coypu.org/administrative-regions/> {
  ?adm spatial:intersectBoxGeom(?flood_geom) .
  ?adm geo:hasGeometry/geo:asWKT ?adm_geom_lit_ .
  filter(geof:sfIntersects(?flood_geom, ?adm_geom_lit_))
  bind(?adm_geom as ?geom) bind("green" as ?geomColor)
}
lateral { select * { ?adm rdfs:label ?adm_label } limit 1 }
bind(?adm_label as ?geomTooltip) bind(?adm as ?geomLabel)
```

Calculate inferences

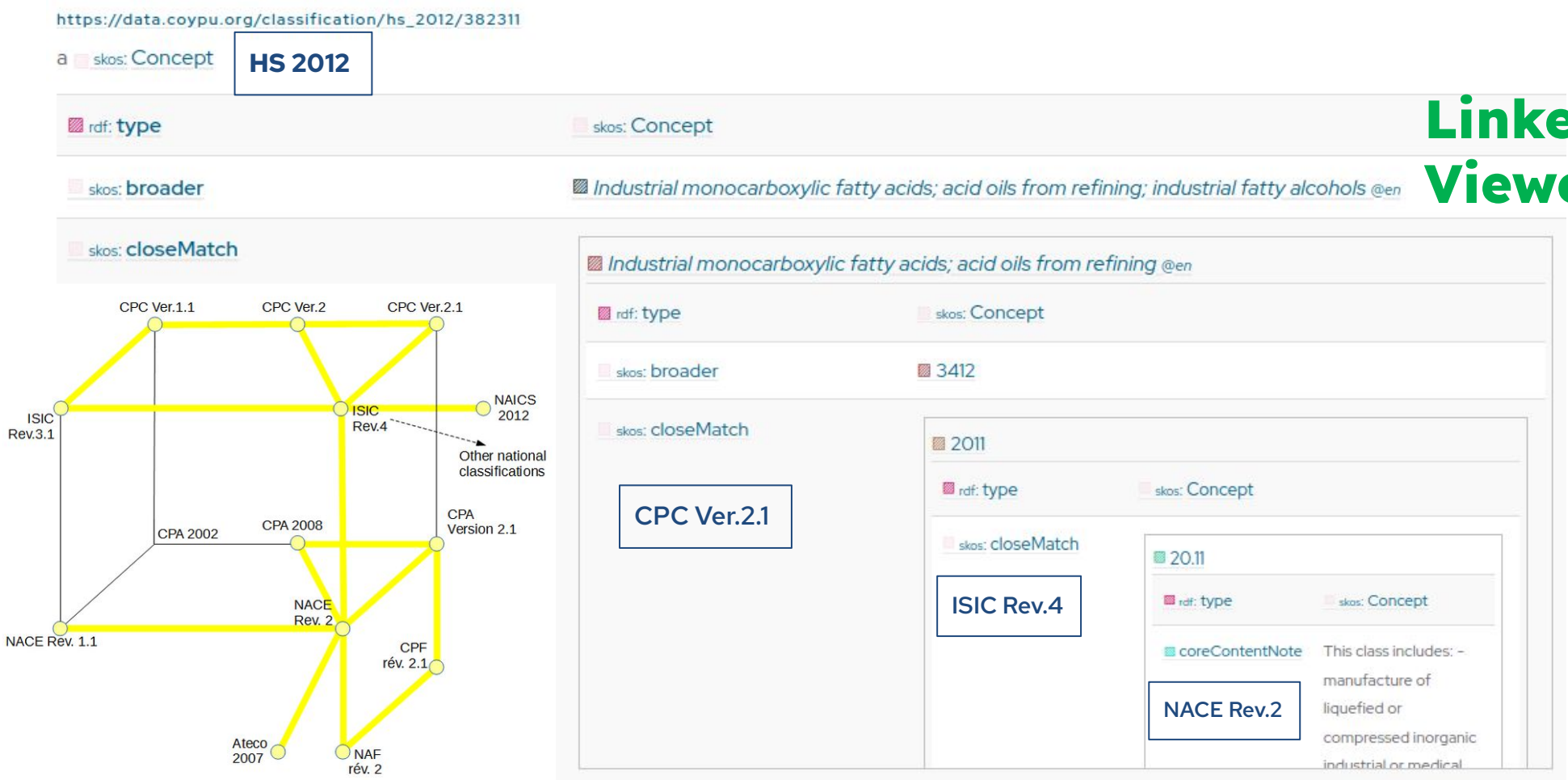



```
CONSTRUCT {
  ?s ?p ?o .
} {
  SERVICE <sameAs+rdfs:> {
# ...
  }
}
```

Jena

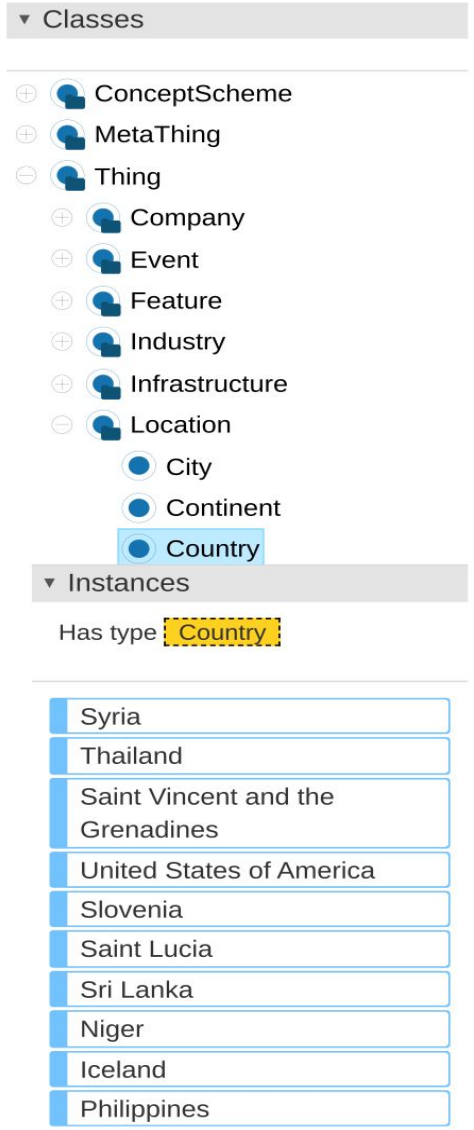
Linked Data Viewer

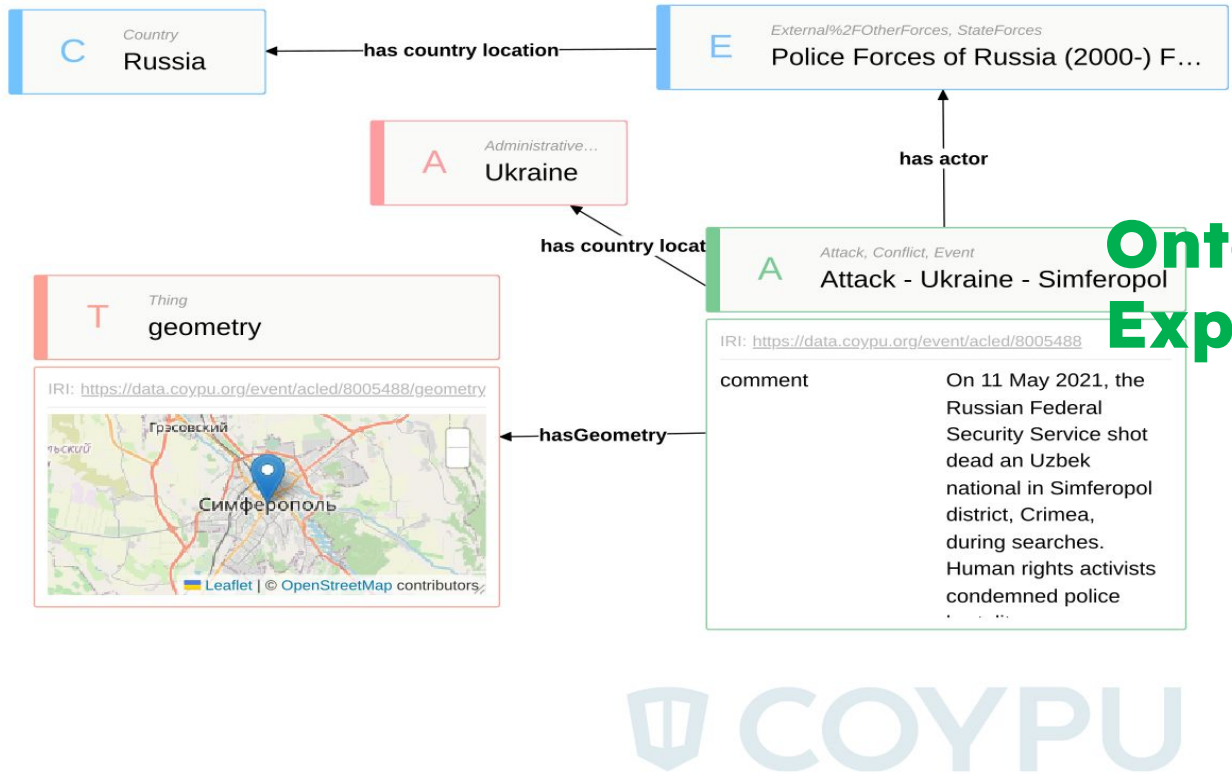
Mapping of product sectors



Linked Data Viewer

Interactive exploration of the graph

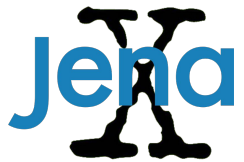




Ontodia Graph Explorer

Caching for fast retrieval of precalculated results

```
SELECT ?g (SUM(?c) AS ?count) WHERE {
    { SERVICE <cache:> { SELECT ("defaultGraph" AS ?g) ?p (COUNT(*) AS ?c) { ?s ?p ?o } GROUP BY ?p } }
  UNION
    { SERVICE <cache:> { SELECT ?g ?p (COUNT(*) AS ?c) { GRAPH ?g { ?s ?p ?o } } GROUP BY ?g ?p } }
}
GROUP BY ?g
ORDER BY DESC(?count)
```



5 minutes ⇨
0.6 seconds

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Markup | Network | Pivot | Timeline | 7 results in 0.597 seconds

Simple view ☐ Ellipse ☑ Filter query results Page size: 50

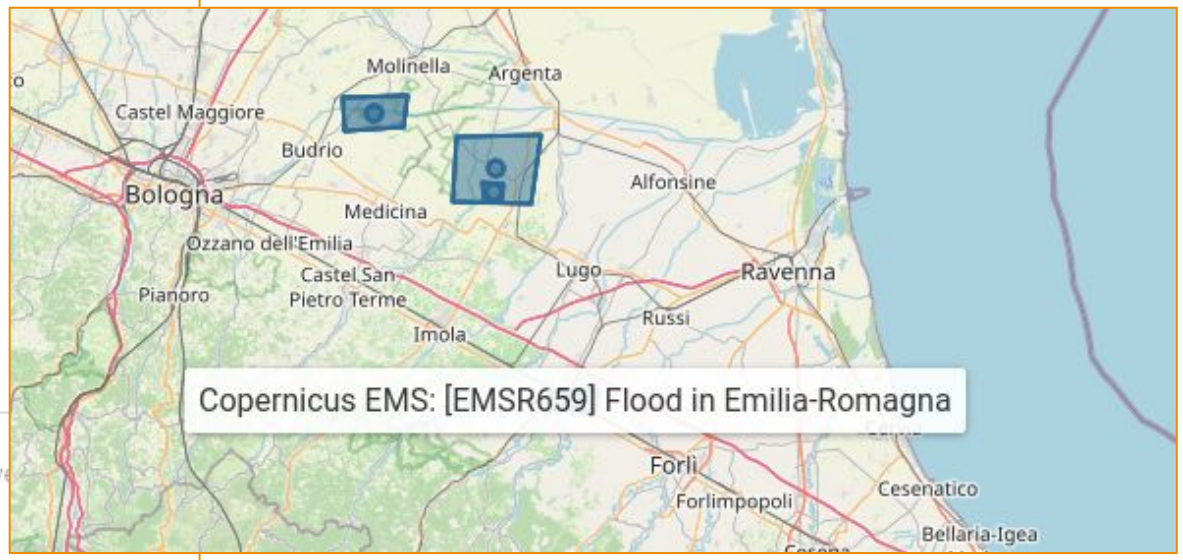
	g	count
1	defaultGraph	"208467870"^^<http://www.w3.org/2001/XMLSchema#integer>
2	<urn:mvn:org.coypu.thirdpary.osm:osm-river-planet:2022-10-05:ttl.bz2>	"163310838"^^<http://www.w3.org/2001/XMLSchema#integer>
3	<urn:mvn:org.coypu.thirdpary.osm:osm-railway-planet:2022-10-05:ttl.bz2>	"43235550"^^<http://www.w3.org/2001/XMLSchema#integer>
4	<urn:mvn:org.coypu.thirdpary.osm:osm-railwaystation-planet:2022-10-05:ttl.bz2>	"1570048"^^<http://www.w3.org/2001/XMLSchema#integer>
5	<urn:mvn:org.coypu.thirdparty:gadm-level-0-2:gadm:4.1:ttl.gz>	"334776"^^<http://www.w3.org/2001/XMLSchema#integer>
6	<urn:mvn:org.coypu.thirdparty.msi.nga.mil:world-port-index:2022-10-05:ttl>	"187363"^^<http://www.w3.org/2001/XMLSchema#integer>
7	<urn:mvn:org.coypu.thirdpary.osm:osm-ports-planet:2022-10-05:ttl.bz2>	"63213"^^<http://www.w3.org/2001/XMLSchema#integer>

Inline Javascript for API transformations

```
27    BIND(xml:path(?event_item, "//georss:polygon/text()") as ?polygon_points
28    BIND(json:jse("""
29        let arr = $0.split(' ')
30        let res=[]
31        for (var i = 0 ; i < arr.length - 1; i = i + 2) {
32           res.push(arr[i + 1] + ' ' + arr[i])
33        }
34        return res.join(', ')
35     """, ?polygon_points) AS ?value)
36    BIND("POLYGON((" + ?value + "))" as ?wkt)
37    BIND(strdt(?wkt, geo:wktLiteral) as ?geom)
38  }
39  LIMIT 100
40
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo ev

	geom
1	"POLYGON((11.77606211830873 44.48668204095276, 11.80291361887779 44.4849722331955...
2	"POLYGON((11.87157062501555 44.27663992525394, 11.93386214668788 44.2763158019823...
3	"POLYGON((11.7330331630454 44.48405330348036, 11.84236179092796 44.48296196729031,...



Thanks

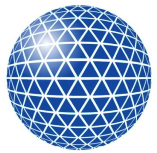
eswc.coypu.org

Simon Bin
sbin@informatik.uni-leipzig.de

Supported by

Federal Ministry
for Economic Affairs
and Climate Action

GA No. 01MK21007A



InfAI®
Institute for Applied Informatics

How does the data end up in the graph?

- Ontology (if required, e.g. with Protégé), upload it into Gitlab
  - Pipeline will document the ontology on https://schema.coypu.org/
- Describe data with metadata: Licence, Source (RDF document is stored in Gitlab)
  - Pipeline will document data on https://datasets.coypu.org/ (in tabular form) and https://metadata.coypu.org/ (as RDF)
- [A1] Find/have Data/API, and map it (e.g. with Python, RML, Tarql, …)
  - Put scripts/configurations into Gitlab, for documentation+reproducibility
- [A2] Upload data onto the server in the right folder (daily/weekly/static) and register it in dataset-mapping.csv
  - Pipeline will document file-sets on https://metadata.coypu.org/ (as RDF)

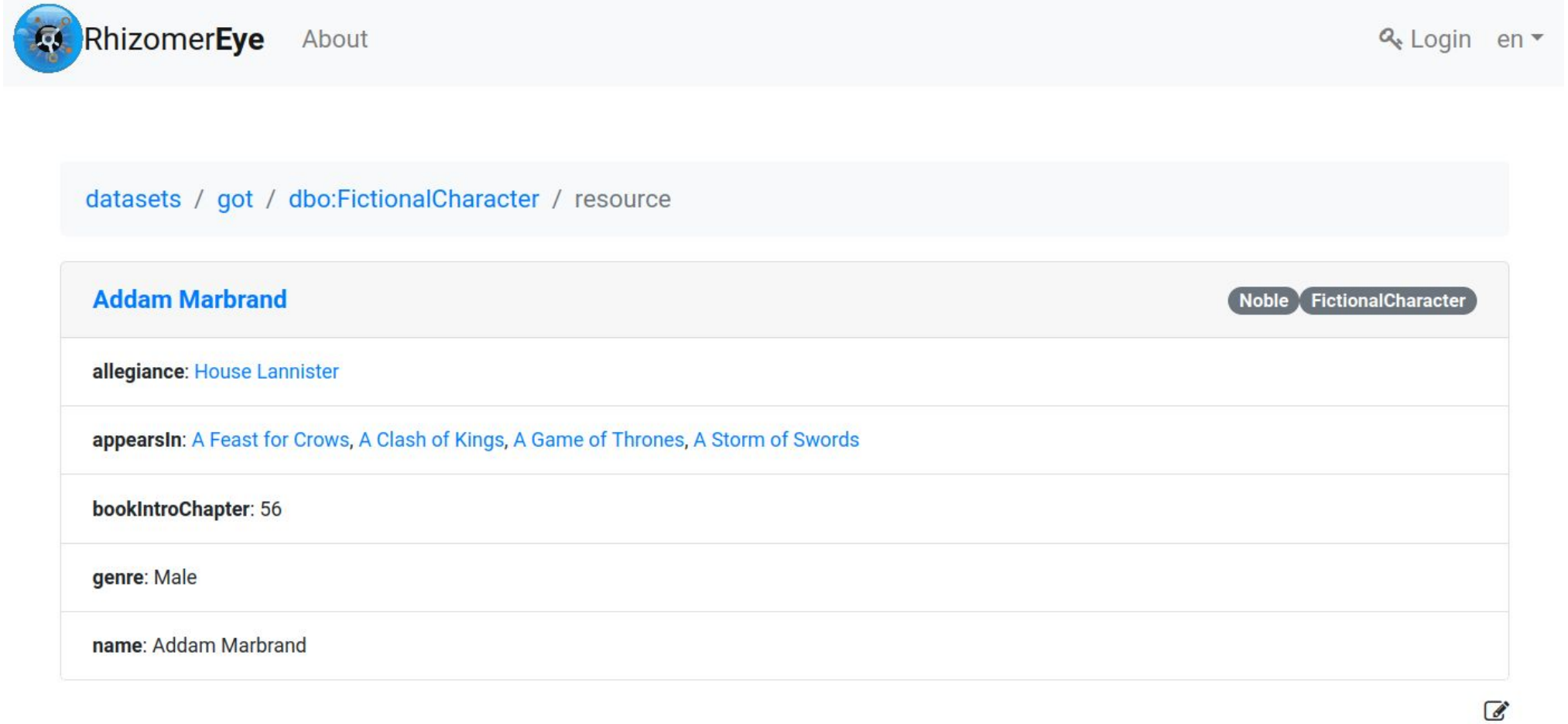
How does the data end up in the graph?

- [B] Create a periodic job comprising [A1] and [A2]
- (Currently) 1× per month the knowledge graph is created from all source data
- Daily updates are processed 1× per day
- Text-index for label search is built/refreshed
- Geo-index for GeoSPARQL queries is built/refreshed
- Statistics about the knowledge graph are calculated

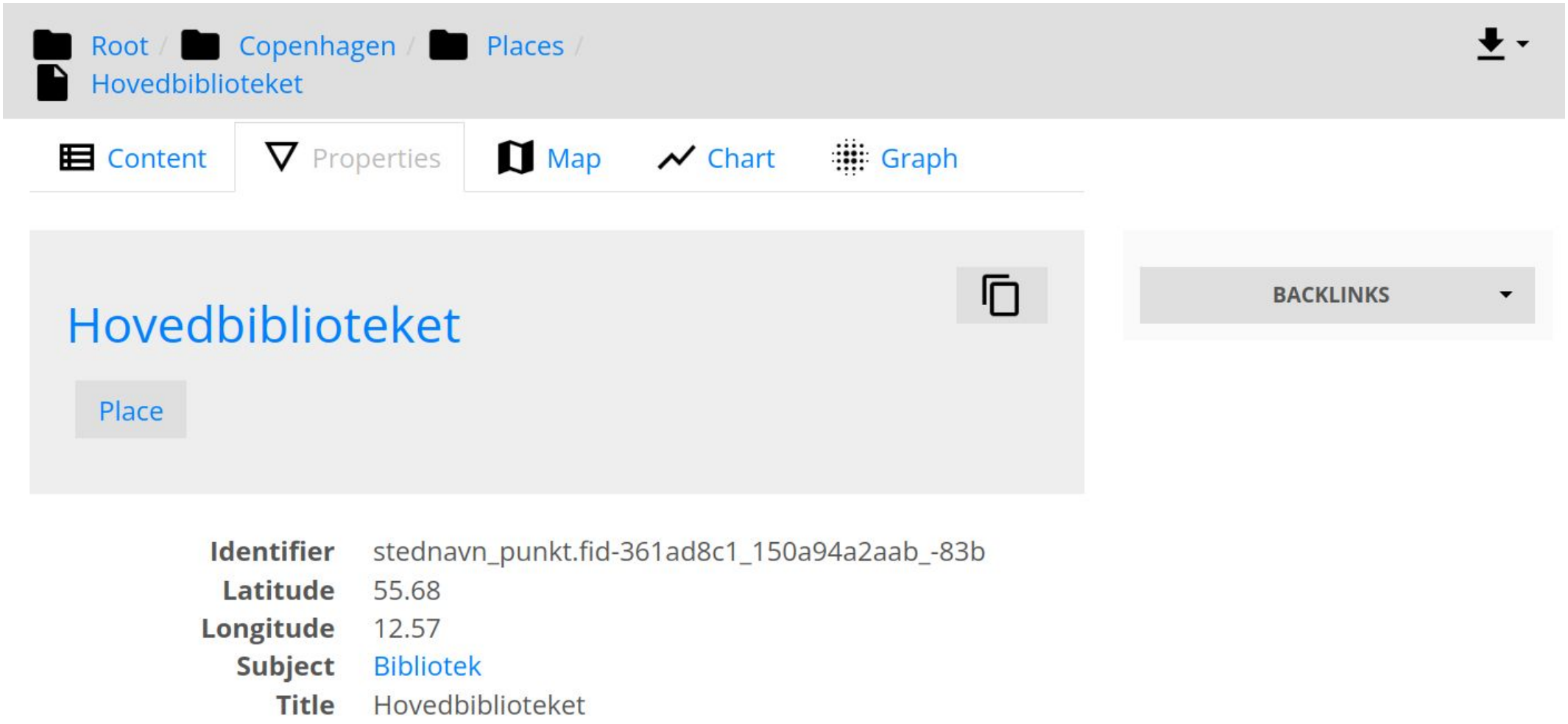
How to get data out of the graph

- Query with SPARQL
- Query with custom API (delegating to SPARQL internally)
- Query with prepared queries as CSV
- Interactive SPARQL results inspection using Triply Yasgui
- Redash-dashboards created on th basis of SPARQL queries
- Dereference and view RDF data using the CoyPu Linked Data Viewer at https://data.coypu.org
- Explore the knowledge graph using Ontodia Graph Explorer at https://explore.skynet.coypu.org/coypu-internal

RhizomerEye



Linked Data Hub



Thanks

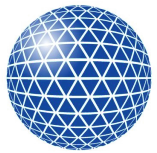
eswc.coypu.org

Simon Bin
sbin@informatik.uni-leipzig.de

Supported by

Federal Ministry
for Economic Affairs
and Climate Action

GA No. 01MK21007A



InfAI®
Institute for Applied Informatics