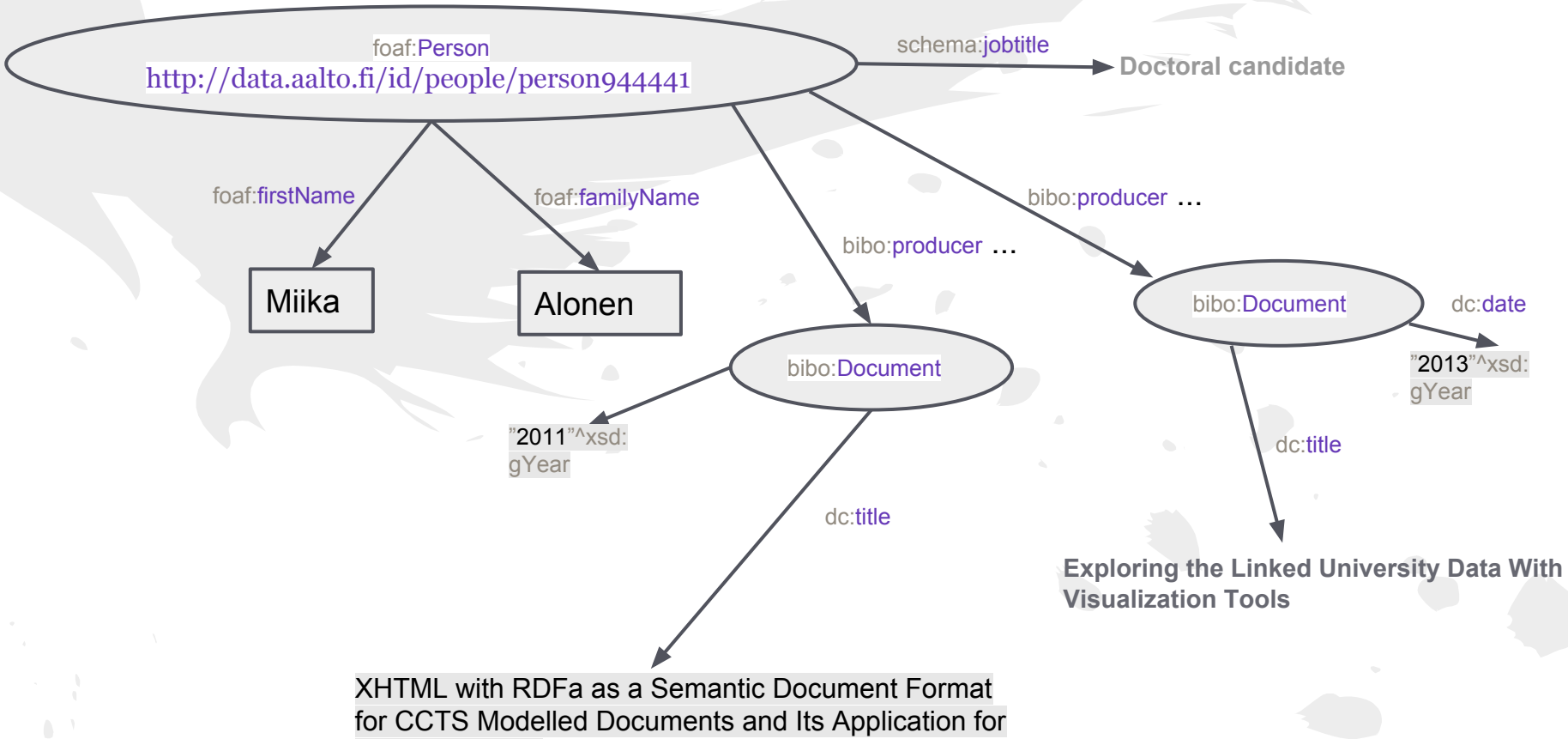
Linkitetyn Datan automaattinen analysointi ja dokumentointipalvelu

vocab.at

Miika Alonen

24.1.2014

About me


foaf:Person
http://data.aalto.fi/id/people/person944441
schema:jobtitle
Doctoral candidate
foaf:firstName
foaf:familyName
Miika
Alonen
bibo:producer …
bibo:producer …
bibo:Document
dc:date
"2013"^xsd:gYear
bibo:Document
"2011"^xsd:gYear
dc:title
dc:title
Exploring the Linked University Data With Visualization Tools
XHTML with RDFa as a Semantic Document Format for CCTS Modelled Documents and Its Application for Social Services

Motivaatio

Quality of Linked Data

2289 datasets

62,224,812,703 triples from **870 datasets**:

- 1,563,527,469 triples from **717 dumps**
- 60,661,285,234 from **156 datasets via SPARQL**

Problems with **1416 datasets** = **61.9%**

(**1061 dumps having errors**, **355 SPARQL endpoints with errors**)

Status from:
http://datahub.io/

at 27.9.2013

analyzed by http://stats.lod2.eu/

What's Wrong with Linked Data?

"The last issue of Semantic Web journal received 27 submissions of Linked Data publications and half of them had to be rejected due to issues with dataset quality or usefulness"

http://blog.semantic-web.at/2012/08/09/whats-wrong-with-linked-data/

- Osaamisen puute?
  > 30 linkitettyyn dataan liittyvää suositusta
- Ei laadunhallintaan liittyviä W3C suosituksia
  > Rakenteen validointi riittämätöntä
- Yksi suuri ongelma on laatupoikkeamien löytämisessä suuresta määrästä dataa

Linkitetyn datan sanastot

- Käytettävissä oleva sanasto määritellään RDF-muodossa ja dokumentoidaan ihmisille HTML-sivustoina
- Datassa käytetty sanasto ei kuitenkaan aina vastaa mitään yksittäistä sanastoa
- Objektiivisen dokumentaation puute johtaa helposti väärinymmärryksiin ja vääriin johtopäätöksiin tietosisällöstä

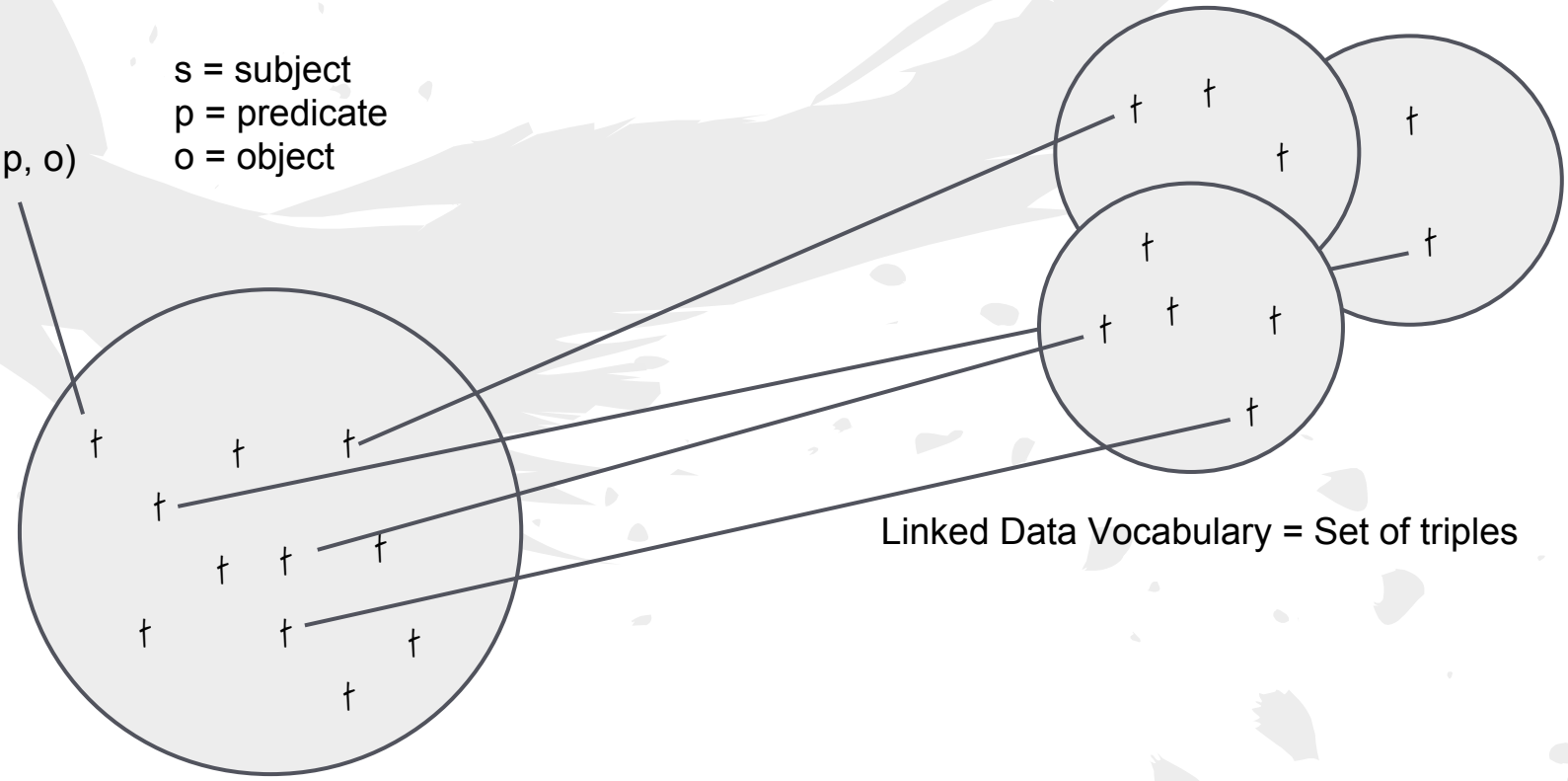
MIKSI?

Linkitetty data on koneluettavaa ja se voidaan analysoida ja dokumentoida automaattisesti!

Linkitetty data

s = subject
p = predicate
o = object

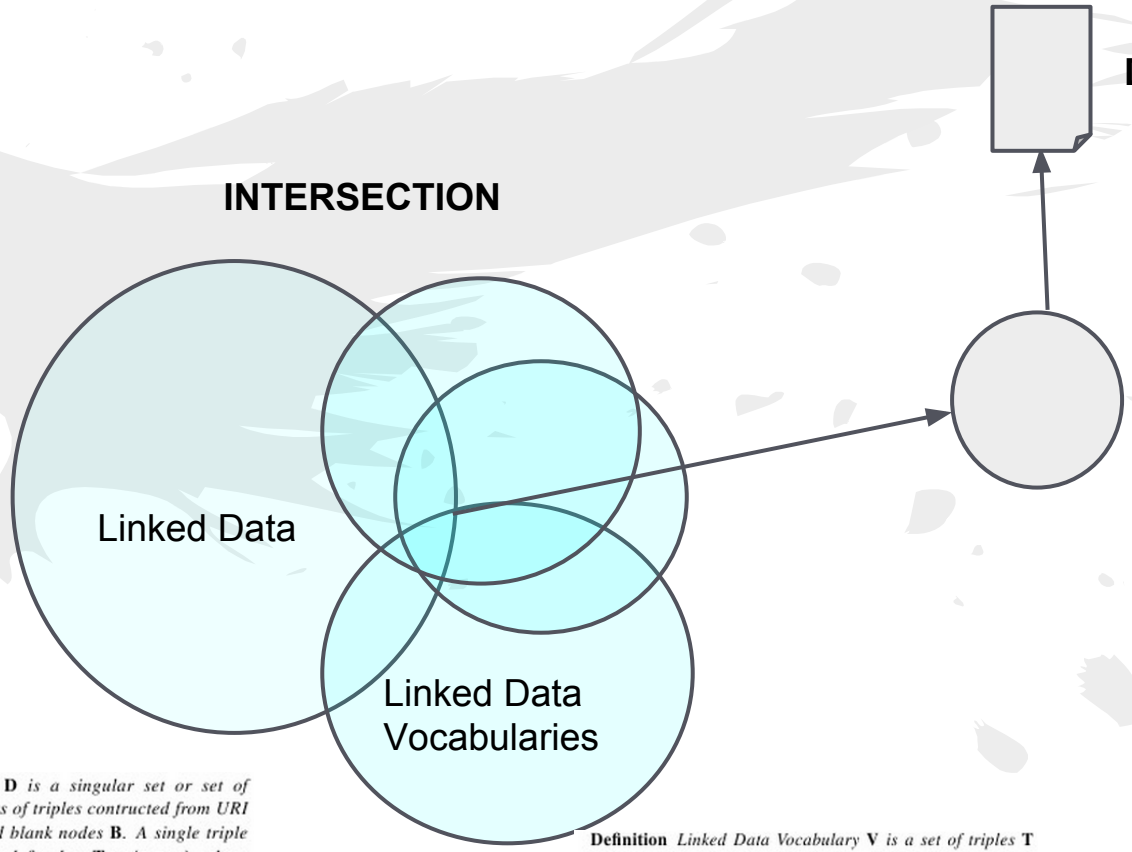
(s, p, o)



Linked Data Vocabulary = Set of triples

Linked Data = Set of triples

Linkitetyn datan analysointi ja dokumentointi



Dataset description

Definition *Dataset description* **DS** *is a set of triples* **T** *that describe the used vocabulary terms from* **V** *and the amount of data in dataset* **D**. *The usage of vocabularies are expressed as class usage* $\mathbf{Cu} = \sum \mathbf{o} \in \mathbf{D}$ *and property usage* $\mathbf{Pu} = \sum \mathbf{p} \in \mathbf{D}$, *where* $\mathbf{p} \cup \mathbf{o} \in \mathbf{V}$. *Dataset description can also include other metrics like Number of entities in* $\mathbf{s} \in \mathbf{T}$, *number of triples* $(\mathbf{s}, \mathbf{p}, \mathbf{o}) \in \mathbf{T}$ *and Unresolved Vocabulary Usage* $\mathbf{U} = \sum(\mathbf{p} \cup \mathbf{o}) \notin \mathbf{V}$ *meaning the sum of triples that has URI references which are not deferenceable.*

Definition *Linked Dataset* **D** *is a singular set or set of graphs* **G**, *which contain sets of triples contructed from URI references* **U**, *Literals* **L** *and blank nodes* **B**. *A single triple* **T** *is a construct that can be defined as* $\mathbf{T} = \{\mathbf{s}, \mathbf{p}, \mathbf{o}\}$, *where* $\mathbf{s} \in \mathbf{U} \cup \mathbf{B}$, $\mathbf{p} \in \mathbf{U}$ *and* $\mathbf{o} \in \mathbf{U} \cup \mathbf{B} \cup \mathbf{L}$.

Definition *Linked Data Vocabulary* **V** *is a set of triples* **T** *that describe the classes* **C** *and properties* **P** *which are identified with URIs meaning* $\mathbf{C} \cup \mathbf{P} \in \mathbf{U}$. *Any dataset* **D** *may use class* $\mathbf{o} \in \mathbf{C}$ *or property* $\mathbf{p} \in \mathbf{P}$ *from the vocabulary* **V** *to describe a resource.*

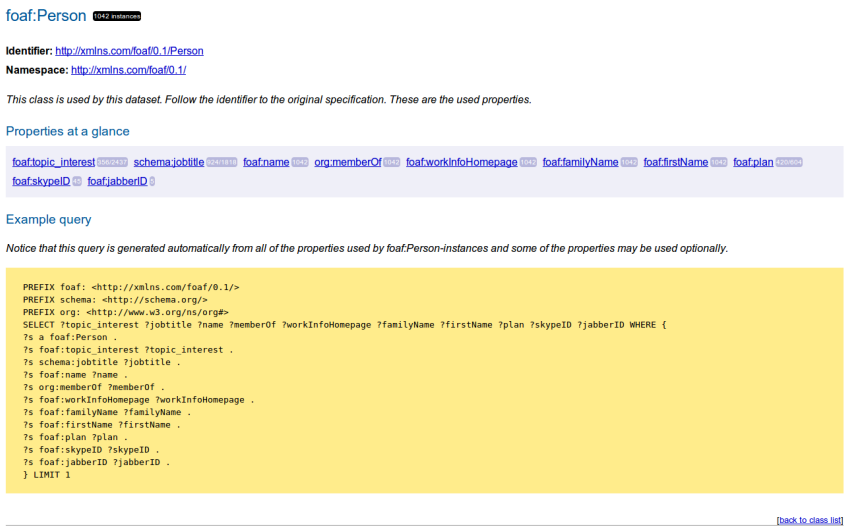
http://vocab.at

Linkitetyn datan automaattinen dokumentointi ja analysointipalvelu

Example 1: Dataset description

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix void: <http://rdfs.org/ns/void#> .

<http://vocab.at/id/3Ivr> a <prov:Activity> ;
 <dct:source> <http://vocab.at/info> ;
 <prov:startedAtTime> "2013-09-24T08:00:46.122Z"^^
     xsd:dateTime ;
 <prov:endedAtTime> "2013-09-24T08:00:48.249Z"^^xsd
     :dateTime ;
 <prov:generated>
    [ a <void:Dataset> ;
      <dct:source> <http://vocab.at/info> ;
      <void:classPartition>
       [ <void:class> foaf:Document ;
         <void:distinctSubjects> 1 ;
         <void:propertyPartition>
           [ <void:entities> 1 ;
             <void:property> dct:subject ;
             <void:triples> 4
           ] ;
       ];
      <void:vocabulary>
      <http://purl.org/dc/elements/1.1/> ,
      <http://xmlns.com/foaf/0.1/> ;
      # continues, see http://vocab.at/data/30@v
    ];
```



foaf:Person 1042 instances

Identifier: http://xmlns.com/foaf/0.1/Person

Namespace: http://xmlns.com/foaf/0.1/

This class is used by this dataset. Follow the identifier to the original specification. These are the used properties.

Properties at a glance

foaf:topic_interest 956/2439 schema:jobtitle 924/1818 foaf:name 1042 org:memberOf 1042 foaf:workInfoHomepage 1042 foaf:familyName 1042 foaf:firstName 1042 foaf:plan 1/1004 foaf:skypeID 48 foaf:jabberID 8

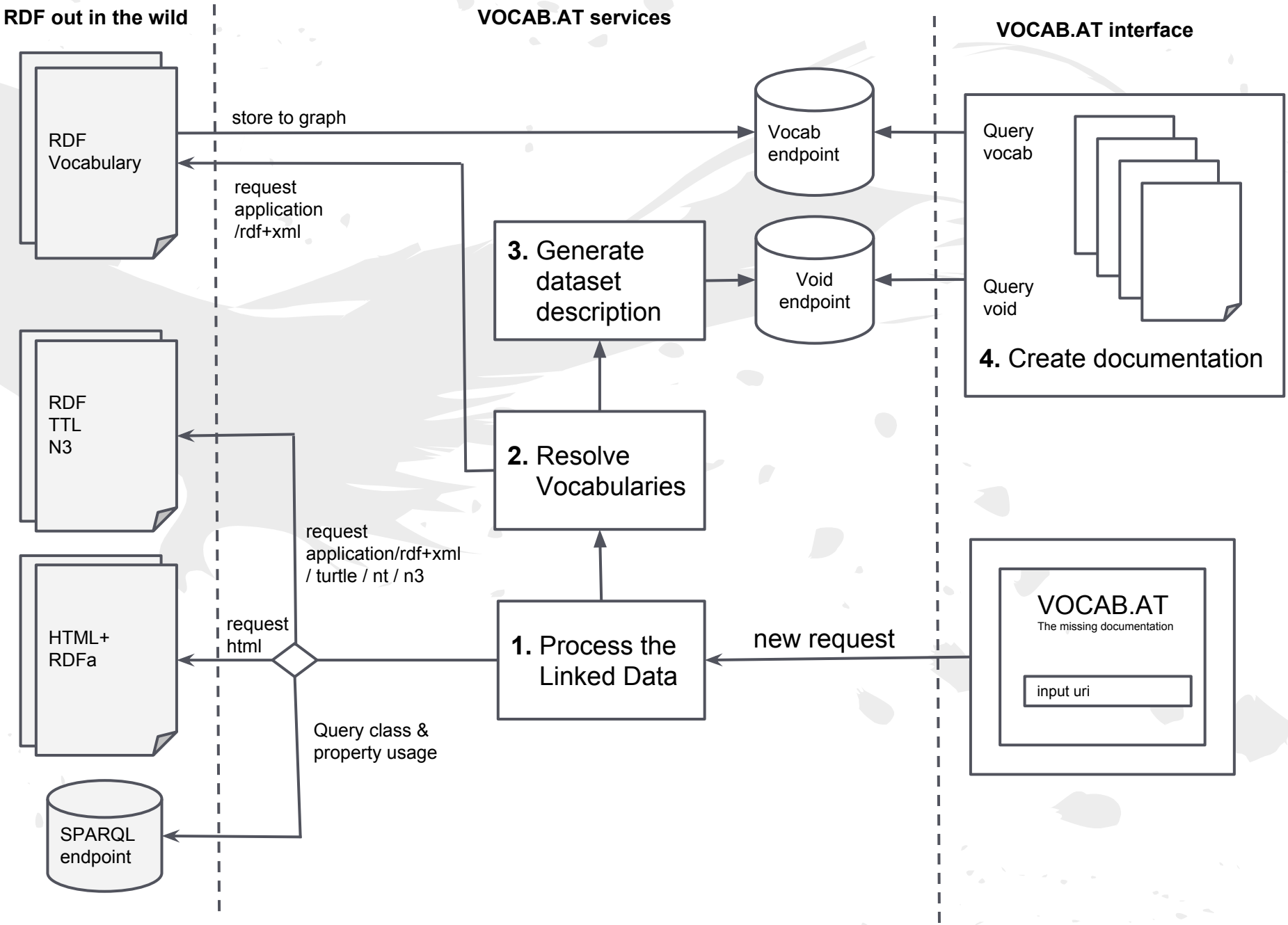
Example query

Notice that this query is generated automatically from all of the properties used by foaf:Person-instances and some of the properties may be used optionally.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX schema: <http://schema.org/>
PREFIX org: <http://www.w3.org/ns/org#>
SELECT ?topic_interest ?jobtitle ?name ?memberOf ?workInfoHomepage ?familyName ?firstName ?plan ?skypeID ?jabberID WHERE {
?s a foaf:Person .
?s foaf:topic_interest ?topic_interest .
?s schema:jobtitle ?jobtitle .
?s foaf:name ?name .
?s org:memberOf ?memberOf .
?s foaf:workInfoHomepage ?workInfoHomepage .
?s foaf:familyName ?familyName .
?s foaf:firstName ?firstName .
?s foaf:plan ?plan .
?s foaf:skypeID ?skypeID .
?s foaf:jabberID ?jabberID .
} LIMIT 1
```

[back to class list]

Esim: http://vocab.at?uri=http://vocab.at/info


RDF out in the wild
VOCAB.AT services
VOCAB.AT interface
RDF Vocabulary
store to graph
Vocab endpoint
Query vocab
request application /rdf+xml
3. Generate dataset description
Void endpoint
Query void
4. Create documentation
RDF TTL N3
2. Resolve Vocabularies
request application/rdf+xml / turtle / nt / n3
VOCAB.AT
The missing documentation
input uri
HTML+ RDFa
request html
1. Process the Linked Data
new request
Query class & property usage
SPARQL endpoint

Preliminary Case Study

Table 1
Dataset Vocabulary Assesment

Dataset	Score	Issues
EARTh Dataset[10]	1	0
Linked Amazon Dataset[11]	~~0.004~~ 1	3
OGOLOD Dataset[12]	0.269	40
LOD EUScreen Dataset[13]	0.96	3
Kirjasampo Dataset[14]	0.173	13
AEMET Dataset[15]	0.556	1
TourMISLOD Dataset[16]	0.17	2

Definition *Vocabulary score* $\mathbf{V} = 1 - \frac{UT}{T}$ *where* UT *is sum of properties and classes not defined in any vocabulary,* $UT = \sum \mathbf{P} \cup \mathbf{C} \notin \mathbf{V}$ *and* $\mathbf{T}$ *is total number of triples in the dataset,* $\mathbf{T} = \sum (\mathbf{s}, \mathbf{p}, \mathbf{o}) \in \mathbf{D}$.

KIITOS!