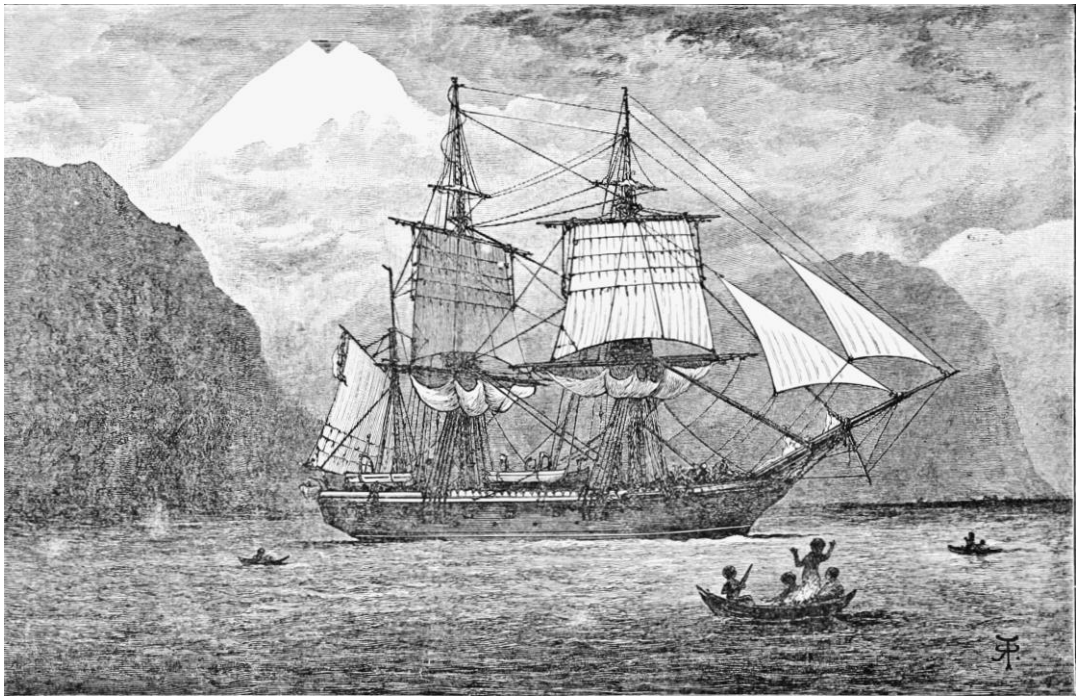
Development of Linked Data Services to Support Widespread Exposure of Data

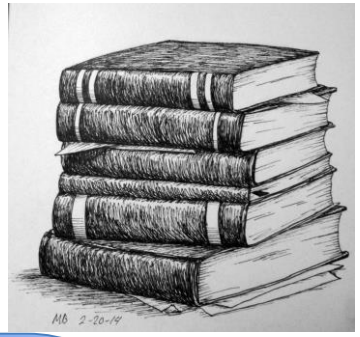
CHRIS WOOD
BRITISH OCEANOGRAPHIC DATA CENTRE

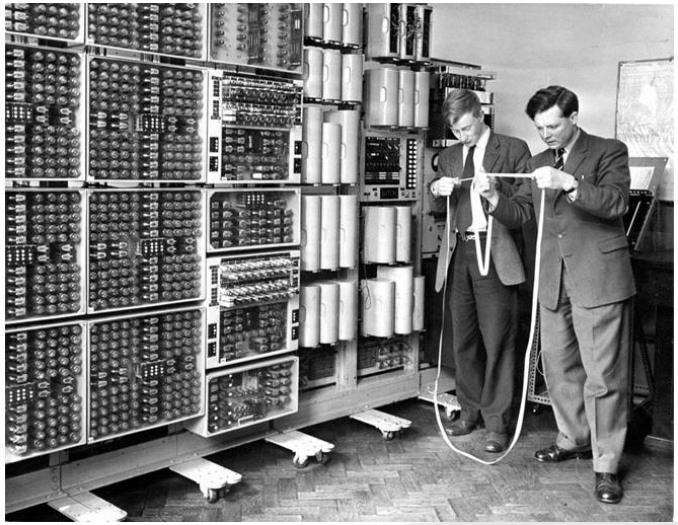
OVERVIEW

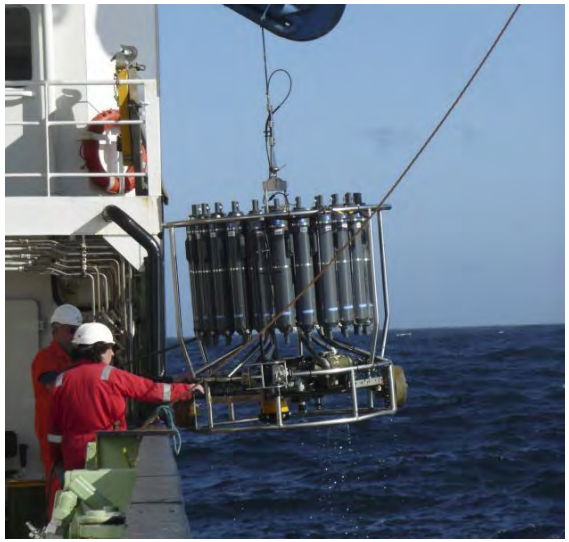
- Evolution of data centres
- What is linked data (and why should we care?)
- What are the technical challenges?
- What are the cultural / societal challenges?
- Two case studies:
    - The BODC triplestore project
    - The Celtic Seas Partnership portal

Evolution of data centres









Linked data

- RDF: W3C standard

> *"RDF is a **standard** model for data interchange on the Web. RDF has features that facilitate data merging even if the **underlying schemas differ**, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed."*

- RDF is a model, ***not*** a file format
- RDF models can be represented in various file formats (e.g. but not limited to, XML, JSON, turtle, n3)
- RDF models can be stored in various types of database

Technical details of RDF

- Individual expressions are called *triples: anything* can be described by a statement with three parts

Subject	Predicate	Object
Person	hasName	Chris
Chris	worksAt	BODC
BODC	locatedIn	Liverpool
Liverpool	hasCathedral	Anglican
Liverpool	hasCathedral	Catholic

- A database of triples is called a *triplestore*, and can be queried using SPARQL:

select ?s where {“Liverpool” hasCathedral ?s . }

The power of predicates

- Discovery of *object*s and filtering of *subject*s is much easier if known *predicates* are used
- Predicates come from external ontologies
- e.g: the *foaf* (‘friend of a friend’) ontology contains a property *name*

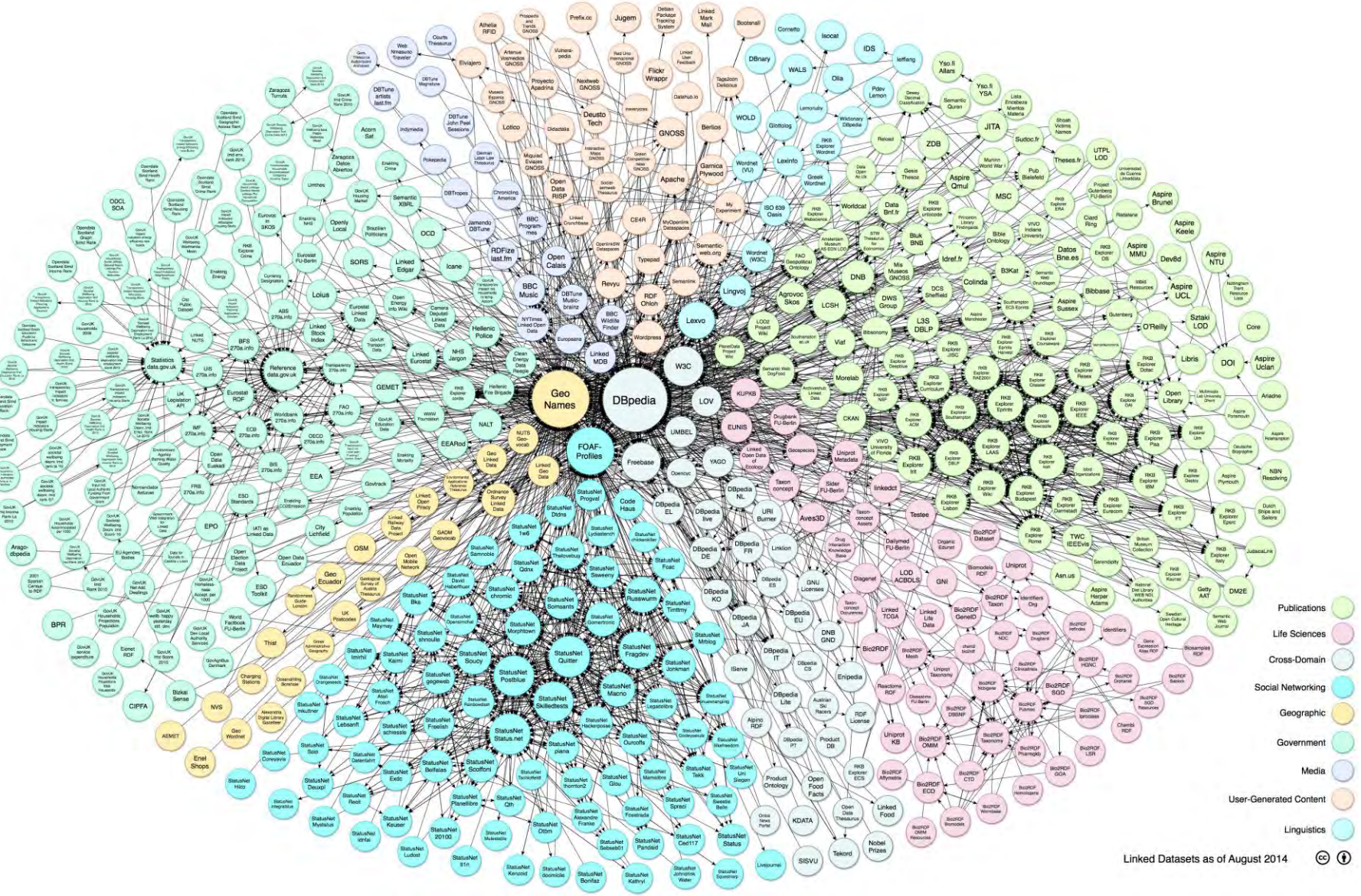
person foaf:name Chris

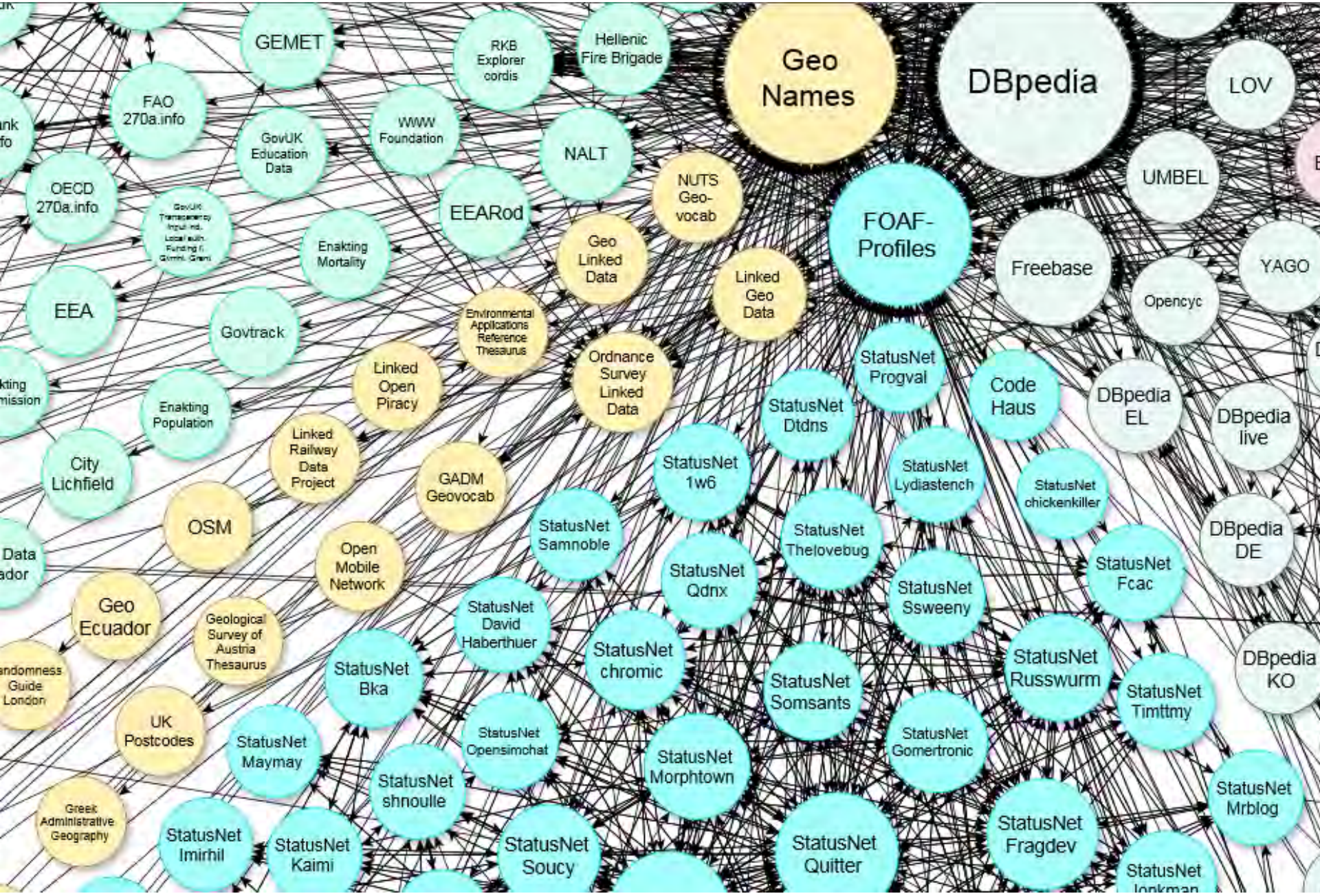
- Exploration and understanding of the dataset becomes easy!

Pros of ontologies: anyone can create them!

Cons of ontologies: anyone can create them!

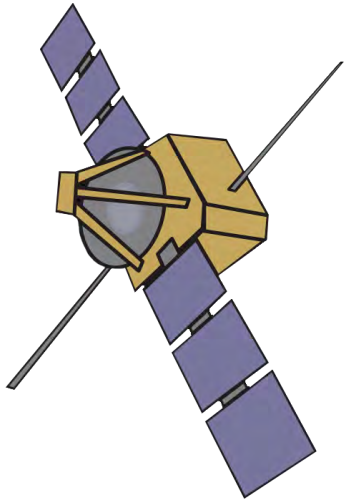
Communication in the community is vital!

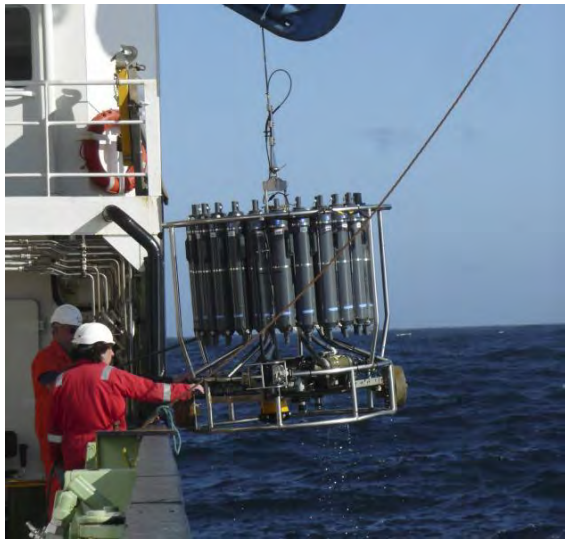



GEMET
RKB Explorer cordis
Hellenic Fire Brigade
Geo Names
DBpedia
LOV
FAO 270a.info
GovUK Education Data
WWW Foundation
NALT
UMBEL
OECD 270a.info
EEARod
NUTS Geo-vocab
FOAF-Profiles
Freebase
YAGO
Enakting Mortality
Geo Linked Data
Linked Geo Data
Opencyc
EEA
Govtrack
Environmental Applications Reference Thesaurus
Ordnance Survey Linked Data
StatusNet Progval
Code Haus
DBpedia EL
DBpedia live
Linked Open Piracy
Enakting Population
StatusNet Dtdns
Linked Railway Data Project
StatusNet 1w6
StatusNet Lydiastench
StatusNet chickenkiller
City Lichfield
GADM Geovocab
OSM
StatusNet Samnoble
StatusNet Thelovebug
DBpedia DE
Open Mobile Network
StatusNet Qdnx
StatusNet Fcac
Geo Ecuador
StatusNet Ssweeny
StatusNet David Haberthuer
Geological Survey of Austria Thesaurus
StatusNet chromic
StatusNet Russwurm
DBpedia KO
StatusNet Bka
StatusNet Somsants
StatusNet Timttmy
UK Postcodes
StatusNet Opensimchat
StatusNet Gomertronic
StatusNet Maymay
StatusNet Morphtown
StatusNet shnoulle
StatusNet Mrblog
Greek Administrative Geography
StatusNet Imirhil
StatusNet Kaimi
StatusNet Soucy
StatusNet Quitter
StatusNet Fragdev

Reflections so far…

- Triplestores are designed to be exposed to the real world (via a SPARQL endpoint)
    - i.e. all data is exposed by default
- Traditional relational databases can normally only be exposed via software (e.g. a web interface)
    - For data to be exposed via traditional databases a custom API needs to be built
- **Standards** need to be followed for triplestores to be most effective
- Triplestores are just a different flavour of database: most **standard transactions** are available





Case study 1: The BODC architecture

- The main BODC schema holds metadata about data series (physical / biological / chemical data from CTD casts / transects / float deployments)
- Currently: ~108000 data series
- Discovery of data to date has been relatively labour intensive

- Triples needed to be created for all relevant metadata
- Triplestore is updated nightly
- Relevant coding for triplestore creation (e.g. data integrity & transactions)
- SPARQL endpoint software installation & configuration: stack of Jena (Fuseki & TDB), and elda

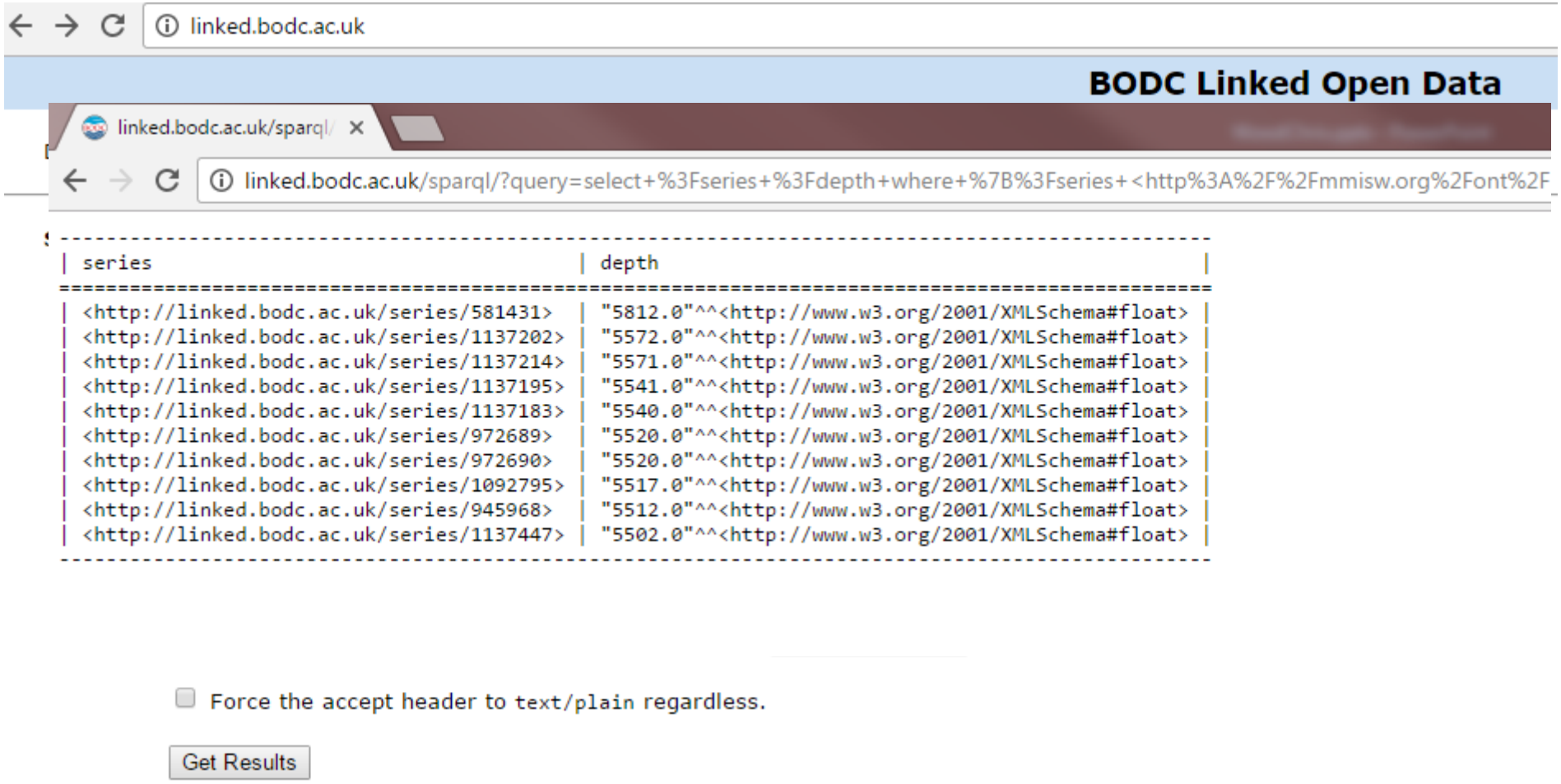
Real life triples…

<http://linked.bodc.ac.uk/series/26229><http://www.w3.org/2004/02/skos/core#notation>26229
<http://linked.bodc.ac.uk/series/26229><http://mmisw.org/ont/ioos/biological#minimumDepthInMeters>"60.0"^^<http://www.w3.org/2001/XMLSchema#float/>0789/>

```
select ?depth where {
        <http://linked.bodc.ac.uk/series/26229>
        <http://mmisw.org/ont/ioos/biological#minimumDepthInMeters>
        ?depth .
}
```

depth
"60.0"^^<http://www.w3.org/2001/XMLSchema#float>

More advanced querying

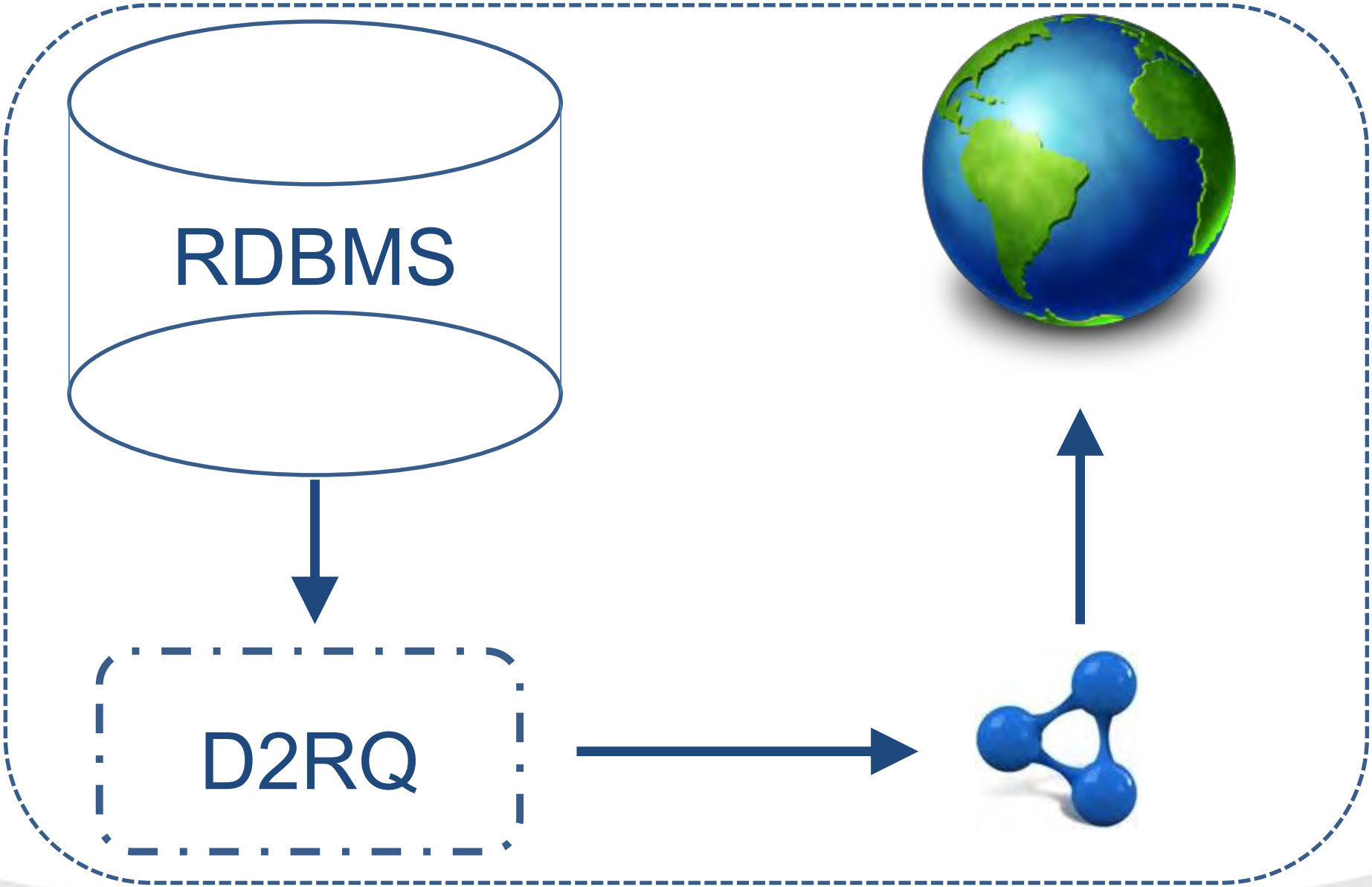


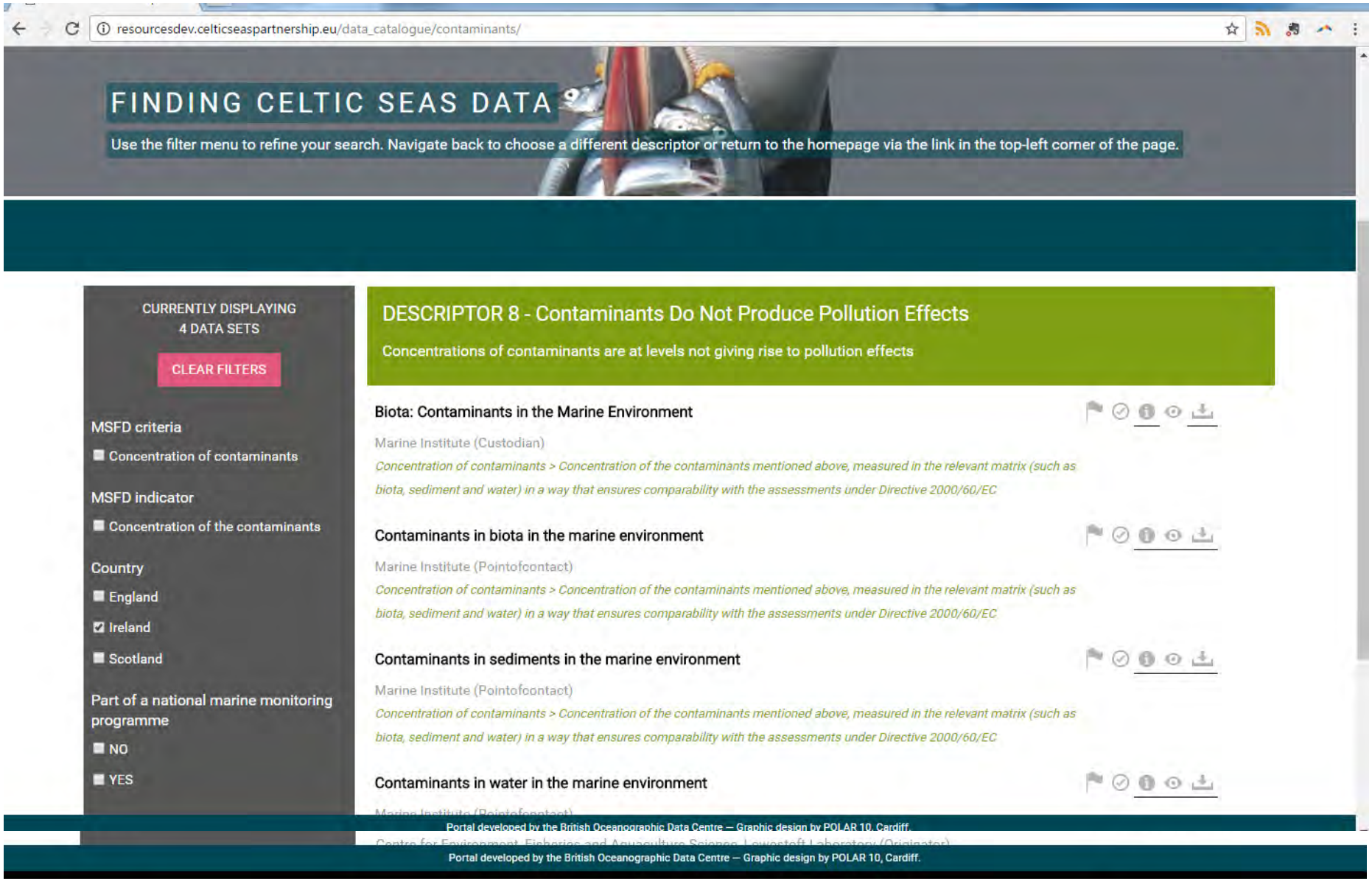
Case study 2: The Celtic Seas Partnership

- Marine Strategy Framework Directive (MSFD) requires EU member states to reach or maintain Good Environmental Status by 2020
- There are various frameworks to help implementation
- The Celtic Seas Partnership is a transboundary management framework to help implementation in the Celtic Seas

Celtic Seas Partnership Portal

- Key deliverable: a web-based portal for the marine community and interested stakeholders to expose datasets and documents that are of relevance
- Metadata stored in Oracle and exposed via a SPARQL endpoint using the D2RQ architecture
- D2RQ: dynamic mapping of an RDBMS to a psuedo-triplestore.
    - Mappings for ontologies can be applied in a configuration file
    - This allows the data to be queried as if it were stored in a native triplestore → the data is publicly exposed
    - SPARQL queries are transformed to SQL


RDBMS
D2RQ



Portal developed by the British Oceanographic Data Centre — Graphic design by POLAR 10, Cardiff.

Conclusions & take home messages

- The RDF model (and associated technologies) are **important** and **powerful tools** for **discovery** and **delivery** of data and metadata
- Implementation can have a steep learning curve, & can be time consuming
- Until access to data via triplestores is widespread, end-users need to learn a new technology
    - 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!

Acknowledgments

- Justin Buck, Rob Thomas, Alexandra Kokkinaki
- Developers of D2RQ, Jena, Fuseki, and elda (and all the people who've answered various questions!)
- Funding: EU LIFE+ (LIFE011 ENV/UK/000392) via the Celtic Seas Partnership

Questions?

chwood@bodc.ac.uk
@c_wood

http://linked.bodc.ac.uk

http://linked.bodc.ac.uk/documentation

To go live in 2 weeks:

http://resources.celticseaspartnership.eu

Documentation