We have big data, but we need big knowledge

Weaving surveys into the semantic web

ASC Big Data Conference

September 26th 2014

So much knowledge, so little time

3 takeaways

- What are linked data and the semantic web?
- What are the principal technologies?
  - Resource Description Framework (RDF)
  - Triple stores (and relation to SQL/NoSQL databases)
  - SPARQL: SPARQL Protocol and RDF Query Language
- How can this be useful in survey practice?
  - Make sure information you want to share is available online as **Linked (Open) Data**
  - You can use ***one*** standardised technology to integrate:
    - Metadata for multiple surveys
    - Individual data for multiple surveys
    - Aggregate data for multiple surveys ***AND***
    - Arbitrary external data

Linked data bottom-up

Converting conventional data to semantic data:

- Spreadsheet
- CSV File

Unifying and linking:

- Universal data model: graph
- Universal naming: URI
- Latent data on your web page: RDFa

World Wide Web → Giant Global Graph

RDF in a nutshell (1)

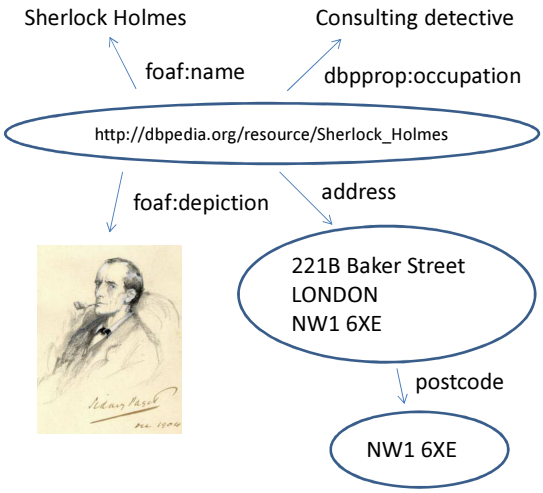
- All your and everyone else's data can be represented by a *directed graph*
- Directed graphs have *nodes* and *edges*
- Directed graphs can be represented as a list of *triples,* one for each edge
- Graphs can be merged simply by combining their triples
- Therefore, in principle, all our data comprise one "Giant Global Graph"

RDF in a nutshell (2)

- RDF provides:
  - Mapping of graphs into WWW concepts:
    - Nodes -> URIs
    - Edges -> Links
  - Formats for serialising graphs into triples
  - A language for managing and querying graphs: SPARQL

Enhancing a postcode – NW1 6XE

Source: Google Streetview



Postcode properties (CSV files)

Ordnance survey Codepoint data

http://www.ordnancesurvey.co.uk/business-and-government/products/code-point-open.html

PC,PQ,EA,NO,CY,RH,LH,CC,DC,WC
…
"NW1 6XB",10,527873,182010,"E92000001","E19000003","E18000007","","E09000033","E05000632"
"NW1 6XE",10,527849,182139,"E92000001","E19000003","E18000007","","E09000033","E05000632"
"NW1 6XN",10,527808,182196,"E92000001","E19000003","E18000007","","E09000033","E05000632"
…

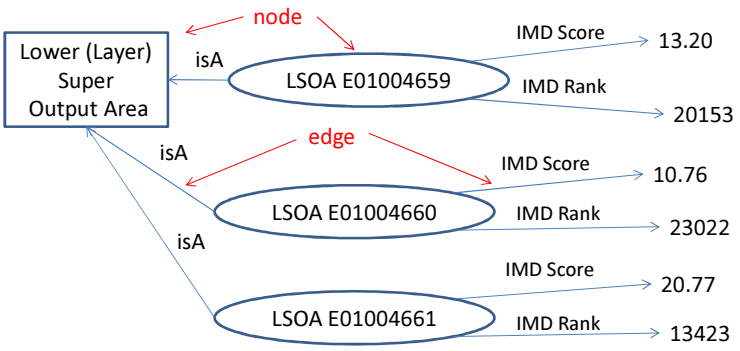
Office for National Statistics geography

http://data.gov.uk/dataset/enumeration-postcodes-2011-to-output-areas-2011-to-lower-layer-super-output-areas-2011-to-middl/resource/8138c00b-37b8-4c8f-b105-a585e4745f74

"PCD7","PCD8","OA11CD","LSOA11CD","LSOA11NM","MSOA11CD","MSOA11NM","LAD11CD","LAD11NM","LAD11NMW","PCDOASPLT"
…
"NW1 6AL","NW1 6AL","E00023522","E01004660","Westminster 008B","E02000967","Westminster 008","E09000033","Westminster","",0
"NW1 6XE","NW1 6XE","E00023522","E01004660","Westminster 008B","E02000967","Westminster 008","E09000033","Westminster","",0
"NW1 6AR","NW1 6AR","E00023522","E01004660","Westminster 008B","E02000967","Westminster 008","E09000033","Westminster","",0
…

Graphing spreadsheet data

	A	B	C	D	E	F	G
1	LSOA CODE	LA CODE	LA NAME	GOR CODE	GOR NAME	IMD SCORE	RANK OF IMD SCORE (where 1 is most deprived)
4660	E01004659	00BK	City of Westminster	H	London	13.20	20153
4661	E01004660	00BK	City of Westminster	H	London	10.76	23022
4662	E01004661	00BK	City of Westminster	H	London	20.77	13423



Source: UK Department for Communities and Local Government, Indices of Deprivation 2010

Deprivation extremes

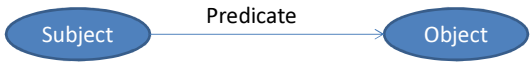
Lowest 0.53:
Chorleywood

Highest 87.8:
Clacton-on-Sea

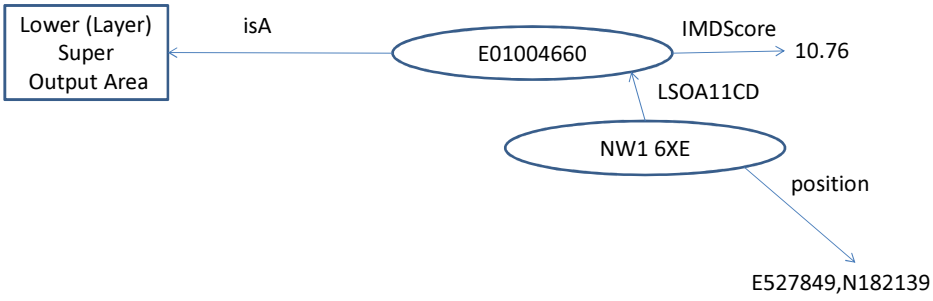
Triples from spreadsheet

Subject	Predicate	Object
E01004659	isA	LSOA
E01004659	IMDScore	13.20
E01004659	IMDRank	20153
E01004660	isA	LSOA
E01004660	IMDScore	10.76
E01004660	IMDRank	23022
E01004661	isA	LSOA
E01004661	IMDScore	20.77
E01004661	IMDRank	13423

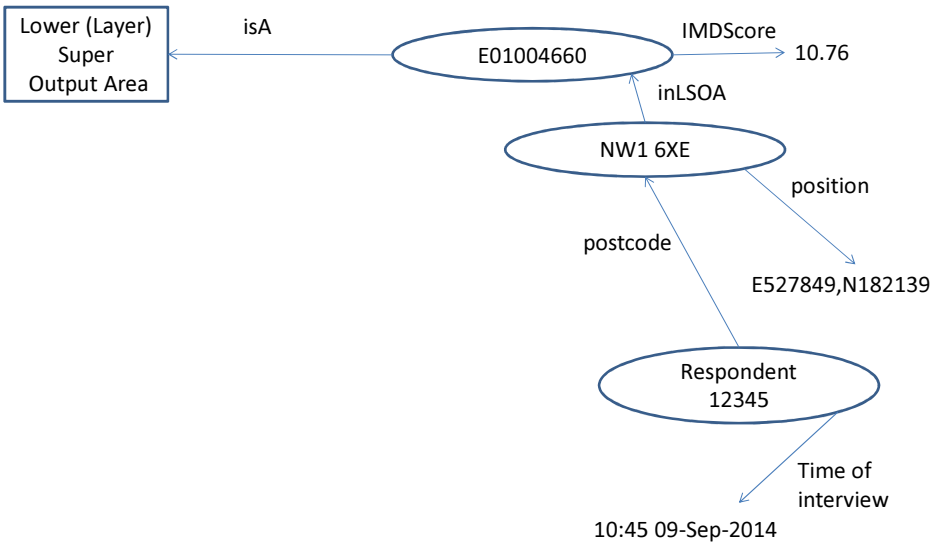
One triple per edge of the (directed)graph



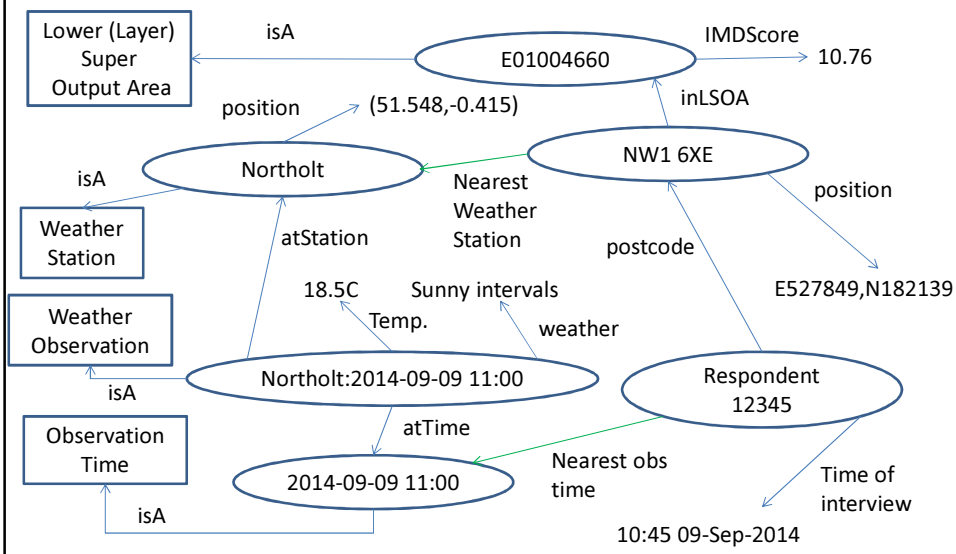
UK geography graph (1)



UK geography graph (2)



UK geography graph (3)



Stable identifiers

The same "thing" should have the same identifier in all contexts so that graphs can merge

Postcode: <http://data.ordnancesurvey.co.uk/id/postcodeunit/NW16XE>

LSOA: http://opendatacommunities.org/id/geography/lsoa/E01004660

LSOA type: <http://opendatacommunities.org/def/geography#LSOA> .

Triple:

<http://opendatacommunities.org/id/geography/lsoa/E01004660>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://opendatacommunities.org/def/geography#LSOA> .

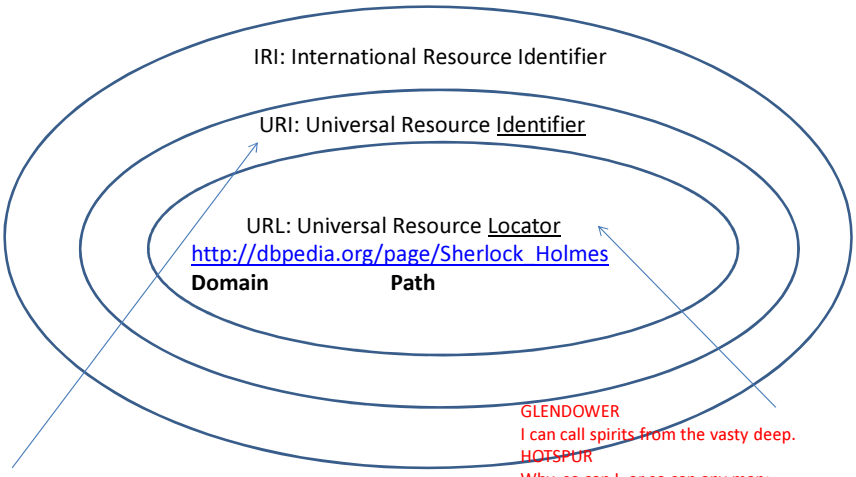
Ontological statement:

<http://opendatacommunities.org/id/geography/lsoa/E0100466>
<http://www.w3.org/2002/07/owl#sameAs>
<http://statistics.data.gov.uk/id/statistical-geography/E01004660> .

Choosing identifiers

- Consistent identifiers unite data
- Universal Resource Identifiers (URIs) are suitable:
  - Two components playing together:
    - Domain name: consensus on ownership
    - Path: owner controls
  - Familiar to everyone
  - Can link to relevant information

URIs, URLs and IRIs



There are more things in heaven and earth, Horatio,
Than are dreamt of in (y)our philosophy.
- Hamlet (1.5.167-8), Hamlet to Horatio

GLENDOWER
I can call spirits from the vasty deep.
HOTSPUR
Why, so can I, or so can any man;
But will they come when you do call for them?
Henry IV part 1, 3.1

URL returns a representation: HTML

About: Sherlock Holmes

An Entity of Type : fictional character, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Sherlock Holmes (/ˈʃɜrlɒk ˈhoʊmz/) FRSC is a fictional detective created by Scottish author and physician Sir Arthur Conan Doyle, a graduate of the University of Edinburgh Medical School.

Property	Value
dbpedia-owl:abstract	▪ Sherlock Holmes (/ˈʃɜrlɒk ˈhoʊmz/) FRSC is a fictional detective created by Scottish author and physician Sir Arthur Conan Doyle, a graduate of the University of Edinburgh Medical School. A London-based "consulting detective" whose abilities border on the fantastic, Holmes is famous for his astute logical reasoning, his ability to adopt almost any disguise, and his use of forensic science skills to solve difficult cases.Holmes, who first appeared in publication in 1887, was featured in four novels and 56 short stories. The first novel, A Study in Scarlet, appeared in Beeton's Christmas Annual in 1887 and the second, The Sign of the Four, in Lippincott's Monthly Magazine in 1890. The character grew tremendously in popularity with the first series of short stories in The Strand Magazine, beginning with "A Scandal in Bohemia" in 1891; further series of short stories and two novels published in serial form appeared between then and 1927. The stories cover a period from around 1880 up to 1914.All but four stories are narrated by Holmes's friend and biographer, Dr. John H. Watson; two are narrated by Holmes himself ("The Blanched Soldier" and "The Lion's Mane") and two others are written in the third person ("The Mazarin Stone" and "His Last Bow"). In two stories ("The Musgrave Ritual" and "The Gloria Scott"), Holmes tells Watson the main story from his memories, while Watson becomes the narrator of the frame story. The first and fourth novels, A Study in Scarlet and The Valley of Fear, each include a long interval of omniscient narration recounting events unknown to either Holmes or Watson.
dbpedia-owl:creator	▪ dbpedia:Arthur_Conan_Doyle
dbpedia-owl:firstAppearance	▪ A Study in Scarlet
dbpedia-owl:relative	▪ dbpedia:Mycroft_Holmes
dbpedia-owl:series	▪ dbpedia:Sherlock_Holmes_(play)
dbpedia-owl:thumbnail	▪ http://commons.wikimedia.org/wiki/Special:FilePath/Sherlock_Holmes_Portrait_Paget.jpg?width=300
dbpedia-owl:wikiPageExternalLink	▪ http://sherlockholmes.stanford.edu/index.html

For humans, when accessed by a browser

URL returns a representation: RDF

```
…
@prefix dbpedia:        <http://dbpedia.org/resource/> .
@prefix foaf:           <http://xmlns.com/foaf/0.1/> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
…
dbpedia:Sherlock_Holmes     dbpedia-owl:wikiPageExternalLink        ns99:SherlockHolmesComplete ,
                            <http://www.chesshistory.com/winter/extra/holmes.html> ;
            dbpedia-owl:firstAppearance     "A Study in Scarlet" ;
            dbpprop:caption"Sherlock Holmes in a 1904 illustration by Sidney Paget"@en ;
            dbpprop:colour  "#DEDEE9"@en ;
            dbpprop:creator dbpedia:Arthur_Conan_Doyle ;
            dbpprop:family  dbpedia:Mycroft_Holmes ;
            dbpprop:first   "A Study in Scarlet"@en ;
            dbpprop:gender "Male"@en ;
            dbpprop:name    "Sherlock Holmes"@en ;
            dbpprop:nationality         "British"@en ;
            dbpprop:occupation          "Consulting detective"@en ;
            dbpprop:series  "Sherlock Holmes"@en ;
            dbpprop:title   "Sherlock Holmes related articles"@en ;
            foaf:depiction
            <http://commons.wikimedia.org/wiki/Special:FilePath/Sherlock_Holmes_Portrait_Paget.jpg>;
…
dbpedia:Sherlock_Holmes     owl:sameAs      <http://lv.dbpedia.org/resource/> Šerloks_Holmss .
…
```

When accessed by an application

Serialising your triples

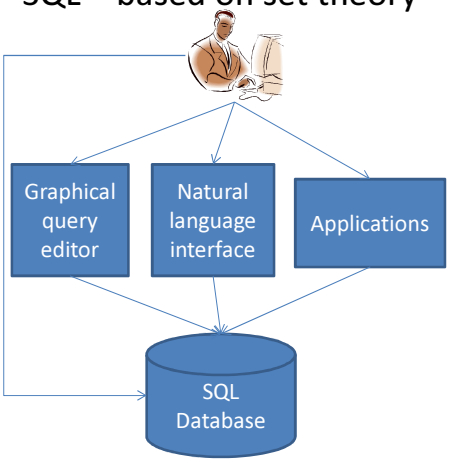
- N3: http://www.w3.org/TeamSubmission/n3/Turtle
- N-triples: http://www.w3.org/TR/n-triples/
- JSON-LD: http://www.w3.org/TR/json-ld/
- Turtle: http://www.w3.org/TR/turtle/
- RDF/XML: http://www.w3.org/TR/REC-rdf-syntax/

Storing RDF triples

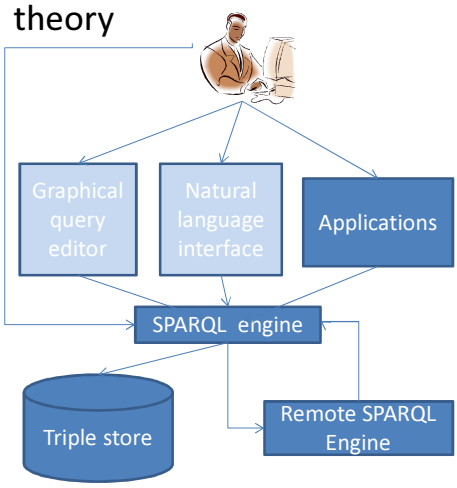
- Storage alternatives
  - Flat files in a serialisation format
  - Custom triple store, e.g.
    - 3Store
    - **Bigdata**
  - Triple store layered on SQL database
  - Triple store layered on graph/NoSQL database
- Libraries to decouple your code from storage strategy:
  - Java: Sesame (http://openrdf.callimachus.net/)
  - Python: **rdflib** (https://github.com/RDFLib/rdflib)

Relationship of SQL and SPARQL

SQL – based on set theory



SPARQL – based on graph theory



Ontologies

Understand ontologies, and you're a data scientist

What is an ontology?

- Ontology is RDF equivalent of a schema
- Words in same space:

Taxonomy
Schema Vocabulary
Thesaurus Dictionary Catalogue
Grammar
Codeframe

- Documents potential relationships between 'resources'
- Differences from database schema:
  - Ontology is optional
  - Permits validation
  - Enables inference
  - Enables meaningful links for sharing data

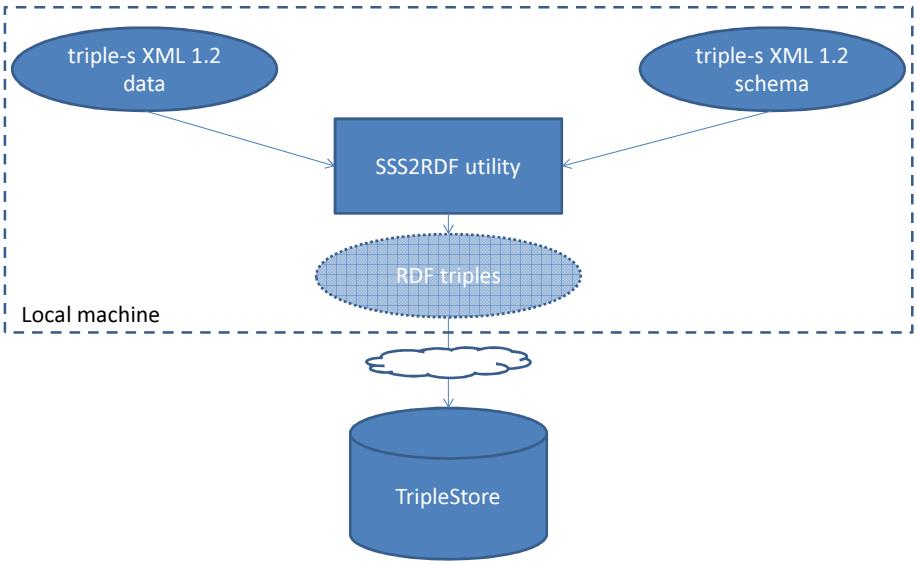
Proof of the pudding…

"The first essential in chemistry is that thou shouldest perform practical work and conduct experiments, for he who performs not practical work nor makes experiments will never attain to the least degree of mastery"

Jabir Ibn Hayyan (Geber)

AD 721 - 815

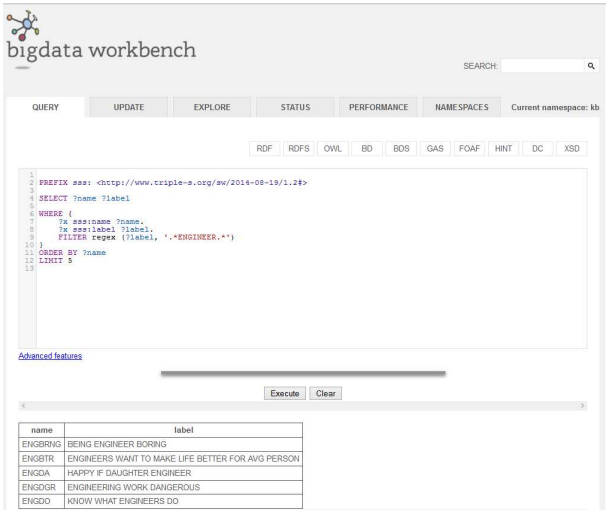
triple-s upload workflow



Some Triple-S RDF

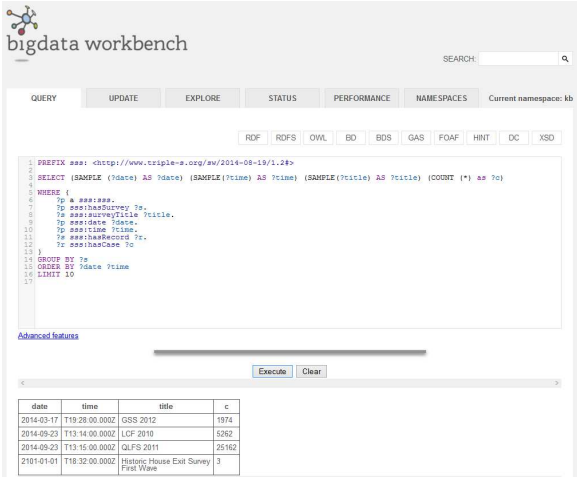
```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sss: <http://www.triple-s.org/sw/2014-08-19/1.2#> .
@prefix survey: <http://rdf.x-mr.com/sss12/example> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://rdf.x-mr.com/sss12/example#sss> a sss:sss ;
    sss:date "2101-01-01"^^xsd:date ;
    sss:hasSurvey <http://rdf.x-mr.com/sss12/example#survey> ;
    sss:origin "Export 1.42" ;
    sss:sssVersion "1.2" ;
    sss:time "18:32:00"^^xsd:time .
...
<http://rdf.x-mr.com/sss12/example#record> a sss:record ;
    sss:hasVariable <http://rdf.x-mr.com/sss12/example#variable_Q1>,
        <http://rdf.x-mr.com/sss12/example#variable_Q2>,
...
        <http://rdf.x-mr.com/sss12/example#variable_Q7>,
        <http://rdf.x-mr.com/sss12/example#variable_Q99> ;
    sss:recordIdent "V" .
...
<http://rdf.x-mr.com/sss12/example#variable_Q2> a sss:variable ;
    sss:finishLocation 10 ;
    sss:hasValues <http://rdf.x-mr.com/sss12/example#values_Q2> ;
    sss:ident "2" ;
    sss:label "Attractions visited" ;
    sss:name "Q2" ;
    sss:startLocation 2 ;
    sss:type "multiple" .
```

All variable titles with “ENGINEER”

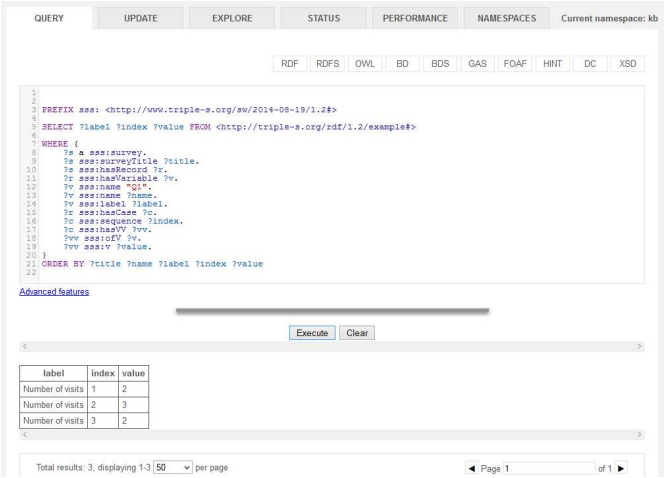


name	label
ENGBRNG	BEING ENGINEER BORING
ENGBTR	ENGINEERS WANT TO MAKE LIFE BETTER FOR AVG PERSON
ENGDA	HAPPY IF DAUGHTER ENGINEER
ENGDGR	ENGINEERING WORK DANGEROUS
ENGDO	KNOW WHAT ENGINEERS DO

List surveys and number of cases



List values of a specific variable



Frequency distribution of a variable

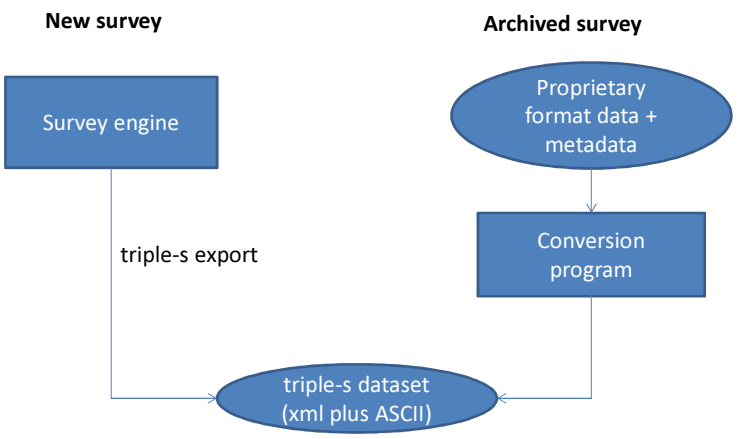
```
PREFIX sss: <http://www.triple-s.org/sw/2014-08-19/1.2#>

SELECT ?vName ?code ?text ?cases
WHERE {
    ?z sss:hasVariable ?v.
    ?v sss:name ?vName.
    ?v sss:name "zodiac".
    { SELECT (SAMPLE (?code) as ?code) (COUNT(*) as ?cases) WHERE {
        ?v rdf:type sss:variable.
        ?v sss:name "zodiac".
        ?vv sss:ofV ?v.
        ?vv sss:iV ?code
      }
      GROUP BY ?code
      ORDER BY ?code
    }
    ?v sss:hasValues ?values.
    ?values sss:hasValue ?value.
    ?value sss:code ?code.
    ?value sss:text ?text.
}
ORDER BY ?code
```

Advanced features

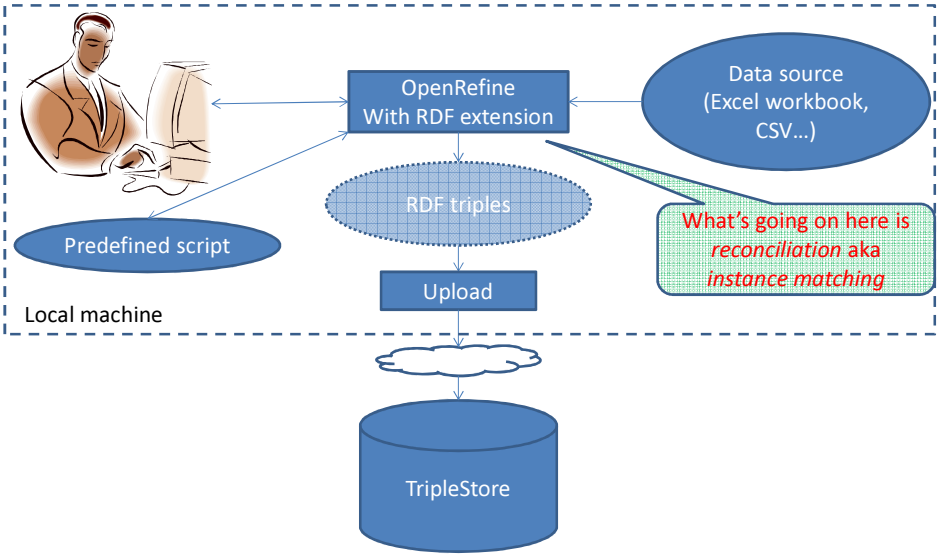
Execute Clear

vName	code	text	cases
zodiac	1	ARIES	145
zodiac	2	TAURUS	162
zodiac	3	GEMINI	163
zodiac	4	CANCER	148
zodiac	5	LEO	186
zodiac	6	VIRGO	161
zodiac	7	LIBRA	178
zodiac	8	SCORPIO	147
zodiac	9	SAGITTARIUS	145
zodiac	10	CAPRICORN	151
zodiac	11	AQUARIUS	170
zodiac	12	PISCES	157

Getting to triple-s



External data upload workflow (OpenRefine example)



Do try this at home

1. Prepare some triple-s data
2. Download the RDF utilities
3. Install a triplestore
4. Populate the store with the survey data
5. Run some queries

Develop some applications you've always needed but were too tedious to build before

Further work - standards

- Survey data ontology
  - Use existing standards:
    - Metadata as DDI RDF
    - aggregate data as SDMX datacube RDF
  - *or* Develop some lean-and-mean alternatives in the spirit of triple-s, on an industry basis
- Product field ontologies – especially in media research

Further work - tools

- User friendly query interfaces
- Workflow tools, e.g. bulk tabulation
- Export of questionnaires and samples
- Interface to statistical systems:
  - Export of data
  - Import of results to reuse as data
- Gateways to industry-wide and public data sources
- Extract data into publication formats