Exposing the SeaDataNet metadata catalogues via SPARQL endpoints

CHRIS WOOD[1,2], ALEXANDRA KOKKINAKI[1], ADAM LEADBETTER[3], ROB THOMAS[3]

1: BRITISH OCEANOGRAPHIC DATA CENTRE, UK
2: (NOW AT) EPCC, UNIVERSITY OF EDINBURGH, UK
3:MARINE INSTITUTE, IRELAND

Why?

In general terms, your research data should be 'FAIR', that is Findable, Accessible, Interoperable and Re-usable.

- “H2020 Programme Guidelines on FAIR Data Management in Horizon 2020”

  Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3 (2016).

Why?

Why?

Why?

SPARQL is to **Linked Data** (graph data) what SQL is to relational databases

- W3C standard query language for Resource Description Framework (RDF) data
- sparql endpoints are publicly available-accessible
- can all be queried in one standard language, SPARQL

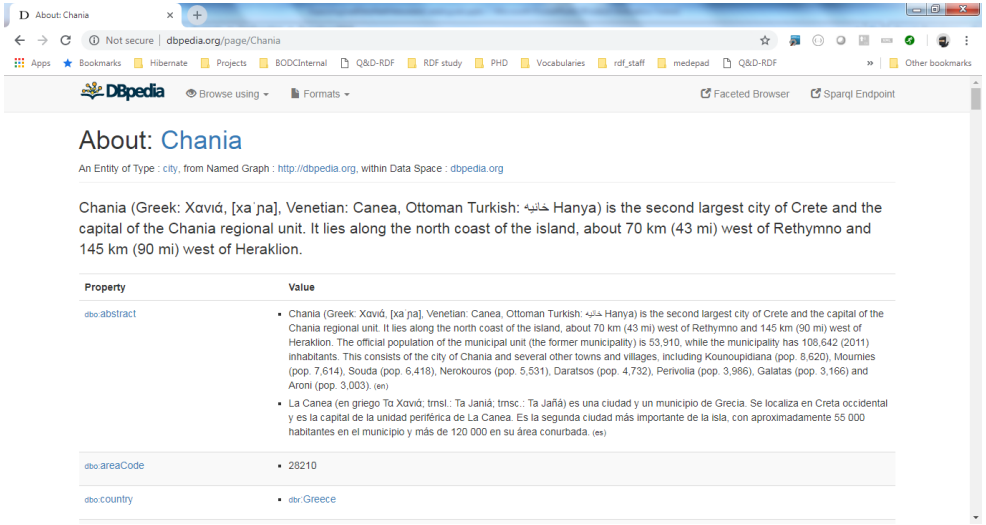
Background

World wide web is evolving from hyper**text** links (link documents) to hyper**data** links (linking data)

"The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With **Linked Data**, when you have some of it, you can find other, related, data."

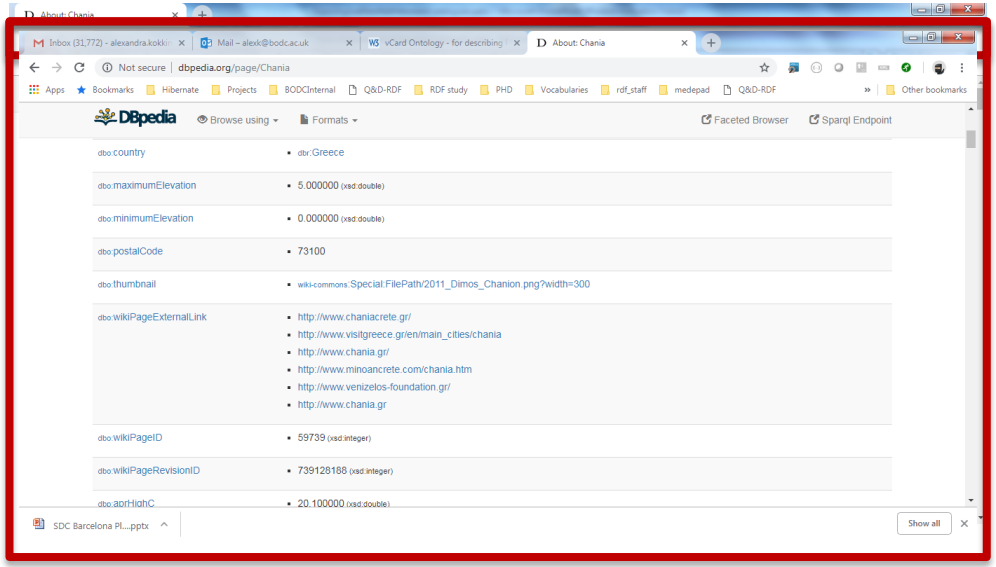
- Sir Tim Berners-Lee, 2006

Background



1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

Background



1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things

Background

EDMERP

Projects (3 079)

EDMO

Organisations (4 523)

CSR

Research cruises (48 836)

EDIOS

Observing programmes (363)

CDI

Data index (2 151 046)

EDMED

Data sets (4 193)

Work so far

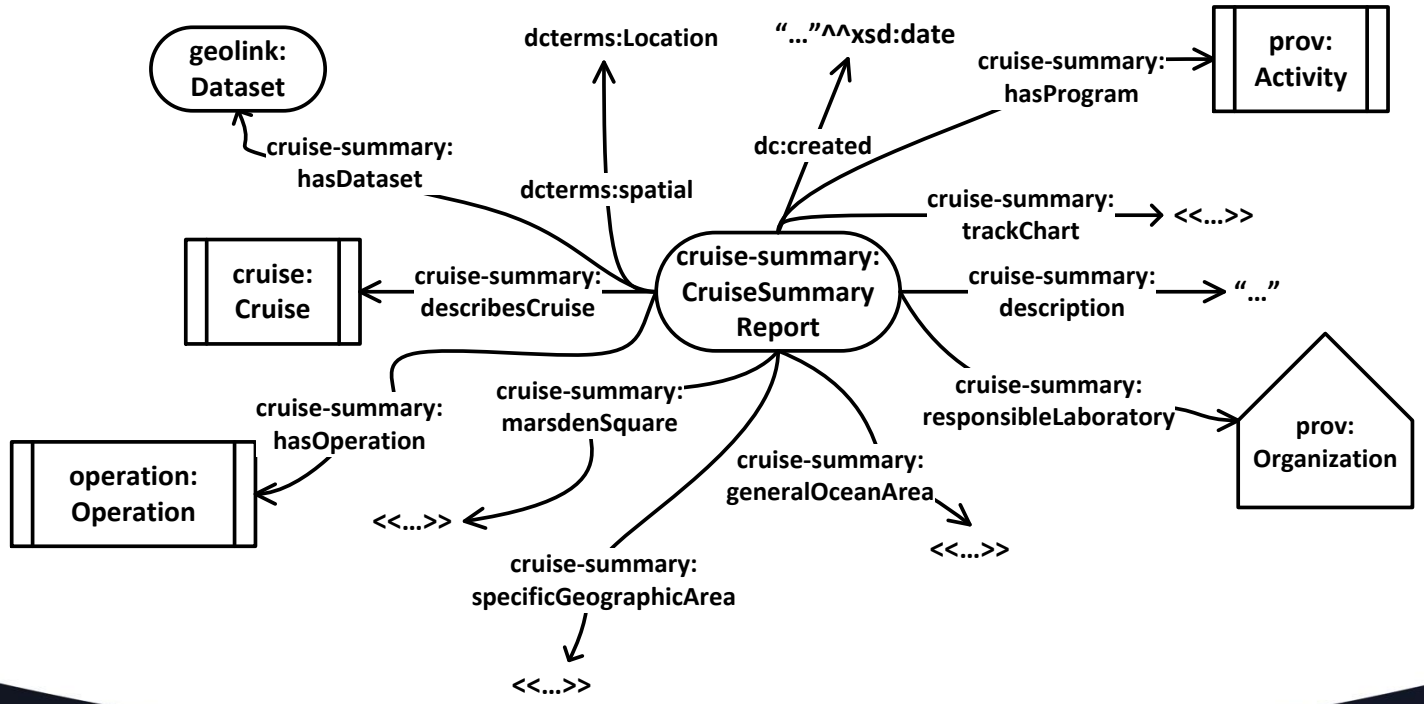
Reusing existing patterns

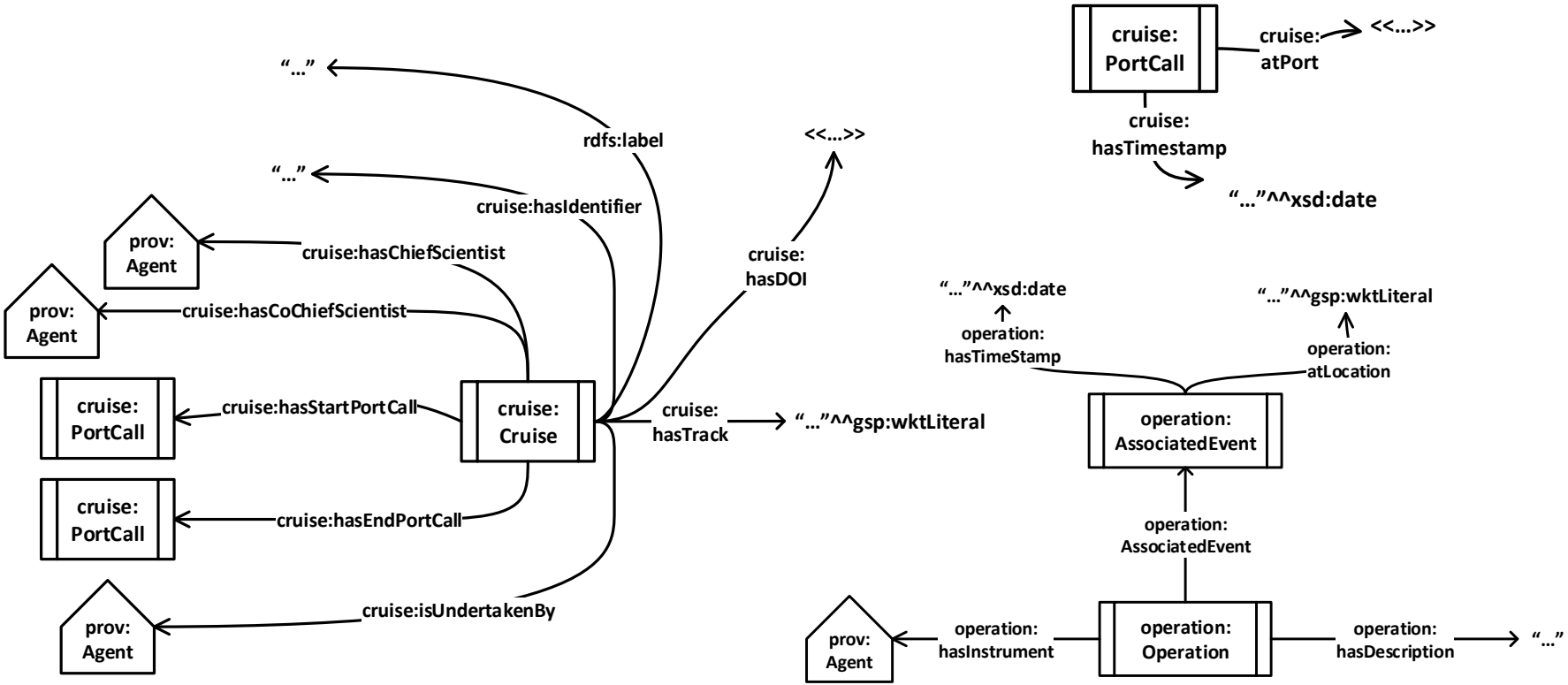
- Better understanding outside of SDN
- Better interoperability with other organisations
- Better INSPIRE compliance

Work so far

- **Reusing existing patterns:**
- EDMO – W3C Organisation
- EDMED – W3C DCAT / W3C Prov
- EDMERP – W3C Prov / DBPedia Research Project
- CDI – W3C DCAT
- ODV metadata to INSPIRE / ISO O&M
- CSR - *Liaised with US-NSF Rolling Deck to Repository & with Australia (through ODIP/SDC)*
- EDIOS – INSPIRE Environmental Monitoring Facilities

CSR: the special catalogue




cruise:
PortCall
cruise:
atPort
<<...>>
cruise:
hasTimestamp
"..."^^xsd:date
"..."
rdfs:label
"..."
cruise:hasIdentifier
<<...>>
cruise:
hasDOI
prov:
Agent
cruise:hasChiefScientist
prov:
Agent
cruise:hasCoChiefScientist
cruise:
PortCall
cruise:hasStartPortCall
cruise:
Cruise
cruise:
hasTrack
"..."^^gsp:wktLiteral
cruise:
PortCall
cruise:hasEndPortCall
prov:
Agent
cruise:isUndertakenBy
"..."^^xsd:date
operation:
hasTimeStamp
"..."^^gsp:wktLiteral
operation:
atLocation
operation:
AssociatedEvent
operation:
AssociatedEvent
prov:
Agent
operation:
hasInstrument
operation:
Operation
operation:
hasDescription
"..."

Work so far

Reusing existing patterns

- Better understanding outside of SDN
- Better interoperability with other organisations
- Better INSPIRE compliance

Implementation

Work so far

- The underlying catalogues are stored at 4 different institutions across Europe
- Technical implementation: the SPARQL endpoints are all available under a shared domain (e.g. http://edmed.seadatanet.org/sparql/)
- No requirement on technical implementations for the catalogue holders
- There are several ways of implementing triplestores from the underlying catalogues: BODC use the Apache Jena stack with custom ingestion code to ensure it's kept up to date

Work so far

- https://edmed.seadatanet.org/
  - https://edmed.seadatanet.org/search/
  - https://edmed.seadatanet.org/sparql/
  - https://edmed.seadatanet.org/report/<ID>
- https://edmo.seadatanet.org/
  - http://edmo.seadatanet.org/sparql/
- https://edios.seadatanet.org/
  - http://linked.bodc.ac.uk/sdn/edios/
- https://edmerp.seadatanet.org/
- https://cdi.seadatanet.org/

Work so far

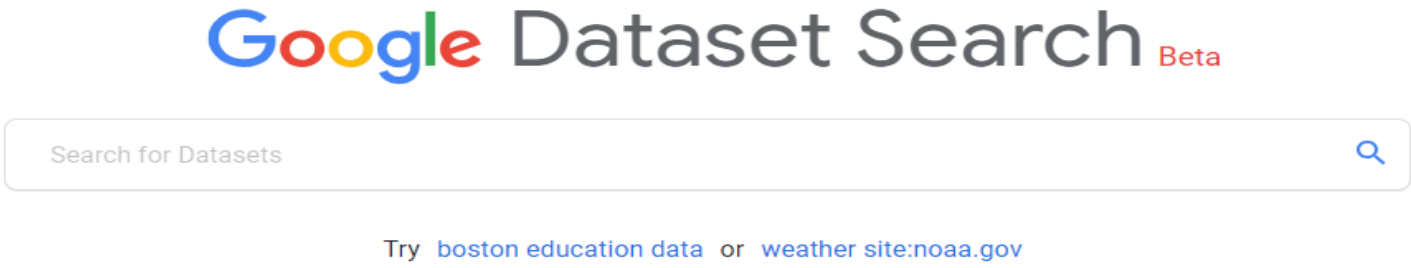
Reusing existing patterns

- Better understanding outside of SDN
- Better interoperability with other organisations
- Better INSPIRE compliance

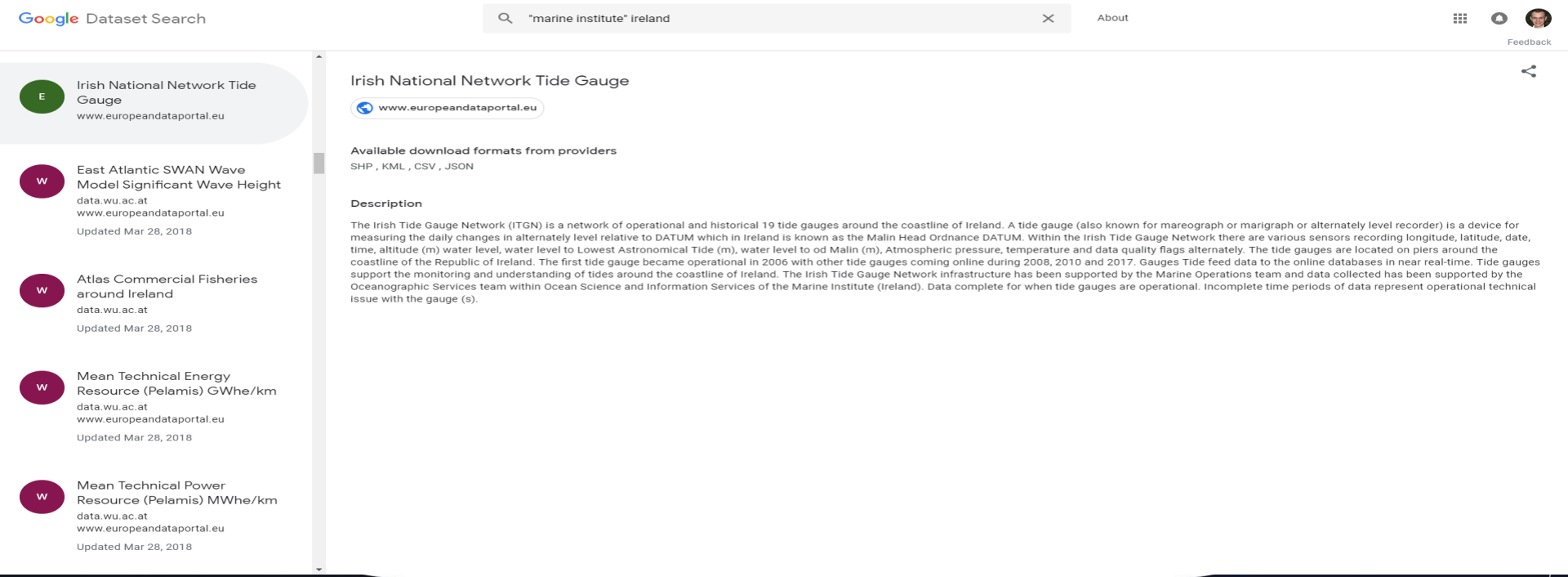
Implementation

Mapping to Schema.org

Work so far


Google Dataset Search Beta
Search for Datasets
Try boston education data or weather site:noaa.gov

Work so far


Google Dataset Search
"marine institute" ireland
About
Feedback
Irish National Network Tide Gauge
www.europeandataportal.eu
East Atlantic SWAN Wave Model Significant Wave Height
data.wu.ac.at
www.europeandataportal.eu
Updated Mar 28, 2018
Atlas Commercial Fisheries around Ireland
data.wu.ac.at
Updated Mar 28, 2018
Mean Technical Energy Resource (Pelamis) GWhe/km
data.wu.ac.at
www.europeandataportal.eu
Updated Mar 28, 2018
Mean Technical Power Resource (Pelamis) MWhe/km
data.wu.ac.at
www.europeandataportal.eu
Updated Mar 28, 2018
Irish National Network Tide Gauge
www.europeandataportal.eu
Available download formats from providers
SHP , KML , CSV , JSON
Description
The Irish Tide Gauge Network (ITGN) is a network of operational and historical 19 tide gauges around the coastline of Ireland. A tide gauge (also known for mareograph or marigraph or alternately level recorder) is a device for measuring the daily changes in alternately level relative to DATUM which in Ireland is known as the Malin Head Ordnance DATUM. Within the Irish Tide Gauge Network there are various sensors recording longitude, latitude, date, time, altitude (m) water level, water level to Lowest Astronomical Tide (m), water level to od Malin (m), Atmospheric pressure, temperature and data quality flags alternately. The tide gauges are located on piers around the coastline of the Republic of Ireland. The first tide gauge became operational in 2006 with other tide gauges coming online during 2008, 2010 and 2017. Gauges Tide feed data to the online databases in near real-time. Tide gauges support the monitoring and understanding of tides around the coastline of Ireland. The Irish Tide Gauge Network infrastructure has been supported by the Marine Operations team and data collected has been supported by the Oceanographic Services team within Ocean Science and Information Services of the Marine Institute (Ireland). Data complete for when tide gauges are operational. Incomplete time periods of data represent operational technical issue with the gauge (s).

Conclusions & lessons learnt

- The RDF model (and associated technologies) are **important** and **powerful tools** for **discovery** and **delivery** of data and metadata
- Like with all new technology, implementation can have a steep learning curve
- Other types of API (e.g. pure JavaScript) will currently be more familiar to 3rd party developers but are often much less powerful
- Need to consider three user-groups: internal users, 3rd party developers, and external users
- The investment required to make and populate a triplestore of your data will be worth it – and your users will appreciate it in the long term!

Future directions

- Data products/applications making use of standardised endpoints
- Libraries e.g.in python
- Example queries and user tools to make sparql endpoints more user friendly
- Data -> Information->Knowledge

Future directions?

Previous work for more explanations

- Presentations at IMDIS 2016 (Gdansk), RSE 2017 (Manchester), Oceanology 2018 (London), ODIP 2018 Meeting, SeaDataCloud 2018 Meeting
- Posters at EGU 2016, EGU 2018, Carhamb'ar 2017
- Paper coming soon!

- c.wood@epcc.ed.ac.uk and/or alexk@bodc.ac.uk for more info!

Acknowledgments

- Justin Buck, Richard Downer, Lesley Rickards (BODC)
- The SeaDataNet community members who have provided valuable feedback
- Developers of D2RQ, Jena, Fuseki, and elda (and all the people who've answered various questions!)
- Funding: SeaDataCloud (2016-2020), EU H2020 programme, grant agreement 730960

Questions?

Code examples for the BODC Series Schema

```
query = "select count(?o) where {
        ?s <http://www.w3.org/2004/02/skos/core#notation> ?o . }"
```

Javascript /jQuery:

```
$.ajax({
     type: "POST",
     url: "http://linked.bodc.ac.uk/sparql?",
     data: "output=json&query=" + query
     dataType: "json"
})
```

Python:

```
r = requests.get('http://linked.bodc.ac.uk/sparql/', {'query':query}, timeout=10)
v = r.json()['results']['bindings'][0]['.1']['value']
print v
```