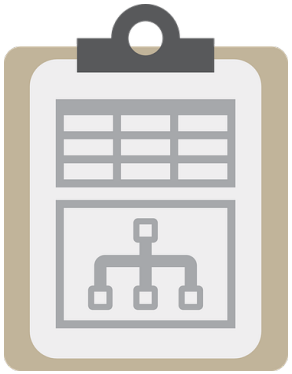
Linked Data and SPARQL

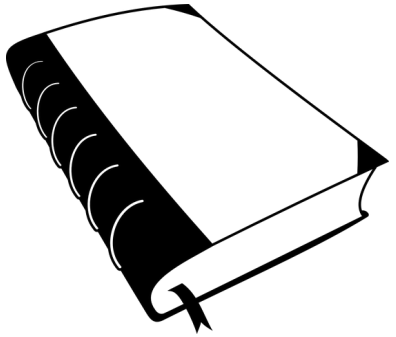
DATA =
RELATIONSHIPS

COW Workshop
IISG, February 23, 2018

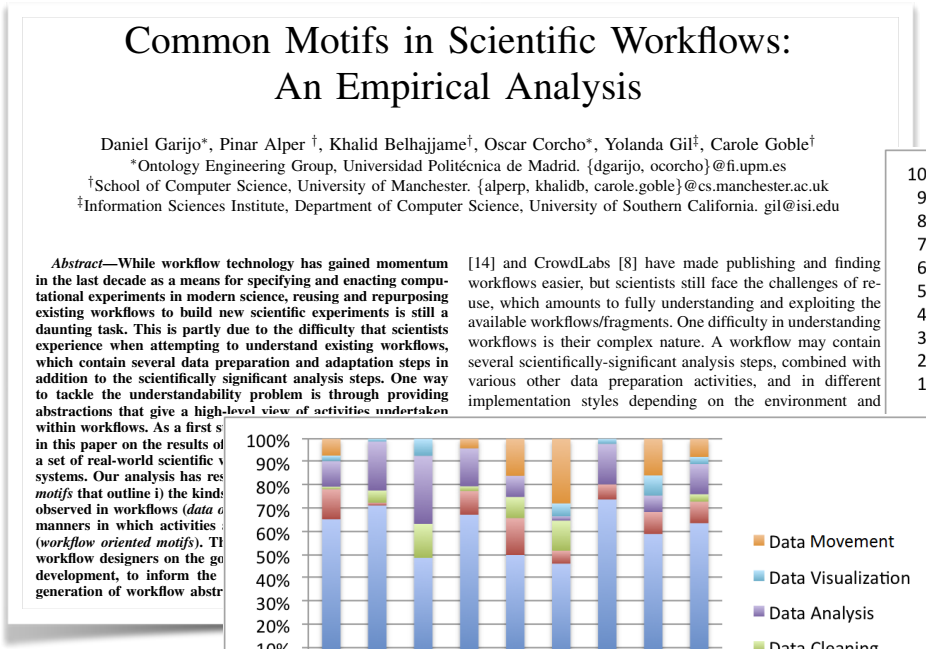
Richard Zijdeman, Albert Meroño Peñuala,
Auke Rijpma, Ruben Schalk, Rinke Hoekstra

Old school research





Data prep takes a lot of work!



Common Motifs in Scientific Workflows:
An Empirical Analysis

Daniel Garijo*, Pinar Alper †, Khalid Belhajjame†, Oscar Corcho*, Yolanda Gil‡, Carole Goble†
*Ontology Engineering Group, Universidad Politécnica de Madrid. {dgarijo, ocorcho}@fi.upm.es
†School of Computer Science, University of Manchester. {alperp, khalidb, carole.goble}@cs.manchester.ac.uk
‡Information Sciences Institute, Department of Computer Science, University of Southern California. gil@isi.edu

***Abstract*—While workflow technology has gained momentum in the last decade as a means for specifying and enacting computational experiments in modern science, reusing and repurposing existing workflows to build new scientific experiments is still a daunting task. This is partly due to the difficulty that scientists experience when attempting to understand existing workflows, which contain several data preparation and adaptation steps in addition to the scientifically significant analysis steps. One way to tackle the understandability problem is through providing abstractions that give a high-level view of activities undertaken within workflows. As a first s**
in this paper on the results of
a set of real-world scientific
systems. Our analysis has res
***motifs* that outline i) the kinds**
observed in workflows (*data*
manners in which activities
(*workflow oriented motifs*). Th
workflow designers on the go
development, to inform the
generation of workflow abstr

[14] and CrowdLabs [8] have made publishing and finding workflows easier, but scientists still face the challenges of reuse, which amounts to fully understanding and exploiting the available workflows/fragments. One difficulty in understanding workflows is their complex nature. A workflow may contain several scientifically-significant analysis steps, combined with various other data preparation activities, and in different implementation styles depending on the environment and

Fig. 3. Distribution of Data-Oriented Motifs per domain

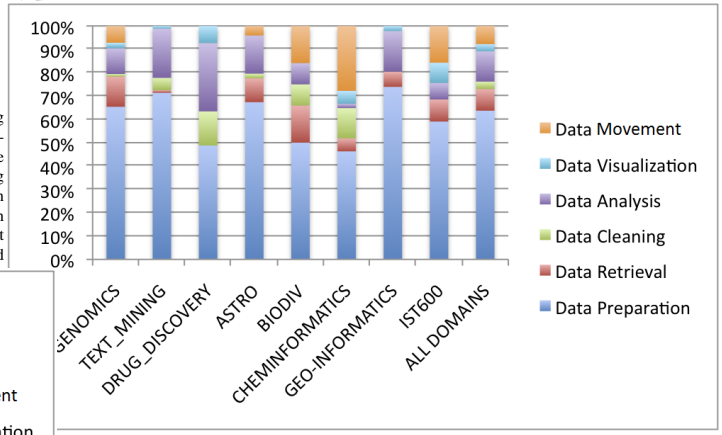
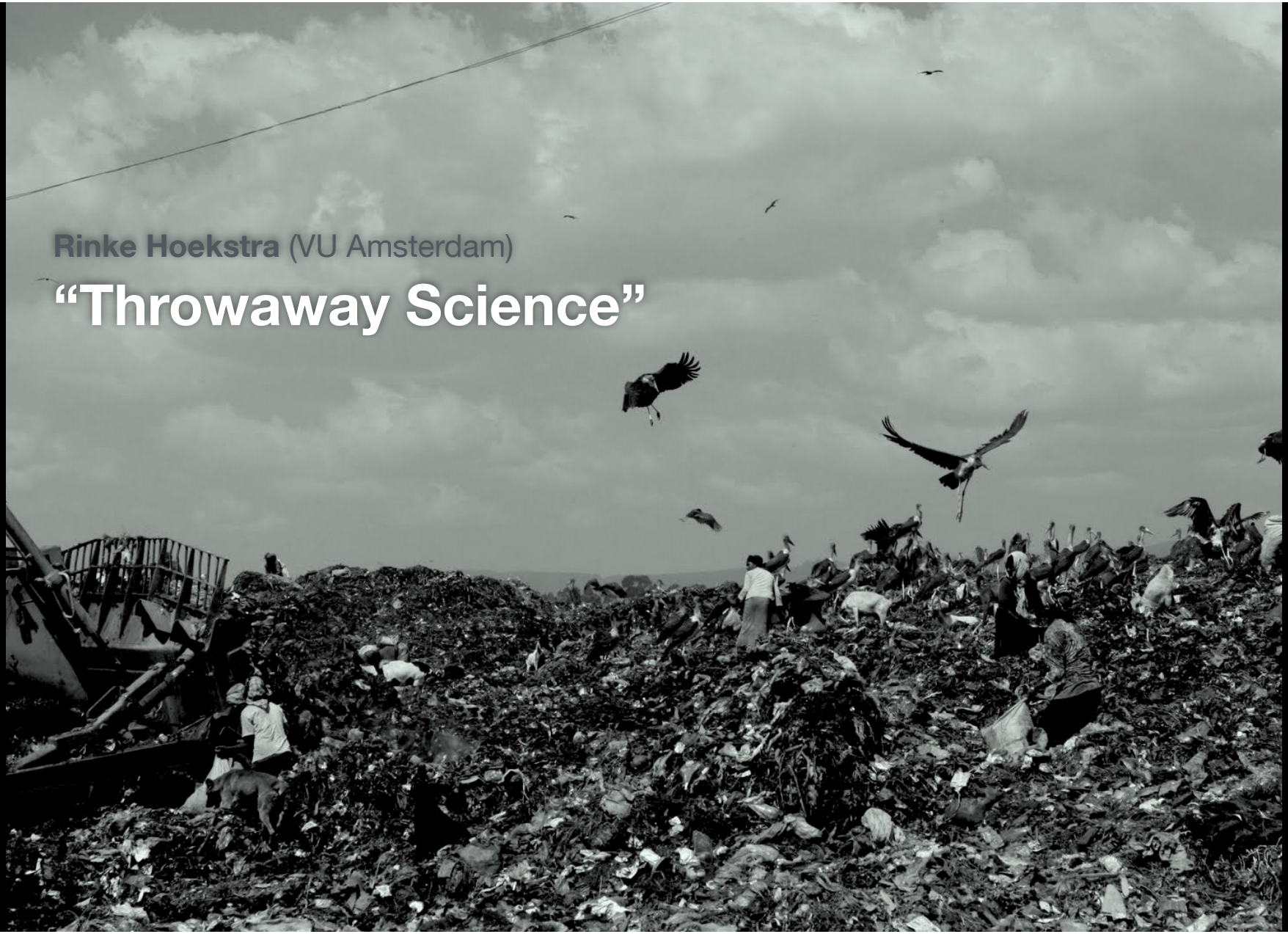


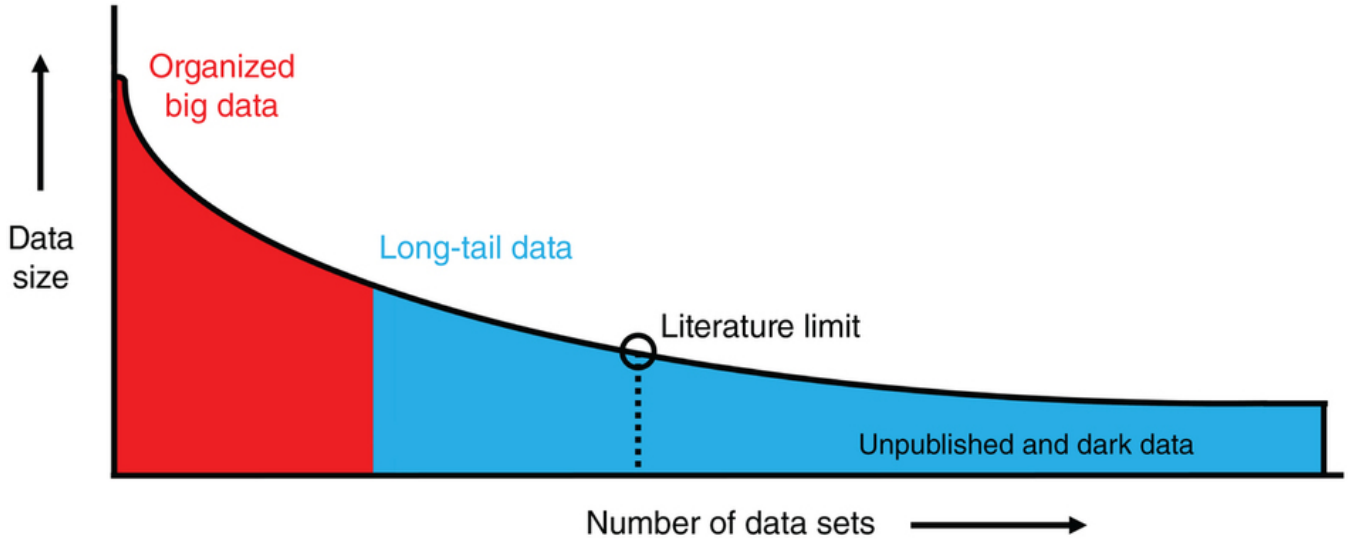
Fig. 3. Distribution of Data-Oriented Motifs per domain

We do this **repeatedly**

for **the same** datasets!

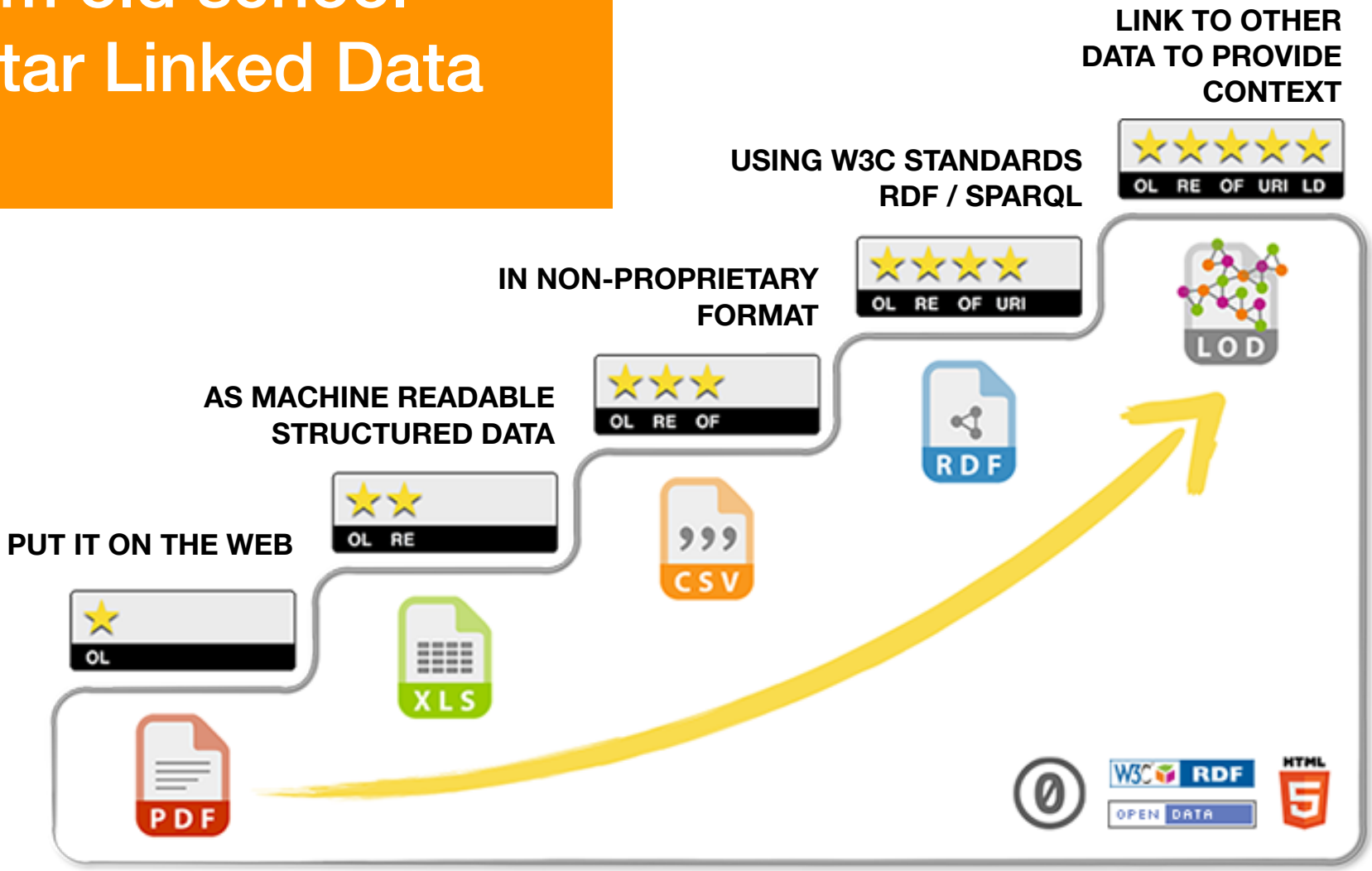

Rinke Hoekstra (VU Amsterdam)
“Throwaway Science”

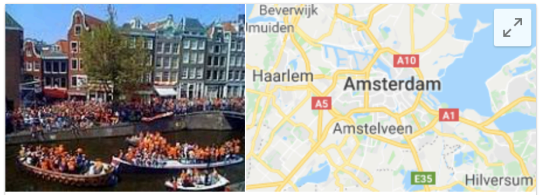
“long tail of big data”-problem




https://www.ted.com/talks/tim_berners_lee_on_the_next_web
The Next Web

From old school to 5-star Linked Data





Amsterdam

Capital of the Netherlands

Amsterdam is the Netherlands' capital, known for its artistic heritage, elaborate canal system and narrow houses with gabled facades, legacies of the city's 17th-century Golden Age. Its Museum District houses the Van Gogh Museum, works by Rembrandt and Vermeer at the Rijksmuseum, and modern art at the Stedelijk. Cycling is key to the city's character, and there are numerous bike paths.

Area: 219.3 km²

Weather: 1°C, Wind NE at 18 km/h, 80% Humidity

Local time: Thursday 21:29

Population: 821,752 (2015) UNdata

Plan a trip

Amsterdam travel guide

3-star hotel averaging €111, 5-star averaging €283

Upcoming Events

Colleges and Universities: University of Amsterdam, MORE

People also search for

View 15+ more

Did you ever use Linked Data?

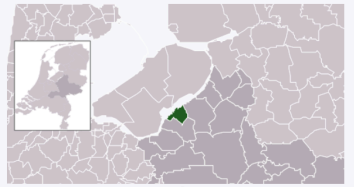
Harderwijk

Municipality

Harbour of Harderwijk seen from windmill De Hoop

Flag

Coat of arms



Location in Gelderland

Coordinates: 52°21′N 5°37′E

Country	Netherlands
Province	Gelderland
Government[1]	
• **Body**	Municipal council
• **Mayor**	Harm-Jan van Schaik (CDA)
Area[2]	
• **Total**	48.27 km² (18.64 sq mi)

RDF - Triples

- All information in RDF is expressed as **triples**; two-placed predicates
- A triple consists of a **subject**, a **predicate** and an **object**:

subject	predicate	object
The Netherlands	has capital	Amsterdam
Amsterdam	has mayor	Jozias van Aartsen
Eberhard van der Laan	birth year	1947

- Another word for a triple is a **statement** or a **fact**
- The elements of an RDF triple are either **URI references**, **blank nodes**, or, **literals**.

RDF - Uniform Resource Identifiers (URIs)

- The Resource Description Framework talks about **resources**
  (almost anything is a resource)

- Resources are **identified by** URIs, or URIs **denote** resources
  (URIs can only refer to a resource, they **are not** the resource, and **multiple** URIs can denote the **same** resource)

The Netherlands	http://dbpedia.org/resource/The_Netherlands
has capital	http://dbpedia.org/ontology/capital
Amsterdam	http://dbpedia.org/resource/Amsterdam
has mayor	http://dbpedia.org/ontology/leaderName
Eberhard van der Laan	http://dbpedia.org/resource/Eberhard_van_der_Laan

- Internationalised Resource Identifiers are URIs that allow unicode characters

VU UNIVERSITY AMSTERDAM

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

Magritte

RDF - URI ≠ URL

URLs are not the only **URIs**

ISBN
urn:isbn:0-486-27557-4

Geo
geo:37.786971,-122.399677

Mail
mailto:rinke.hoekstra@vu.nl

And many more…

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - URIs and CURIES (or QNames)

- URIs are often long and hard to read and write
- Most **serialisations** use an abbreviation mechanism: **namespaces** and **prefixes**

```
@prefix dbpedia:    <http://dbpedia.org/resource/>
@prefix dbo:        <http://dbpedia.org/ontology/>
```

- We can then map **compact URI**s (CURIs) to **full URIs**

dbpedia:The_Netherlands	http://dbpedia.org/resource/The_Netherlands
dbo:capital	http://dbpedia.org/ontology/capital

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - URIs and Triples

- We can now state that the capital of The Netherlands is Amsterdam

<http://dbpedia.org/resource/The_Netherlands> <http://dbpedia.org/ontology/capital> <http://dbpedia.org/resource/Amsterdam> .

- Or use prefixes

dbpedia:The_Netherlands dbo:capital dbpedia:Amsterdam .

- But what if we want to state that Amsterdam's total area is 219320000?

- Cannot have one URI for every integer, decimal number, string, etc.

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - Literals

- **Literals** are used to represent "literal" data values
- All literals have a **datatype**
- Datatypes are also **resources**, referenced via URIs, and written as:

```
dbpedia:Amsterdam dbo:areaTotal "219320000"^^xsd:double.
```

- Default: if no datatype is specified, the datatype is assumed to be **xsd:string**

```
dbpedia:Amsterdam dbo:officialName "Amsterdam".
```

- One can specify the **language** of a string using a **language tag**:

```
dbpedia:The_Hague rdfs:label "Den Haag"@nl.
dbpedia:The_Hague rdfs:label "The Hague"@en.
```

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - Graphs

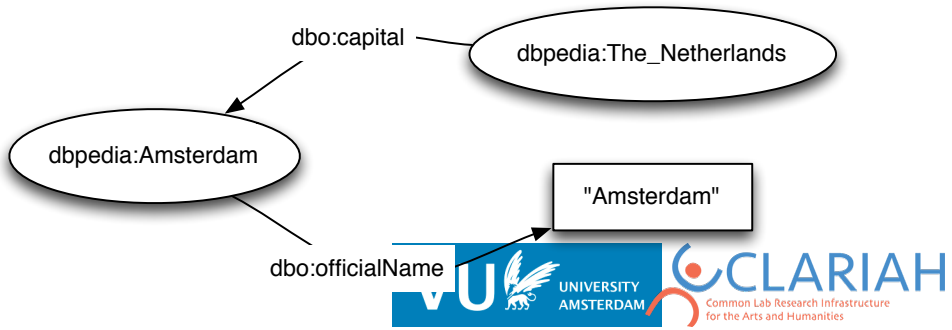
- An **RDF graph** is a **set of triples**, e.g.:

```
dbpedia:The_Netherlands dbo:capital       dbpedia:Amsterdam .
dbpedia:Amsterdam       dbo:officialName  "Amsterdam".
```

  … is a graph that contains two triples

- In practice, many RDF Graphs **have URIs** themselves

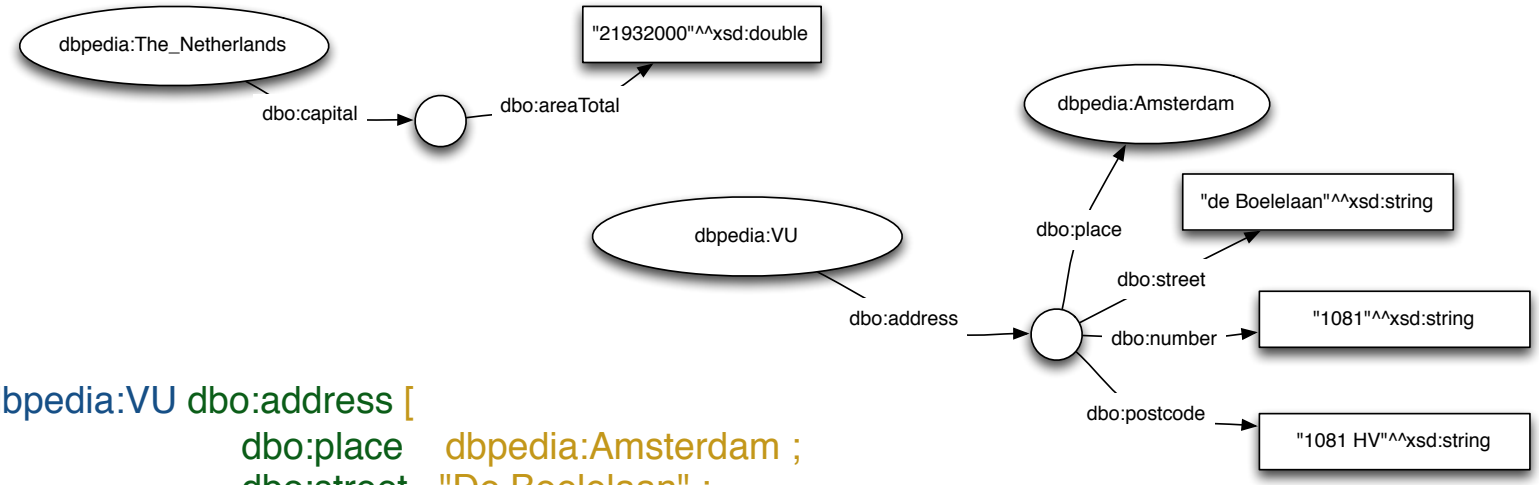
- RDF graphs are often represented as a **directed labelled graph**:



RDF - Blank Nodes

- Blank nodes are resources **without a URI**

- Use them when resource is **unknown**, or has **no (natural) identifier**




```
dbpedia:VU dbo:address [
                    dbo:place    dbpedia:Amsterdam ;
                    dbo:street   "De Boelelaan" ;
                    dbo:number   "1081" ;
                    dbo:postcode "1081 HV"
                   ] .
```

VU UNIVERSITY AMSTERDAM

CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - Triple Grammar

- **Literals** and **BNodes** may not appear in every position of a triple

	subject	predicate	object
URI References	✓	✓	✓
Literals	✗	✗	✓
Blank Nodes	✓	✗	✓

- … literals are **just values**, no relationships from literals allowed

- … blank nodes in predicate position "**too meaningless**" and confusing

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

SPARQL - Query Syntax

PREFIX: the namespace prefixes used in the SPARQL query

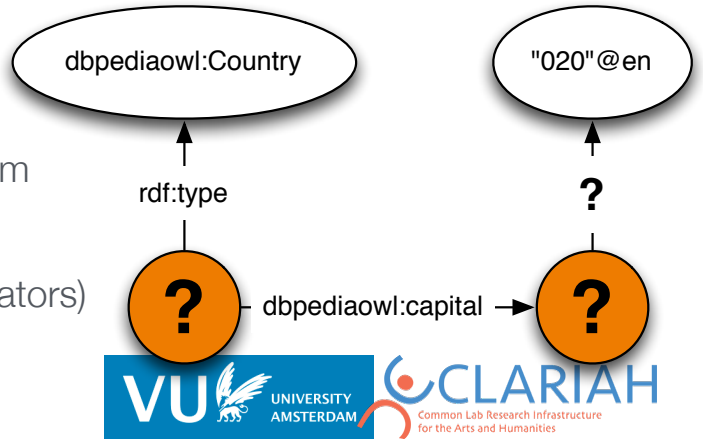
```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?city WHERE {
        ?city dbo:areaCode        "020" .
} LIMIT 10
```

SELECT: the entities (variables) you want to return

WHERE: the (sub)graph you want to get information from

... including additional constraints on results (using operators)

http://www.w3.org/TR/sparql11-query

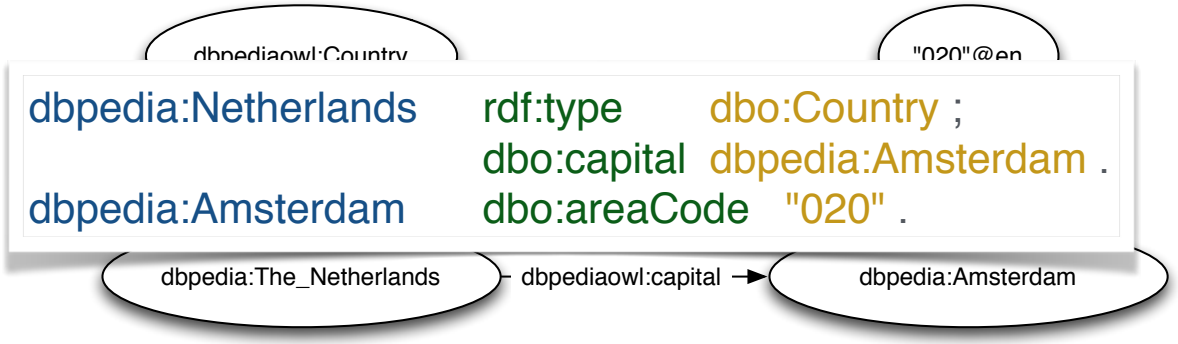


SPARQL - Triple Patterns

- A **graph pattern** consists of multiple **triple patterns**

- A **triple pattern** is a triple with **zero** or **more** variables

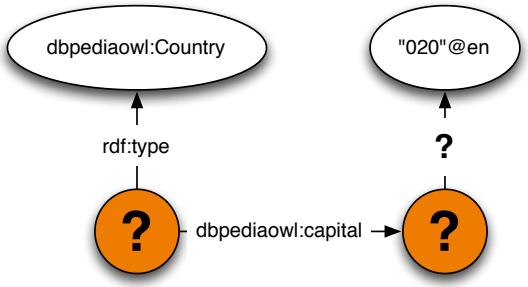
```
?x  dbo:capital  dbpedia:Amsterdam .
?x  dbo:capital  ?y .
?x  dbo:areaCode  "020" .
?x  ?p  ?y
```

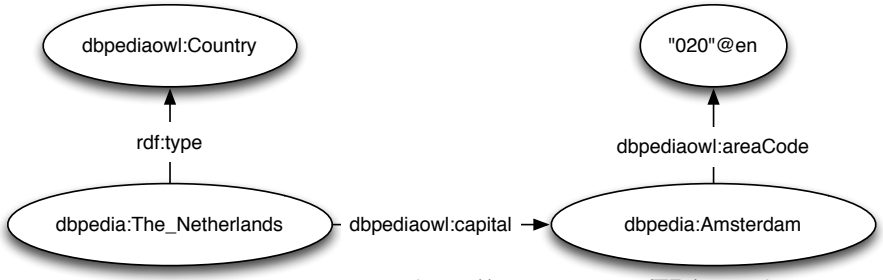


SPARQL - Triple patterns form a **conjunction**

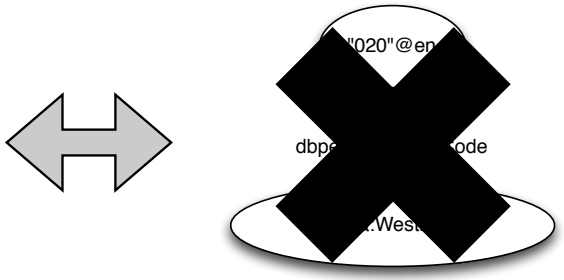
Every triple pattern in the graph pattern should **match**

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?x WHERE {
      ?x        dbo:capital        ?y .
      ?y        dbo:areaCode       "020" .
} LIMIT 10
```



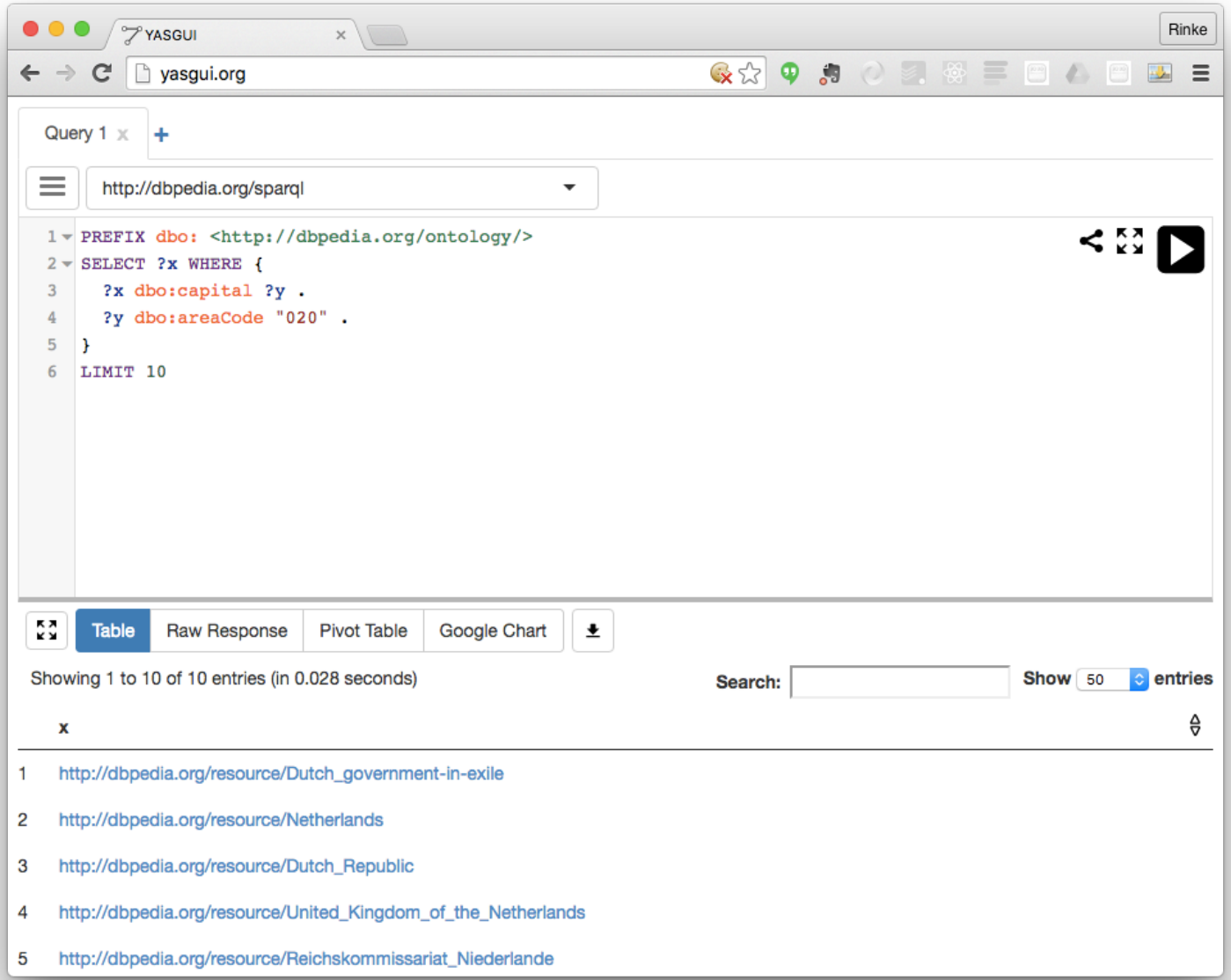


http://www.w3.org/TR/sparql11-query



VU UNIVERSITY AMSTERDAM

CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

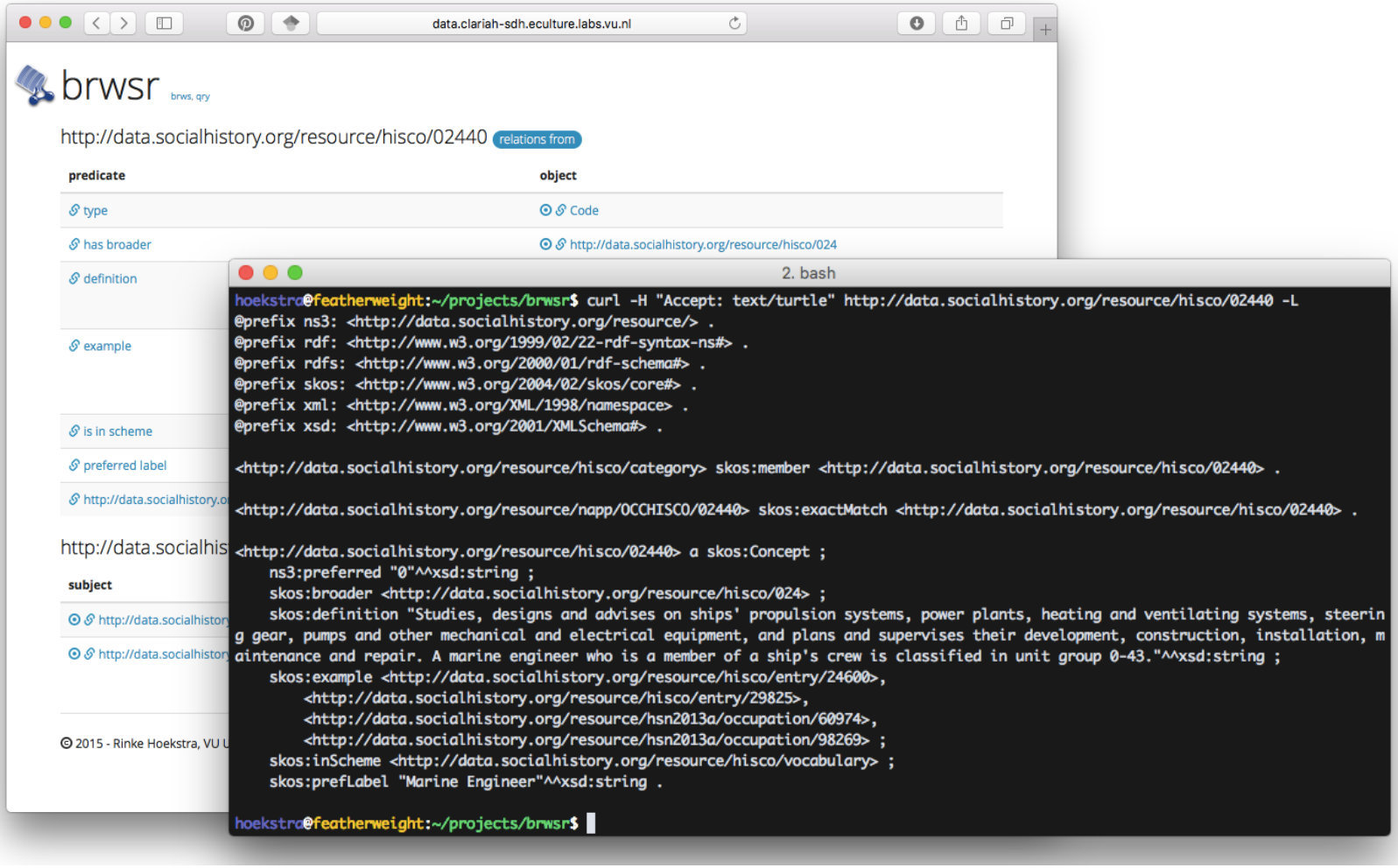

YASGUI
Rinke
yasgui.org
Query 1
http://dbpedia.org/sparql
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?x WHERE {
?x dbo:capital ?y .
?y dbo:areaCode "020" .
}
LIMIT 10
Table
Raw Response
Pivot Table
Google Chart
Showing 1 to 10 of 10 entries (in 0.028 seconds)
Search:
Show 50 entries
x
1 http://dbpedia.org/resource/Dutch_government-in-exile
2 http://dbpedia.org/resource/Netherlands
3 http://dbpedia.org/resource/Dutch_Republic
4 http://dbpedia.org/resource/United_Kingdom_of_the_Netherlands
5 http://dbpedia.org/resource/Reichskommissariat_Niederlande

RDF - Why HTTP URIs?

- HTTP URIs have a **global scope**, unique throughout the Web
  (c.f. e.g. keys in relational databases which are only unique within a table)
  - Helps to avoid **name clashes** http://abc-co.com/category/item/123 http://xyz-co.com/product/123
  - Grounded in **society** … have to pay for a registered domain name
- HTTP URIs are also **addresses**
  - Exploit the well-functioning machinery of Web **browsing**
  - Track data by **following** the resource identifiers found in triples

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

RDF - HTTP Content Negotiation (Cool URIs)



Flexibility - Why triples?

- Very expressive: all other data structures can be **transformed** to graphs

Tables:	row	column	cell
Trees:	parent	path	child

- Minimal **ontological commitment**, the data model has **no fixed keywords**

- This results in **maximal flexibility** with regard to data merging.

In RDF, types and relationships are **part of the data**
(unlike **database columns**, **binary predicates**, and **XML elements**)

VU UNIVERSITY AMSTERDAM CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

Example: from **Tables** to **Graphs** …

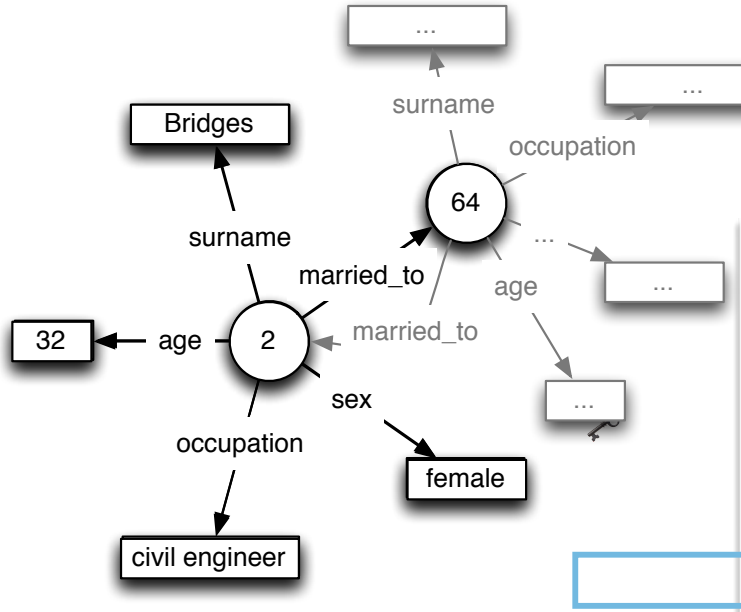
- Most **tables** actually express **graphs**
- Each **row** is a **record** *about* something
- Each **column** is an **attribute** of that thing
- A **primary key** identifies a record (local)
- A **secondary key** can identify an (external) record

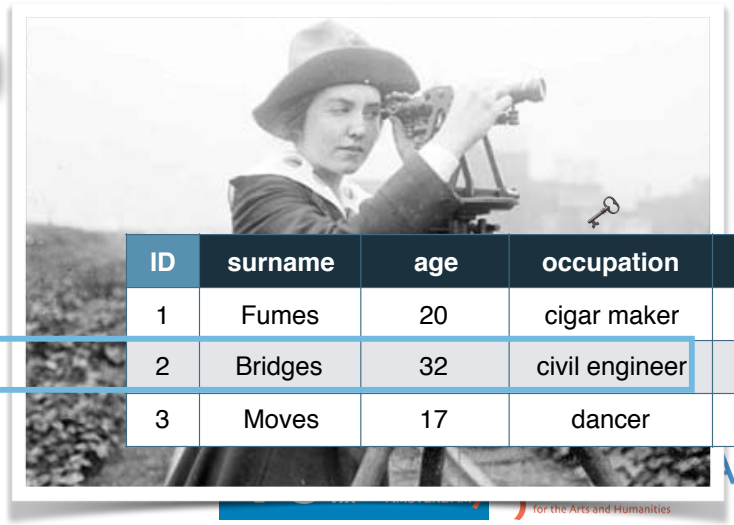
ID	surname	a
1	Fumes	
2	Bridges	
3	Moves	

ID	surname	age	occupation	sex	married_to
1	Fumes	20	cigar maker	female	25
2	Bridges	32	civil engineer	female	64
3	Moves	17	dancer	male	325

Example: from **Tables** to **Graphs**… (the simple version)

- Each **node** represents a **thing** (*resource*) or **value** (*literal*)
- Each **edge** represents an **attribute** (*property*)

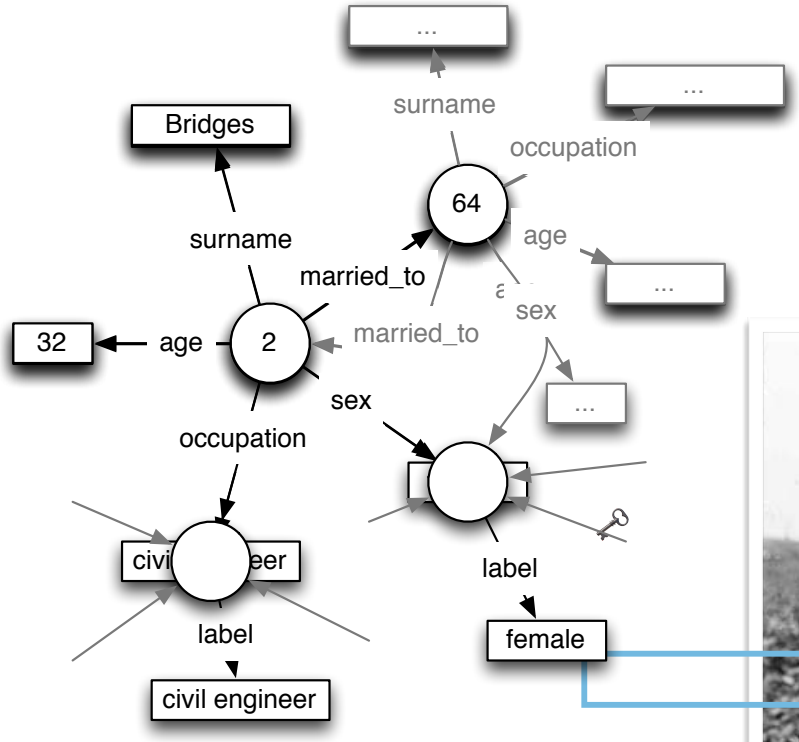


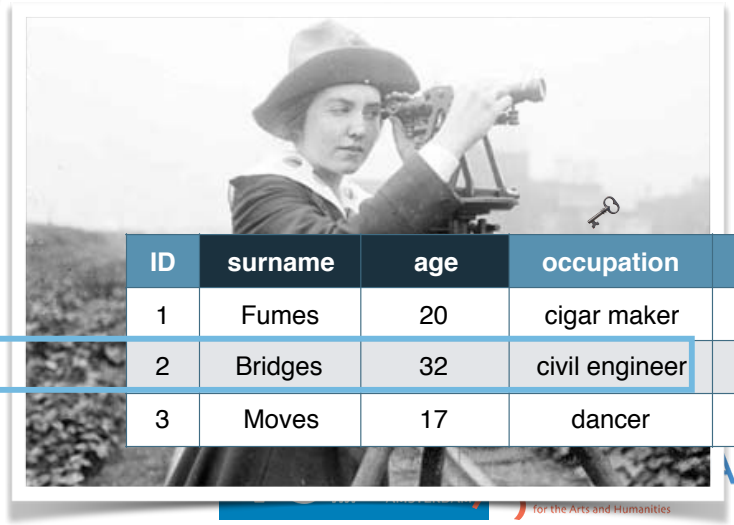


ID	surname	age	occupation	sex	married_to
1	Fumes	20	cigar maker	female	25
2	Bridges	32	civil engineer	female	64
3	Moves	17	dancer	male	325

Example: from **Tables** to **Graphs**… (the “better” version)

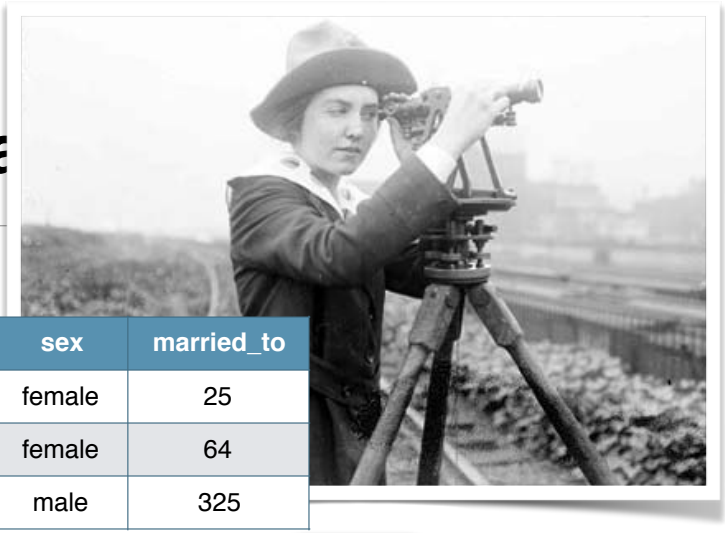
- Some **values** are actually **keys** (literals vs URIs)





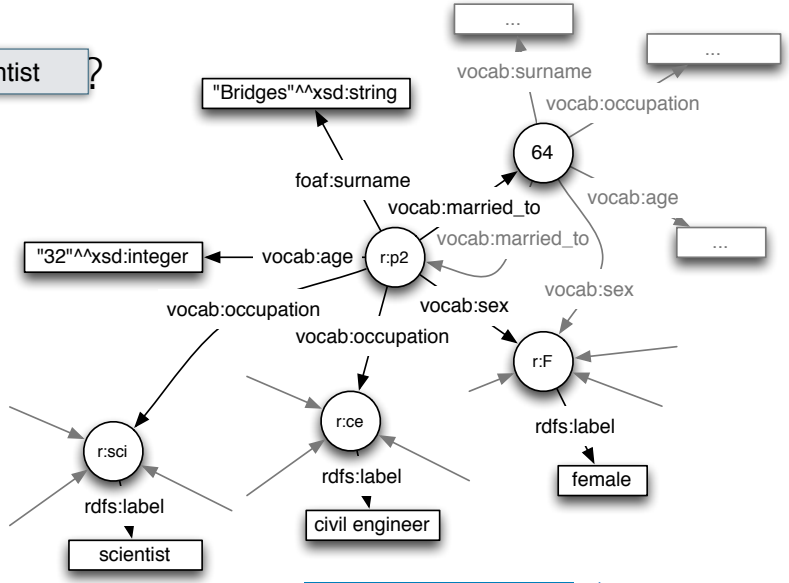
ID	surname	age	occupation	sex	married_to
1	Fumes	20	cigar maker	female	25
2	Bridges	32	civil engineer	female	64
3	Moves	17	dancer	male	325

Example: from **Tables** to **Gra**

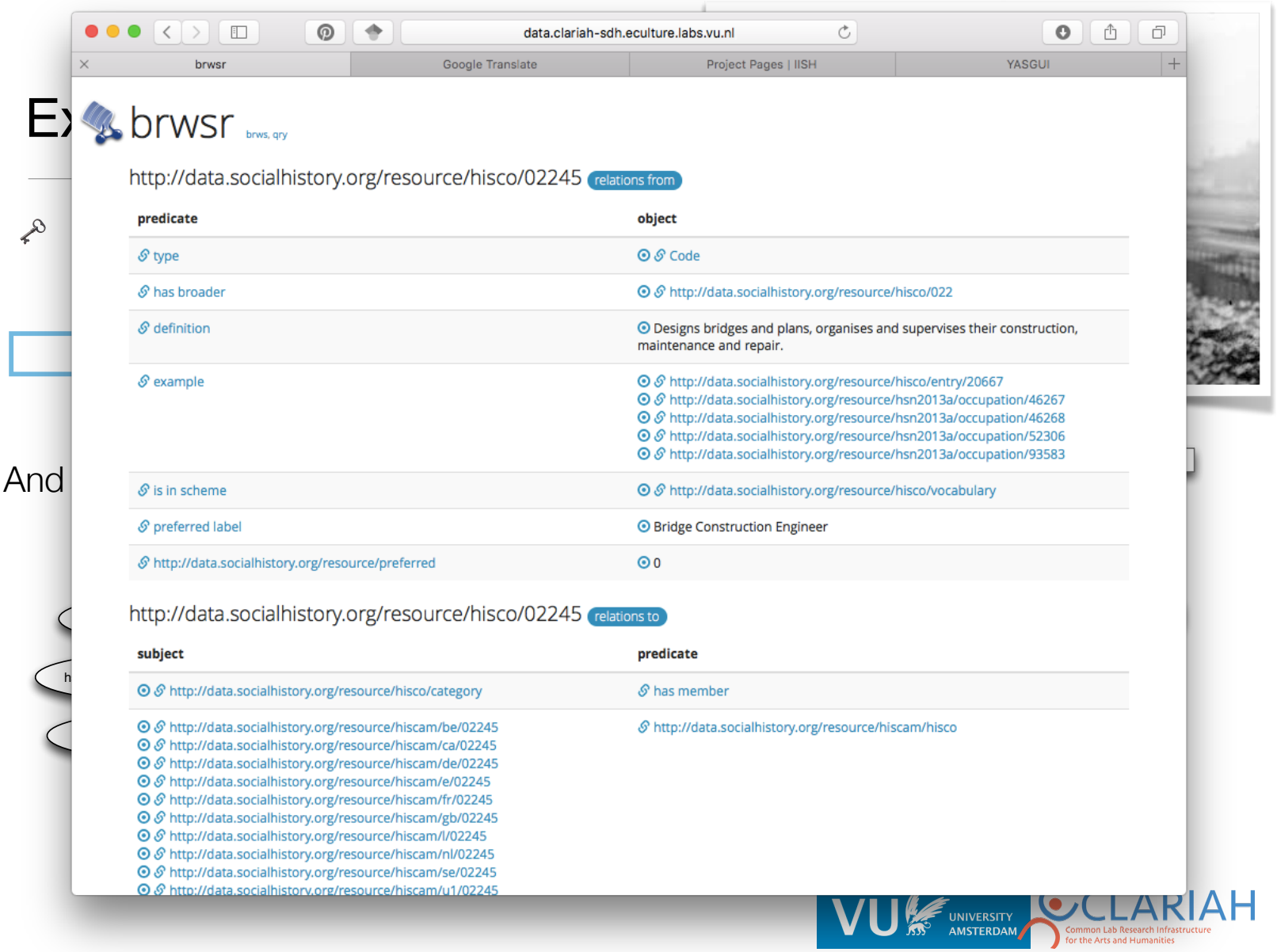


ID	surname	age	occupation	sex	married_to
1	Fumes	20	cigar maker	female	25
2	Bridges	32	civil engineer	female	64
3	Moves	17	dancer	male	325

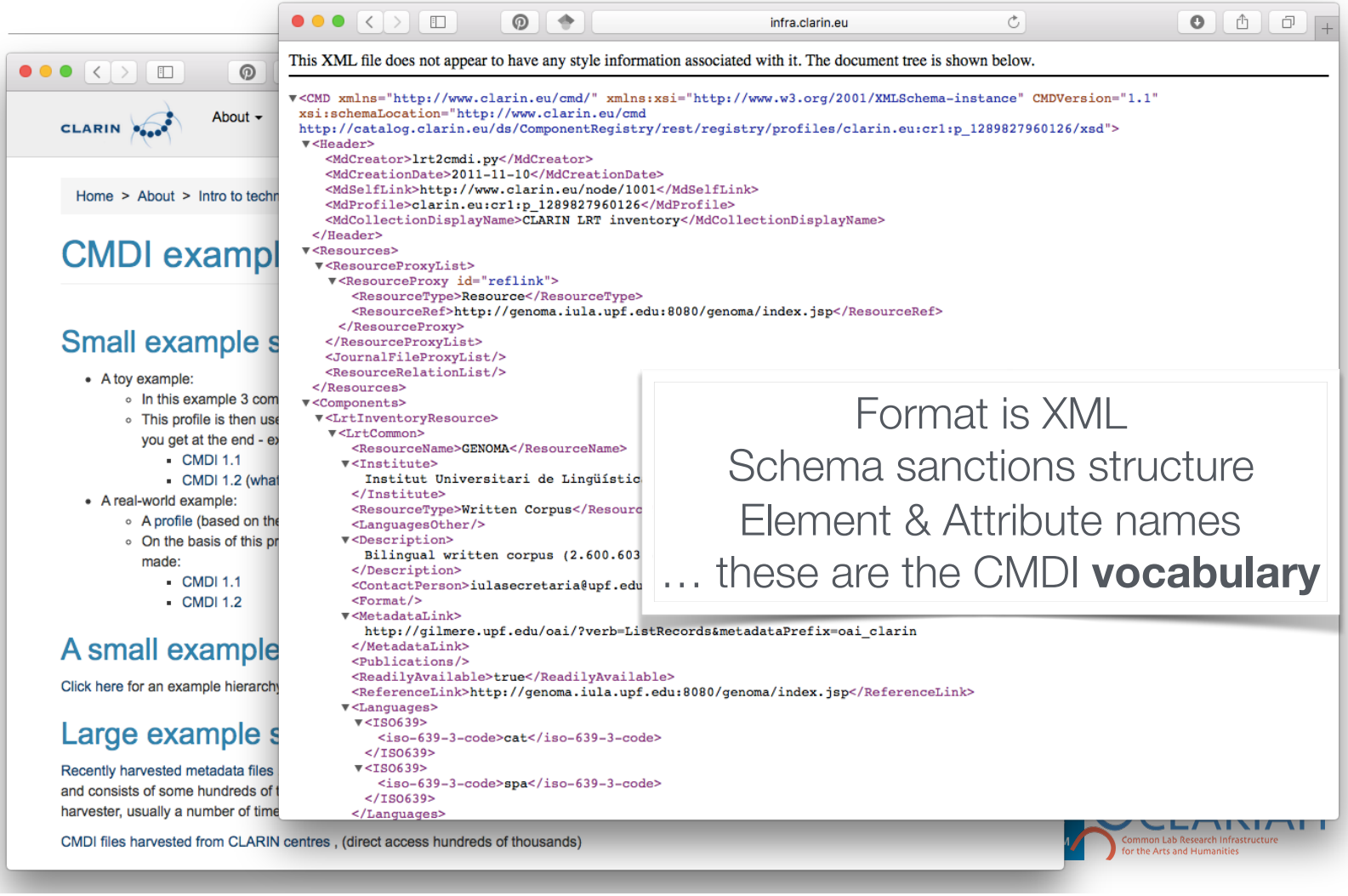
But what if Bridges is also a scientist?
… or add multiple languages?



VU UNIVERSITY AMSTERDAM
CLARIAH Common Lab Research Infrastructure for the Arts and Humanities


data.clariah-sdh.eculture.labs.vu.nl
brwsr
Google Translate
Project Pages | IISH
YASGUI
brwsr
brws, qry
http://data.socialhistory.org/resource/hisco/02245 relations from
predicate
object
type
Code
has broader
http://data.socialhistory.org/resource/hisco/022
definition
Designs bridges and plans, organises and supervises their construction, maintenance and repair.
example
http://data.socialhistory.org/resource/hisco/entry/20667
http://data.socialhistory.org/resource/hsn2013a/occupation/46267
http://data.socialhistory.org/resource/hsn2013a/occupation/46268
http://data.socialhistory.org/resource/hsn2013a/occupation/52306
http://data.socialhistory.org/resource/hsn2013a/occupation/93583
is in scheme
http://data.socialhistory.org/resource/hisco/vocabulary
preferred label
Bridge Construction Engineer
http://data.socialhistory.org/resource/preferred
0
http://data.socialhistory.org/resource/hisco/02245 relations to
subject
predicate
http://data.socialhistory.org/resource/hisco/category
has member
http://data.socialhistory.org/resource/hiscam/be/02245
http://data.socialhistory.org/resource/hiscam/ca/02245
http://data.socialhistory.org/resource/hiscam/de/02245
http://data.socialhistory.org/resource/hiscam/e/02245
http://data.socialhistory.org/resource/hiscam/fr/02245
http://data.socialhistory.org/resource/hiscam/gb/02245
http://data.socialhistory.org/resource/hiscam/l/02245
http://data.socialhistory.org/resource/hiscam/nl/02245
http://data.socialhistory.org/resource/hiscam/se/02245
http://data.socialhistory.org/resource/hiscam/u1/02245
http://data.socialhistory.org/resource/hiscam/hisco
And
VU UNIVERSITY AMSTERDAM
CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

Shared - Vocabularies



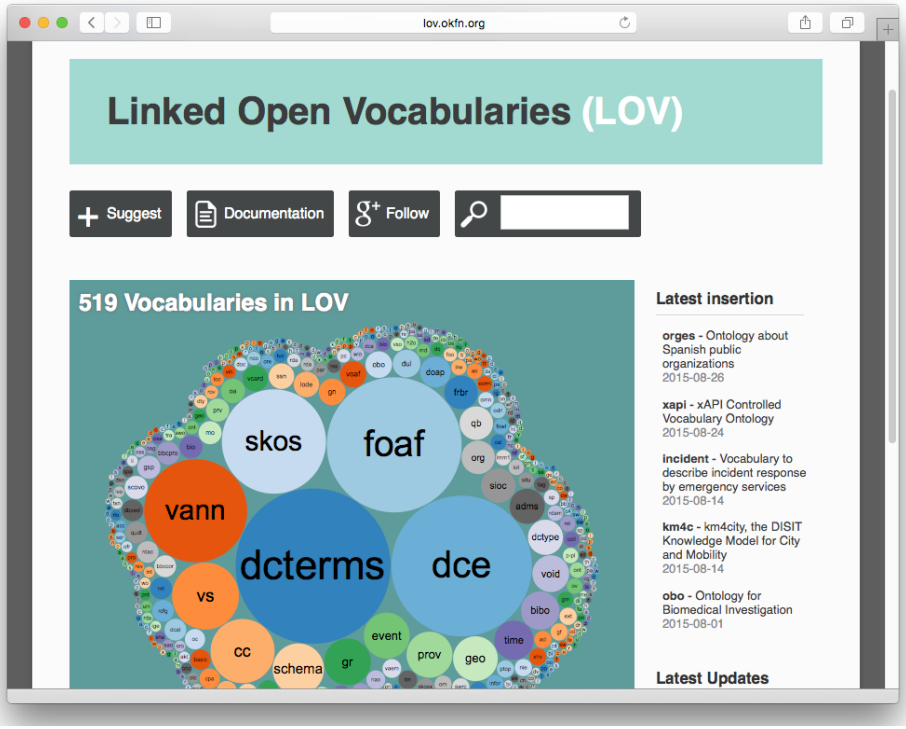

```
<CMD xmlns="http://www.clarin.eu/cmd/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" CMDVersion="1.1"
 xsi:schemaLocation="http://www.clarin.eu/cmd
 http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1289827960126/xsd">
  <Header>
    <MdCreator>lrt2cmdi.py</MdCreator>
    <MdCreationDate>2011-11-10</MdCreationDate>
    <MdSelfLink>http://www.clarin.eu/node/1001</MdSelfLink>
    <MdProfile>clarin.eu:cr1:p_1289827960126</MdProfile>
    <MdCollectionDisplayName>CLARIN LRT inventory</MdCollectionDisplayName>
  </Header>
  <Resources>
    <ResourceProxyList>
      <ResourceProxy id="reflink">
        <ResourceType>Resource</ResourceType>
        <ResourceRef>http://genoma.iula.upf.edu:8080/genoma/index.jsp</ResourceRef>
      </ResourceProxy>
    </ResourceProxyList>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <LrtInventoryResource>
      <LrtCommon>
        <ResourceName>GENOMA</ResourceName>
        <Institute>
          Institut Universitari de Lingüístic
        </Institute>
        <ResourceType>Written Corpus</Resourc
        <LanguagesOther/>
        <Description>
          Bilingual written corpus (2.600.603
        </Description>
        <ContactPerson>iulasecretaria@upf.edu
        <Format/>
        <MetadataLink>
          http://gilmere.upf.edu/oai/?verb=ListRecords&metadataPrefix=oai_clarin
        </MetadataLink>
        <Publications/>
        <ReadilyAvailable>true</ReadilyAvailable>
        <ReferenceLink>http://genoma.iula.upf.edu:8080/genoma/index.jsp</ReferenceLink>
        <Languages>
          <ISO639>
            <iso-639-3-code>cat</iso-639-3-code>
          </ISO639>
          <ISO639>
            <iso-639-3-code>spa</iso-639-3-code>
          </ISO639>
        </Languages>
```

Shared - Vocabularies

- RDF vocabularies (or ontologies) define **semantics** of **classes** and **properties**

(… not prescribed structure …)

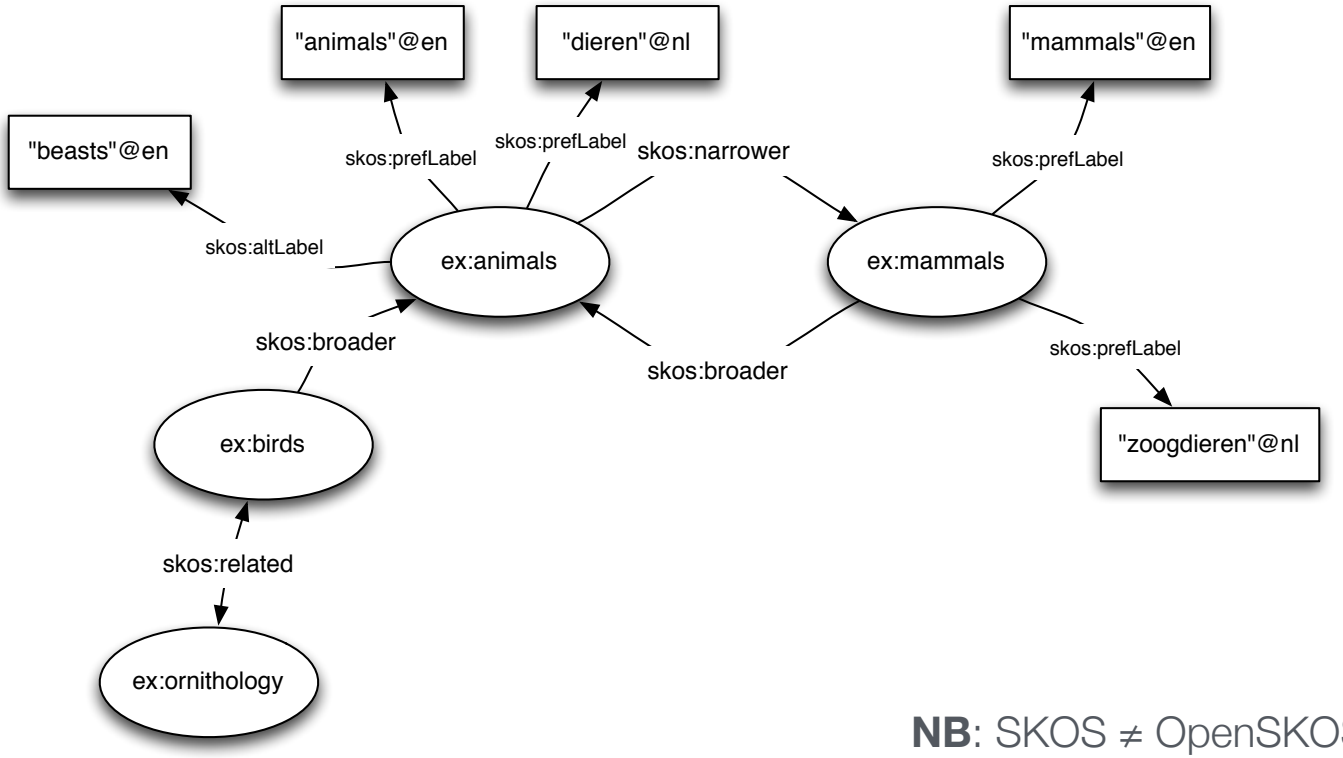
- They are **also** expressed as RDF graphs



VU UNIVERSITY AMSTERDAM

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

SKOS - Vocabulary for Thesaurus Modeling

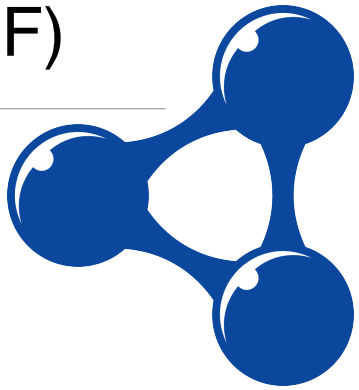


NB: SKOS ≠ OpenSKOS

VU UNIVERSITY AMSTERDAM

CLARIAH Common Lab Research Infrastructure for the Arts and Humanities

The **Resource Description Framework** (RDF)



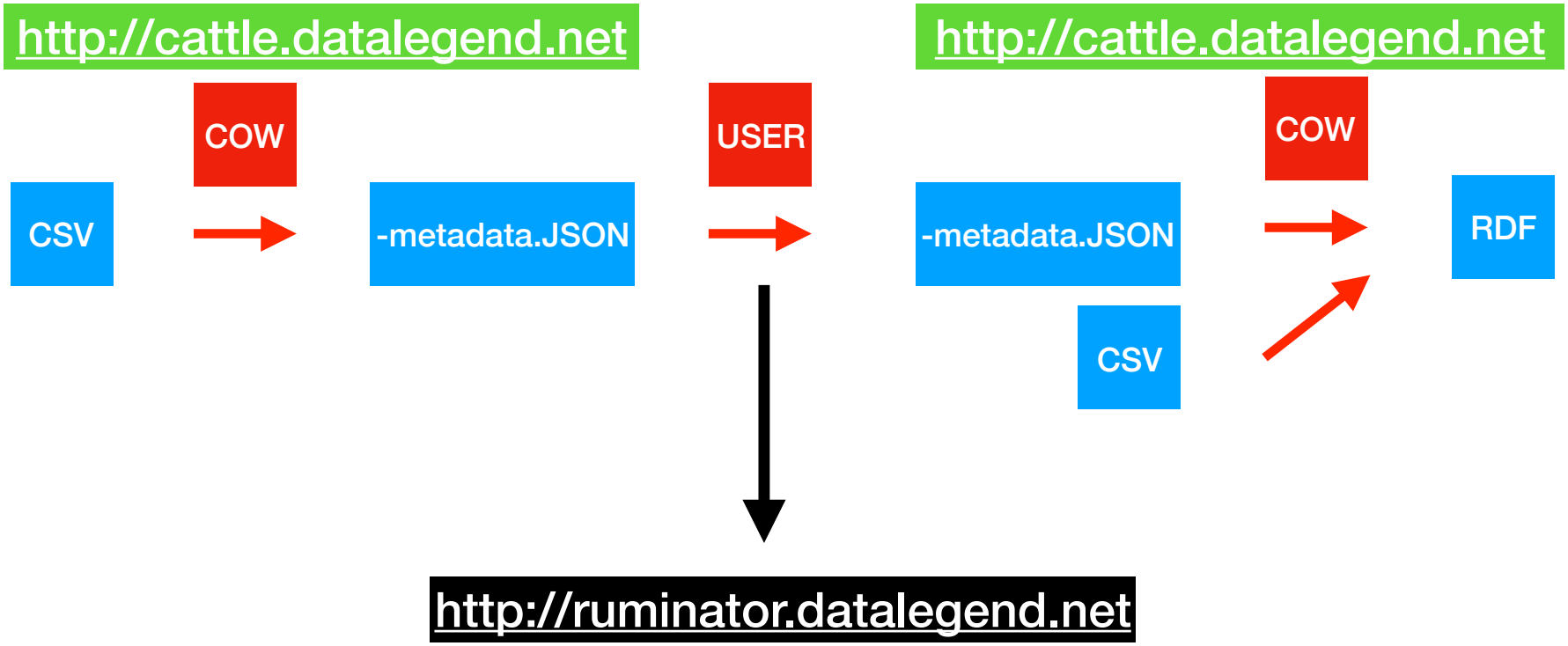
- RDF is a standard **data model** for data interchange on the Web
- Expressive: its graph structure can **express** more than other models
- Flexible: it facilitates **data merging** even if the underlying schemas differ.
- Flexible: it facilitates **data reuse** even if you understand only part of the data.
  … data interpretation relies less on custom code.
- Shared: It extends the **linking structure** of the Web to use URIs.
- It allows data to be mixed, exposed and shared across different applications

http://www.w3.org/RDF/

VU UNIVERSITY AMSTERDAM

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

From CSV to RDF



what did I do today?

- learn why people prefer Linked Data over ‘wasted’ data
- Linked Data terms: triples, subject, predicate, object, sparql, 5-star, URI, URL, prefix, SELECT, *, WHERE, {}, {{}}, |, @en
- terms I will immediately forget: SWAGGER (really?), Numpy
- created your (first) Linked Data
- created your (first) endpoint
- created your (first) github commit
- created your (first) SPARQL query
- created your (first) API
- created your (first) online-shared-executable SPARQL query (for those who weren’t here today)