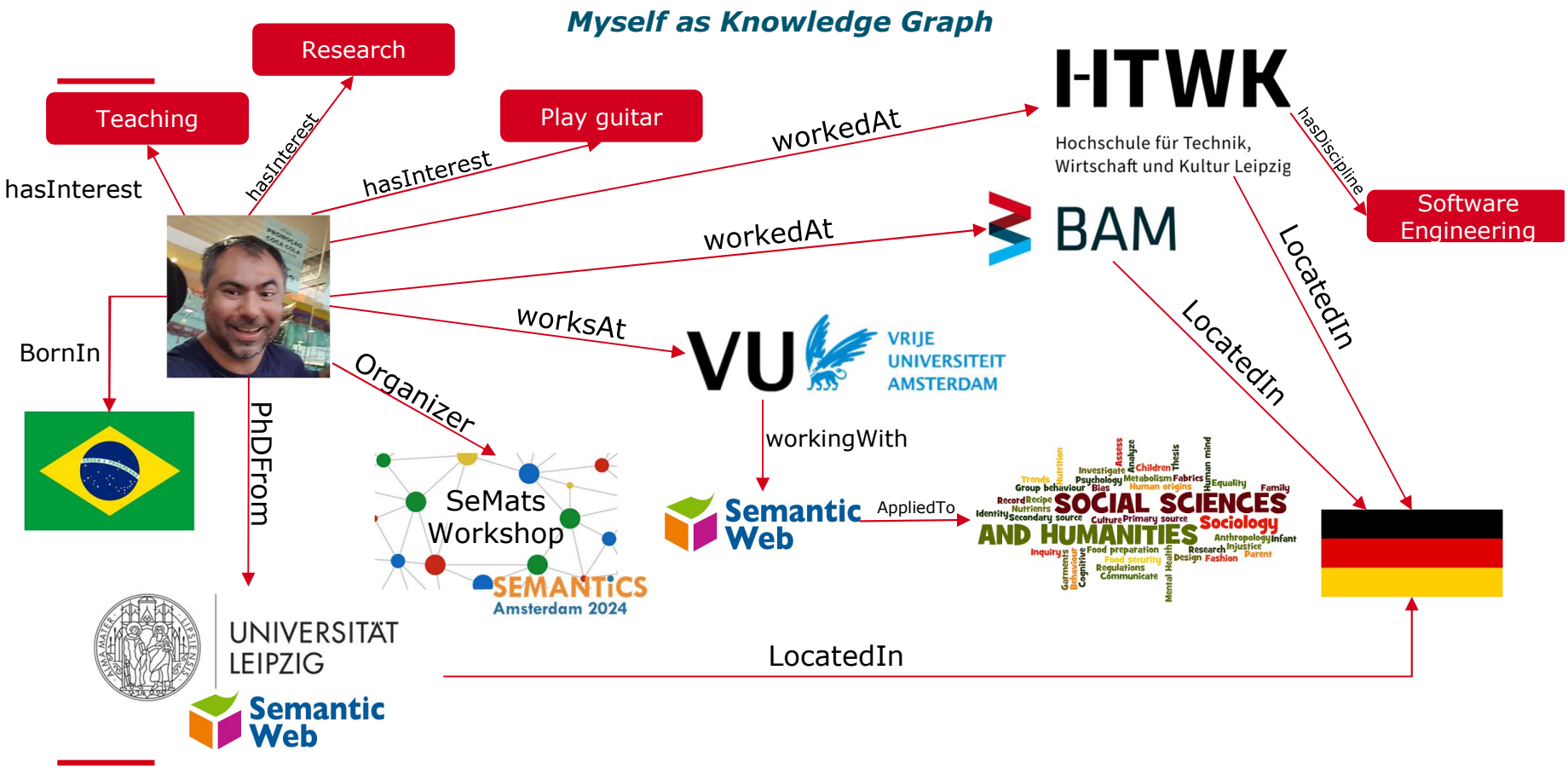
Knowledge Graphs applied to Social Sciences and Humanities on SSHOC-NL project

February 27th, 2025

The Andre Valdestilhas Knowledge Graph (KG)



The SSHOC-NL KG

- What is it?
- Where does the data come from?
- How is it organized?

SSHOC-NL Knowledge Graph: Proof of Concept

Overview

Objective: Develop an initial proof-of-concept knowledge graph (KG).

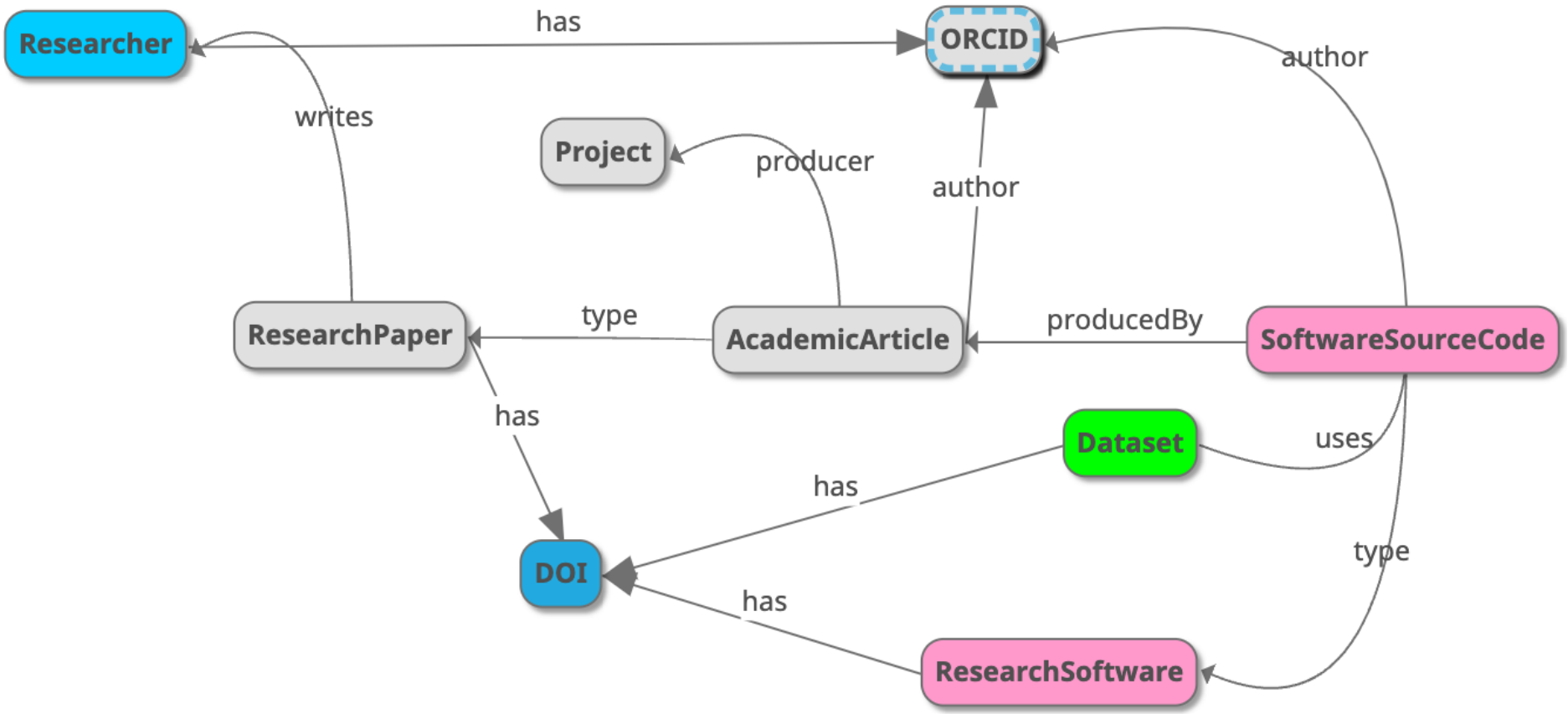
Key Aspects:

- Models the academic ecosystem in social science & humanities.
- Captures relationships between **researchers**, **datasets**, **research software**, and **research papers**.
- Reuses publicly available data.
- Uses persistent identifiers (DOIs for papers and datasets, ORCIDs for authors, etc.).

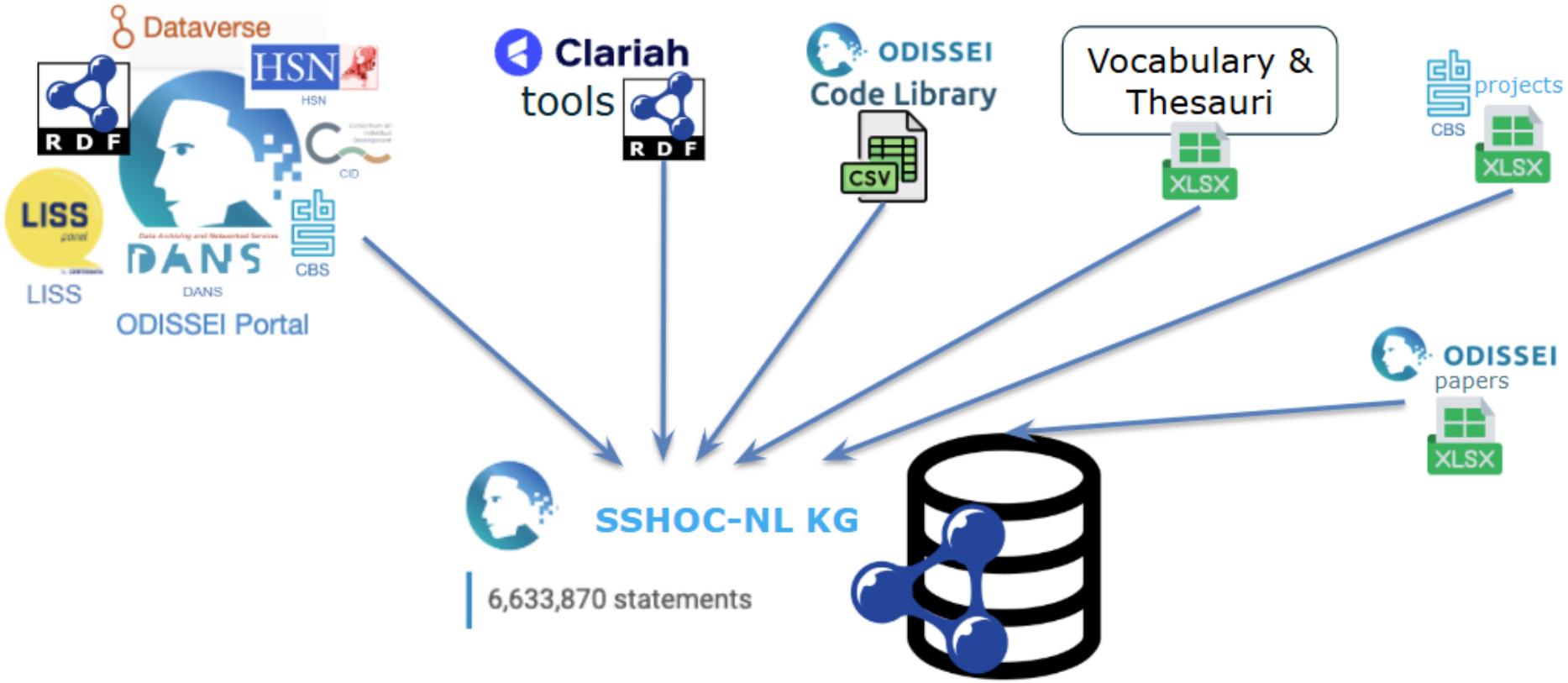
More Information: Read our data story about this knowledge graph.
`https://kg.odissei.nl/odissei/-/stories/ODISSEI-Knowledge-Graph-the-story`

Concepts and one instance

Researchers, datasets, research software, and research papers

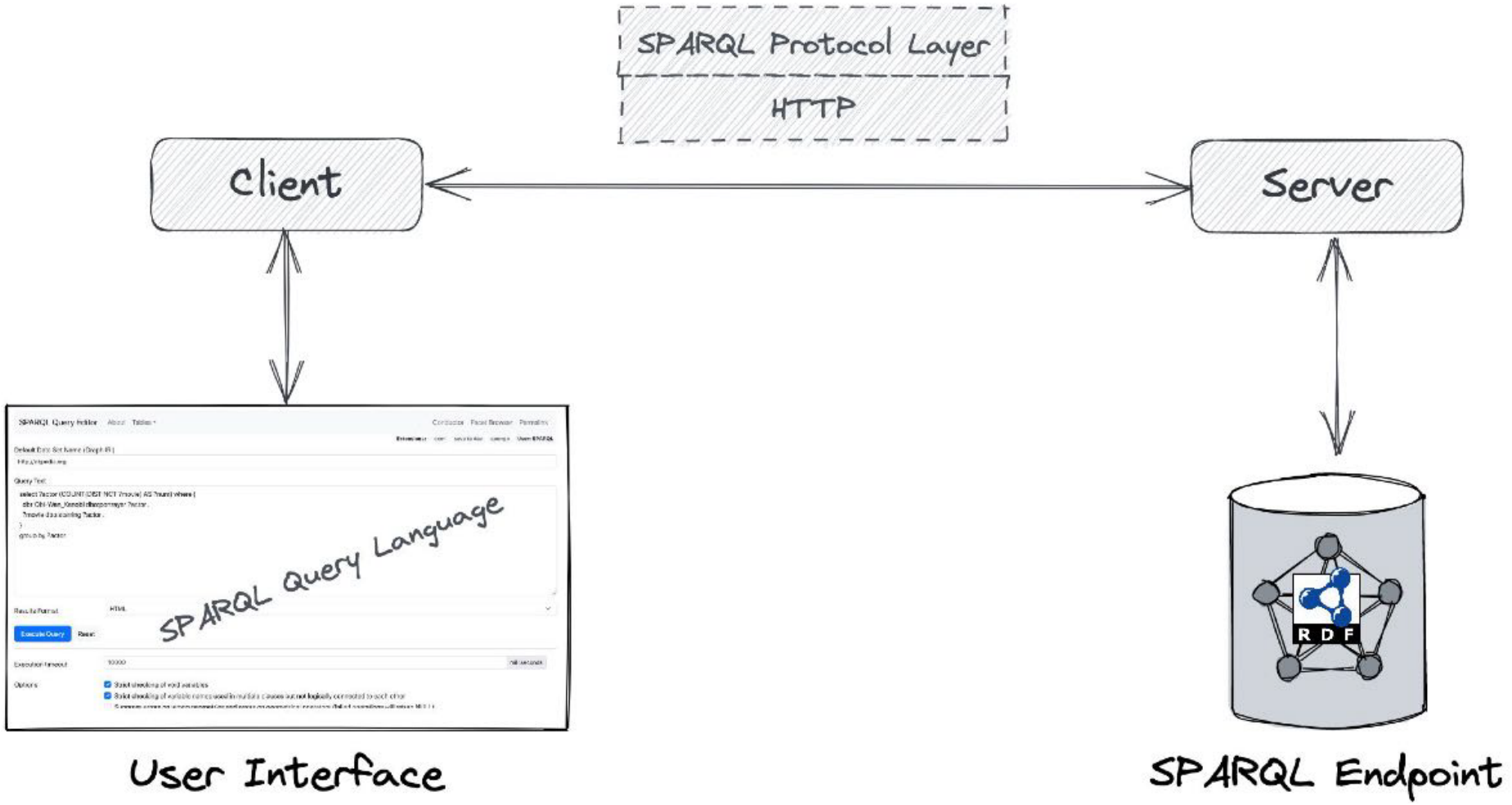


Where does the data come from?



SPARQL Endpoint

Sack, 2024



 https://zenodo.org/records/10062306

Simple SPARQL query

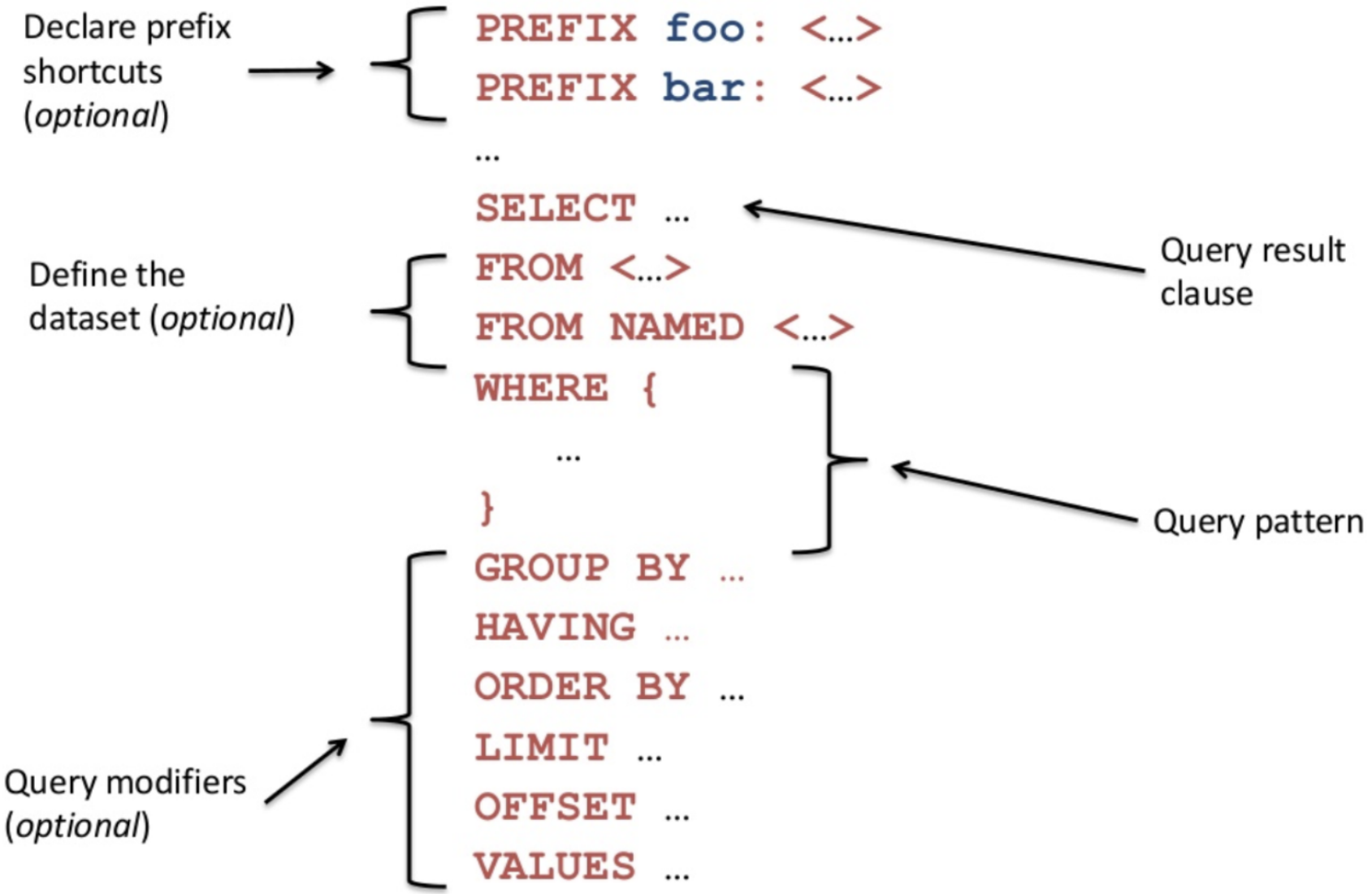
Starting to query the SSHOC-NL KG - Tools/Software from ODISSEI Code Library and CLARIAH tool repository [12]

```
1  prefix sdo: <https://schema.org/>
2  prefix dct: <http://purl.org/dc/terms/>
3
4  Select distinct ?softwareID ?author ?title WHERE
5  {
6    ?softwareID a sdo:SoftwareSourceCode.
7    ?softwareID dct:title ?title .
8    ?softwareID sdo:author ?author .
9  } order by ?author
```

Table | Response | Visualization | 38 results in 0.036 seconds | EXPORT

softwareID	author	title
filter	filter	filter
https://osf.io/ygs72/	https://orcid.org/0000-0001-6978-4737	Multiple environmental exposures along daily mobility paths and depressive symptoms: A smartphone-

Anatomy of a query



 https://www.mkbergman.com/2358/cwpk-25-querying-kbpedia-with-sparql/

Example: Simple SPARQL query

Starting to query the SSHOC-NL KG - Tools/Software from ODISSEI Code Library and CLARIAH tool repository [12]

```
1  prefix sdo: <https://schema.org/>
2  prefix dct: <http://purl.org/dc/terms/>
3
4  Select distinct ?softwareID ?author ?title WHERE
5  {
6    ?softwareID a sdo:SoftwareSourceCode.
7    ?softwareID dct:title ?title .
8    ?softwareID sdo:author ?author .
9  } order by ?author
```

Table | Response | Visualization | 38 results in 0.036 seconds | EXPORT

softwareID	author	title
filter	filter	filter
https://osf.io/ygs72/	https://orcid.org/0000-0001-6978-4737	Multiple environmental exposures along daily mobility paths and depressive symptoms: A smartphone-

Complex SPARQL query

Which projects use datasets with an ICD10-encoded main diagnosis? [9]

```
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix bibo: <http://purl.org/ontology/bibo/>
prefix odissei_kg_schema: <https://kg.odissei.nl/schema/>
prefix var: <https://portal.odissei.nl/schema/variableInformation#>
prefix dct: <http://purl.org/dc/terms/>

select distinct ?project ?dsScheme ?contextVarLabel ?shortTitle where {
 VALUES ?conceptVarLabel {"Diagnose gebaseerd op ICD10"@nl}
 ?conceptVar skos:prefLabel ?conceptVarLabel .
 ?contextVar skos:broader ?conceptVar .
 ?contextVar skos:altLabel ?contextVarLabel .
 ?var var:odisseiVariableVocabularyURI ?contextVar .
 ?dsScheme var:odisseiVariable ?var .
 ?dsScheme dct:alternative ?shortTitle .
 ?cbsdataset dct:alternative ?shortTitle .
 ?project dct:requires ?cbsdataset .
}
```

Table | Response | Charts | 1,172 results in 0.431 seconds | EXPORT | CONFIGURE

project	dsScheme	contextVarLabel	shortTitle
cbs_project:9740	doi:10.57934/0b01e410805d9385	LBZIcd10diagimp	LBZDIAGNOSENTAB
cbs_project:9740	doi:10.57934/0b01e41080395c06	LBZIcd10diagimp	LBZDIAGNOSENTAB

 https://kg.odissei.nl/odissei/-/queries/icd10-projects/5

Infrastructure

What do you need to ”do” knowledge graphs?

1. a triple store
2. a KG generation framework
3. other services that provide RDF import/export functionality
4. applications using KGs (mostly under the hood)

1. Triple store

- All triple stores provide the basic standardized functionality you need:
  - a ”database” to store RDF triples
  - query interface (SPARQL REST API over HTTP)
- Many (open source) triple stores to choose from
- For the SSHOC-NL KG demo, we use several triple stores in parallel
  - Fuseki (comes with SKOSMOS installation, hosted by DANS)
  - Virtuoso, Speedy (come with TriplyDB, hosted by Triply)
  - QLever (experimental phase)
  - also used GraphDB in the past, hosted on SURF Research Cloud
- limited vendor lock-in risk because of standardization, risks are mainly in the ”extra” functionality
  - publish: make URLs resolvable, graphs downloadable
  - user interface: RDF browsers, graph visualizations, full text search
  - For the SSHOC-NL KG demo, we also used Triply’s ”data stories”, ETL & CI/CD infrastructure

2. KG generation framework

How do you create triples for your KG?
Short answer: any method that produces valid RDF will do!

- Reuse useful RDF already published by others (including many thesauri, vocabularies, ...)
- Use RDF-export capabilities of tools you already use
- Use ”triplification” software (e.g. CLARIAH’s COW, RML) to convert CSV/Database tables to RDF
- Use RDFlib and other packages in Python to generate custom RDF from your own code
- For the SSHOC-NL KG, we used Triply’s extract, transform & load (ETL) infrastructure

3. RDF import/export services

Many web services we use may or may not be based on RDF. But many do provide RDF imports or exports:

- ODISSEI's Dataverse metadata portal, other data stations operated by DANS
- CLARIAH's tool registry
- Many vocabulary publishers (NDE, CESSDA, Getty, BARTOC, ...)
- ...

4. Applications

Many applications use knowledge graphs ”under the hood”

- Google’s ”knowledge panel” in search results
- Question answering systems (Watson, Siri)
- Annotation tasks (CLARIAH’s vocabulary recommender, Dataverse’s vocabulary-based keyword annotation)
- ...

Discussion

Questions

Contact: a.valdestilhas@vu.nl

Data Story

edu.nl/xrjkj

Simple Query

edu.nl/ajq69

Complex query

edu.nl/q6899

Slides

edu.nl/e7vx8