Linked Data breakout session

Jonathan Yu, Mark Hedley
CSIRO and UK Met Office
25 May 2016 | EarthCube netCDF-CF Workshop

Session agenda

1. Context + Purpose
2. What is Linked Data? (aka. web of data)
3. How is it relevant for netCDF and CF conventions? Benefits?
4. Work done to date
5. Discuss its value for this community (and broader) and draft use cases
6. Draft a plan for activities to engineer prototype(s), test and validate against use cases

We're not data poor

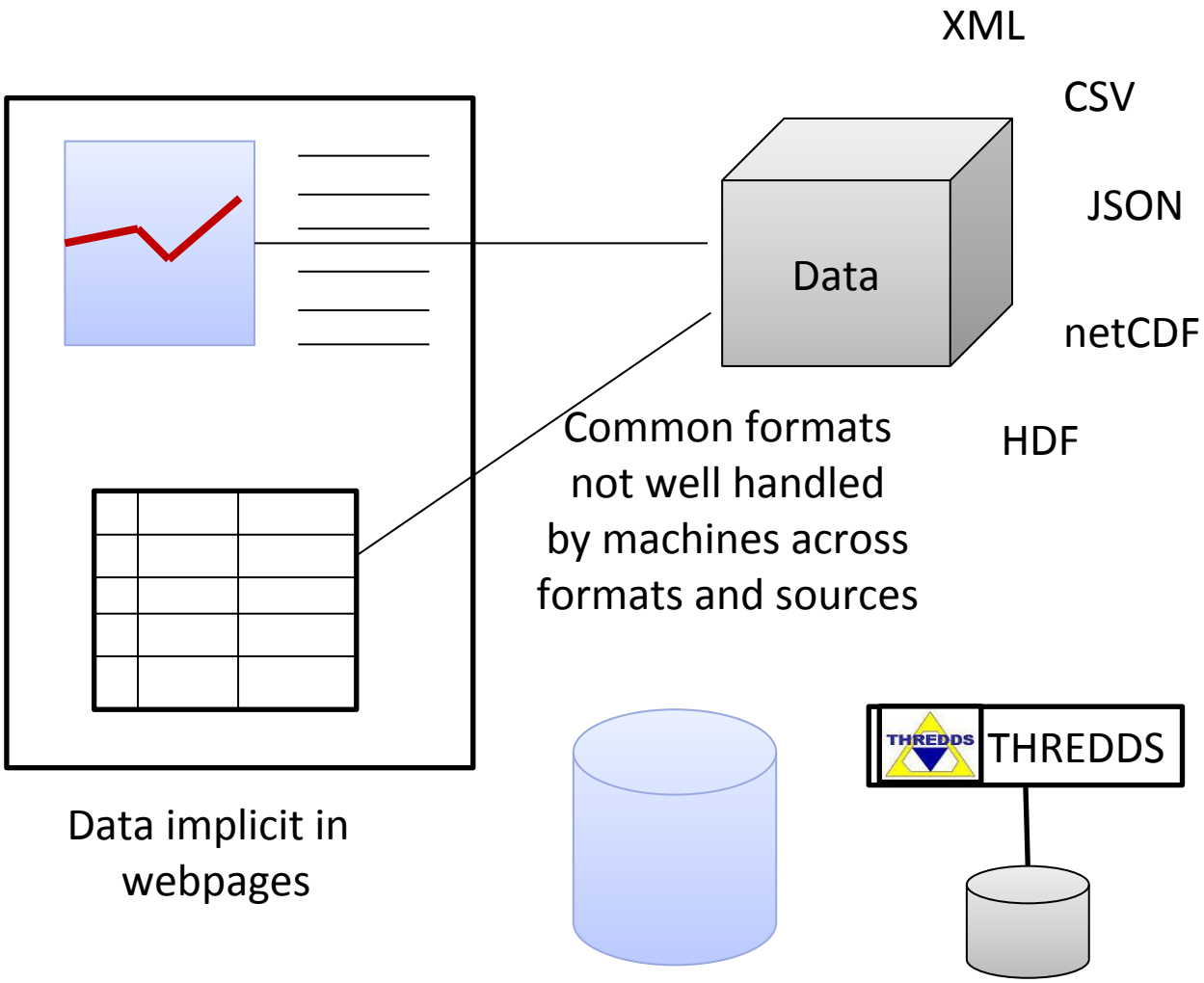
"90% of the world's data has been produced over the last two years"

Problem

Users - find the right data, access, use it, (cite it?)

Data providers – collect data, describe, publish, (update)

Problem



Discovering, accessing, parsing data held in databases and via APIs

netCDF conventions – level of agreement

Individual
Scientist/Researcher

Company

Agencies

Working group/
Committee/CoP

International/
Global bodies

Argo

SeaDataNet

ACDD

OceanSites

My.C.

Project
Teams

eReefs

RAF

CF

Streamflow
forecast

Local
Private

Internal
but shared

Cross organisation

Community wide
(intl')

Challenges with conventions

Keeping up to date/Updating them/Need something now

Suitability – which one?

Validation – have I done the right thing?

Compatibility between versions

Cost/Benefits of adopting conventions – why should I?

Tooling – help me adopt the convention? Make data useful...

netCDF not alone in these challenges

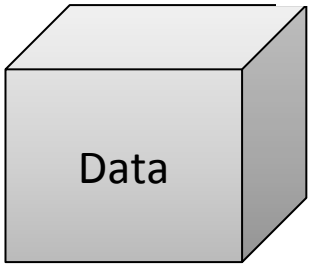
Too many

json-rpc

json-ld

geojson

Numerous bespoke json



netCDF

CF
ACDD
OceanSites
RAF
Temperature
Project/initiative
conventions

csv-au-geo

seadas

Numerous bespoke csv…

HDF

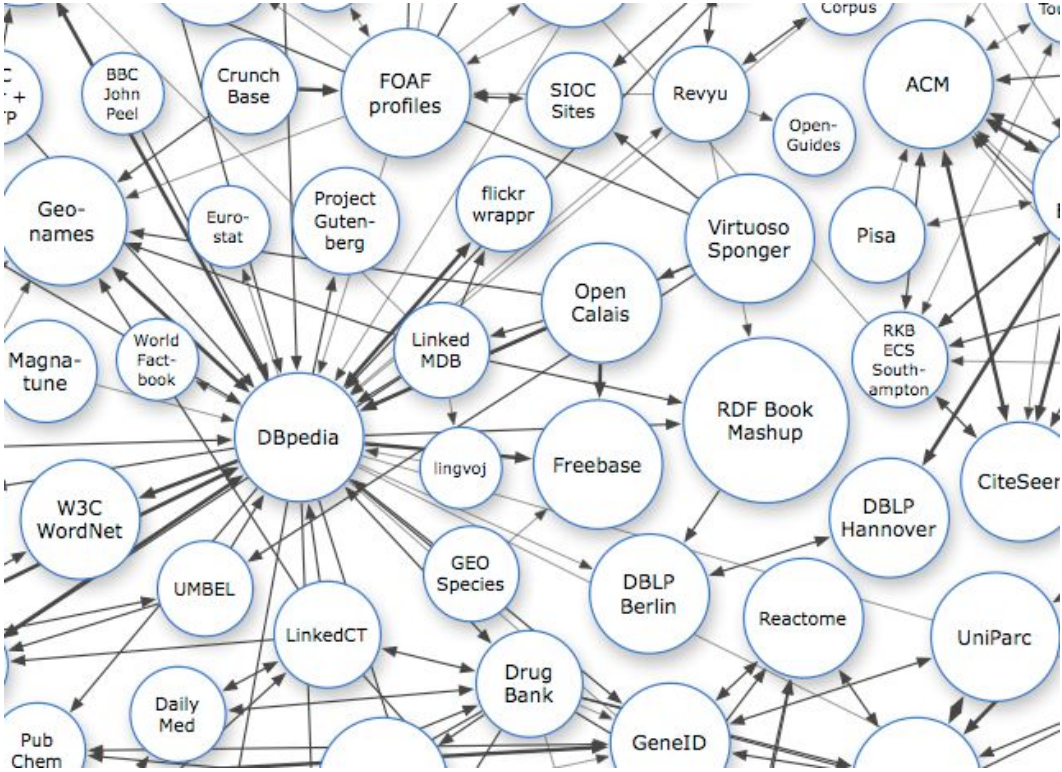
Linked Data / Web of Data

Method to connect related data and semantics using web links (HTTP URIs)

Data is self-describing

Standardised – HTTP + RDF

Applications can then lookup embedded web links to get more info, find more connections, and infer new insights from the data



Linked Data principles

1. Use URIs as names for things.

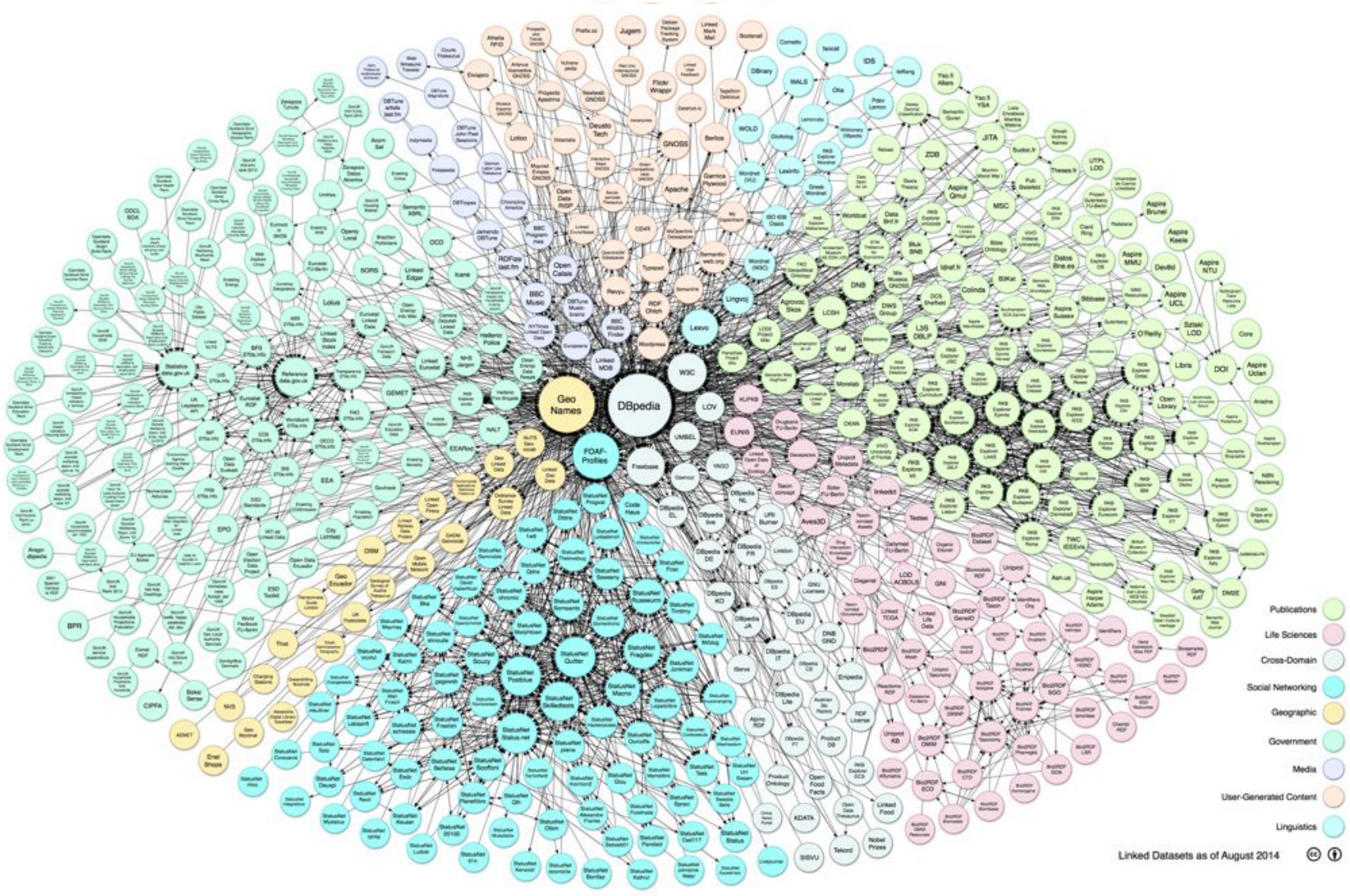
2. Use HTTP URIs, so that people can look up those names.

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.

http://linkeddata.org/
http://linkeddatabook.com

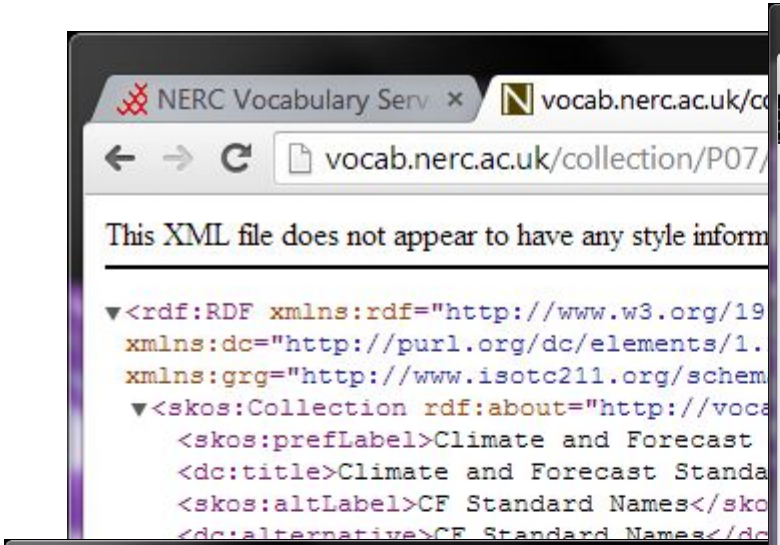
Linked Open Data Cloud

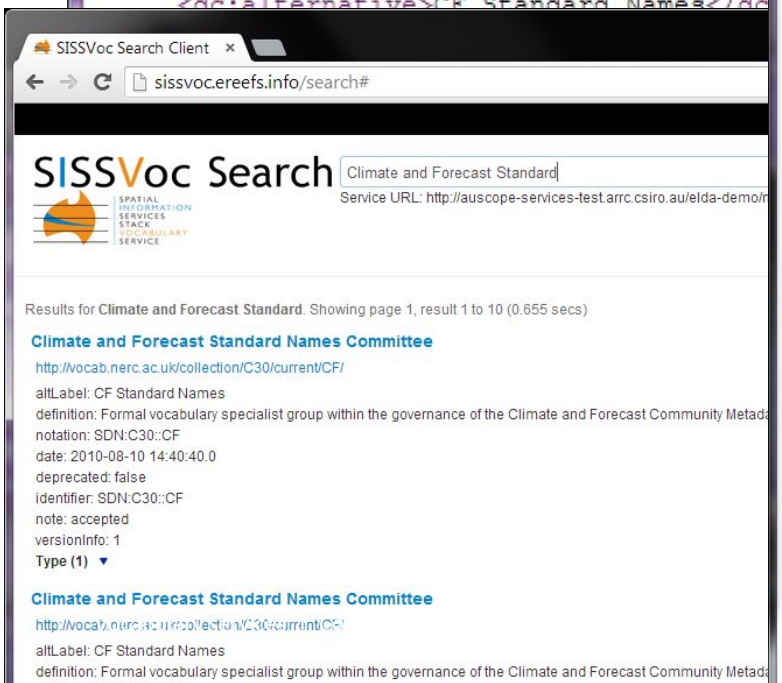


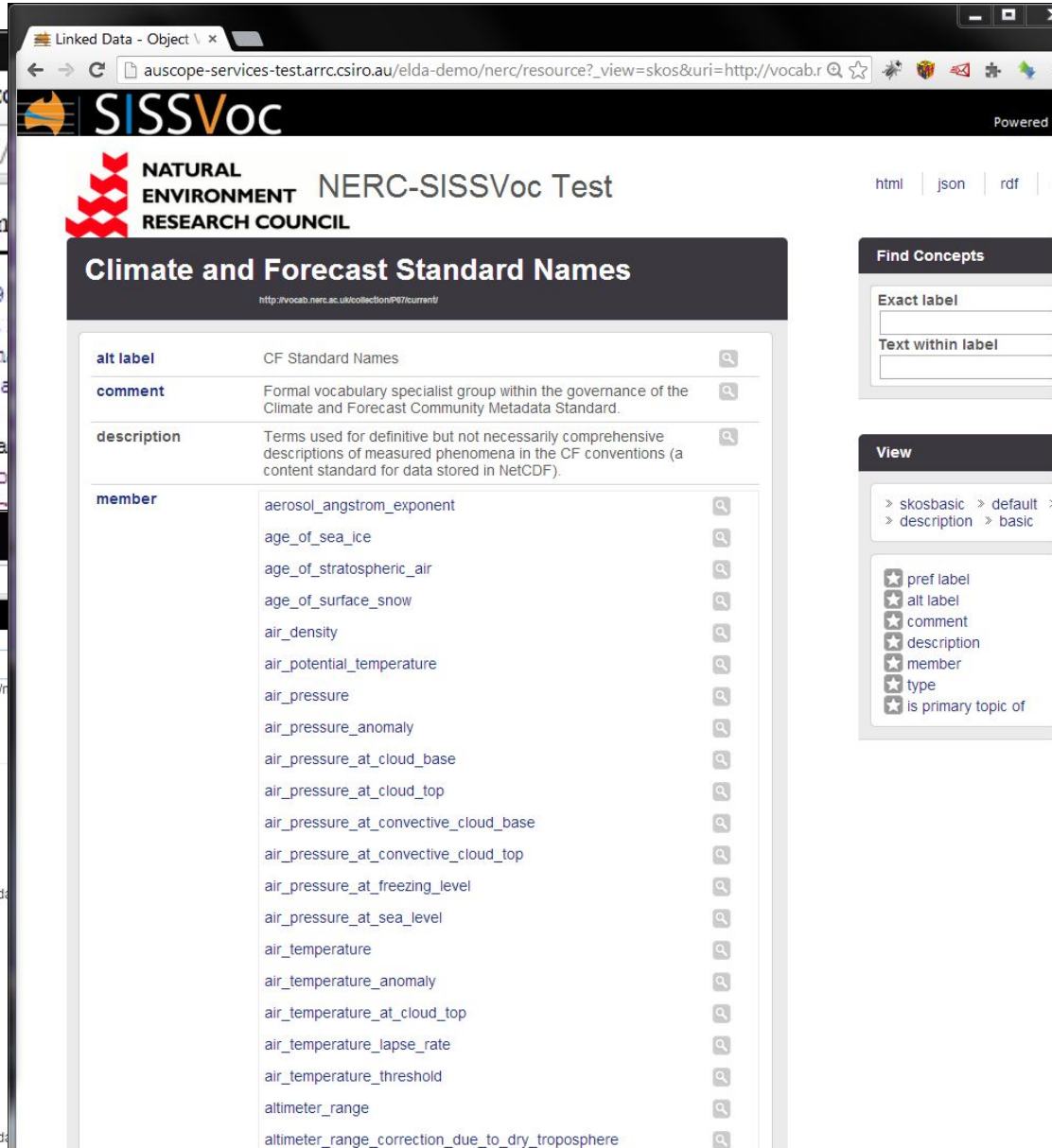
http://lod-cloud.net/

~~95 Datasets~~ ~~295 Datasets~~ 570 Datasets

Science/Domain vocabularies as Linked Data







stable

Register: observable properties

URI: http://registry.it.csiro.au/def/environment/property

A collection of observable properties. This vocabulary defines terms for observed properties originally used for groundwater, surface water and marine water quality observations. Most PropertyKinds are associated with a Species object via the objectOfInterest property or a real-world Feature via the featureOfInterest property. The sub-class of PropertyKinds that can be measured are ScaledQuantityKinds, which have appropriate units of measure (qudt:unit property). This water quality ontology re-uses the Quantities, Units, Dimensions, Data Types (QUDT) ontology which is developed by TopQuadrant and NASA.

Core metadata
Reg metadata
All properties
Download
Send comment

Contents (tree view)

- chemistry observable properties A collection of observable chemistry properties — stable
- life form observable properties A collection of observable organism properties — stable
- major element observable properties A collection of observable major elements as specified by Australia... — stable
- minor or trace element observable properties A collection of observable minor trace elements as specified by Aus... — stable
- nutrient observable properties A collection of observable nutrients — stable
  - ammonia and ammonium as NH4 concentration ammonia and ammonium as NH4 concentration — stable
  - ammonia and ammonium as N concentration ammonia and ammonium as N concentration — stable
  - ammonia and ammonium concentration ammonia and ammonium concentration — stable
  - ammonia as N concentration ammonia as N concentration — stable
  - ammonia concentration ammonia concentration — stable
  - ammonium (NH4) concentration ammonium (NH4) concentration — stable
  - nitrate N concentration nitrate N concentration — stable
  - nitrate and nitrite N concentration nitrate and nitrite N concentration — stable
  - nitrite N concentration nitrite N concentration — stable
  - nitrogen concentration nitrogen concentration — stable
  - nutrient concentration nutrient concentration — stable

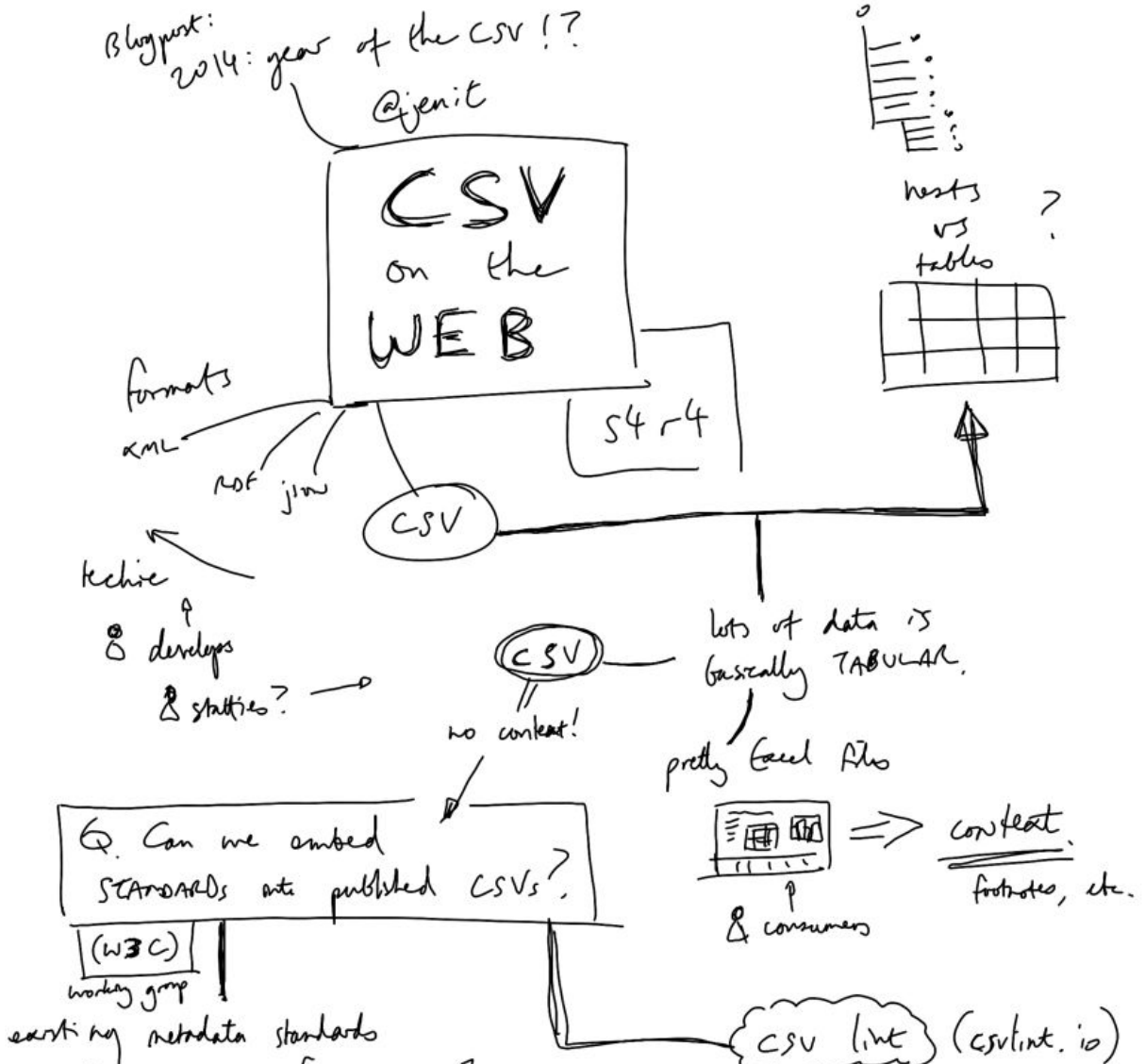
JSON-LD

```
{
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse":
"http://dbpedia.org/resource/Cynthia_Lennon"
}
```

JSON-LD Decorators

```
{
  "@context": "http://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

CSV-on-the-web



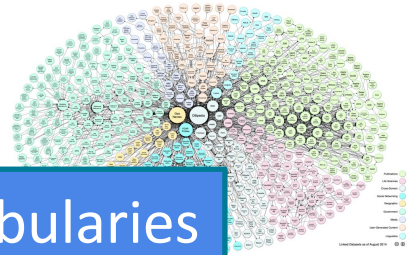
Linked data for CSV tabular data

Add context to tables via metadata file

Use cases: documentation, validation, transformation (e.g. RDF, JSON, XML…), annotate semantics, enhanced discovery

How is all this Linked Data stuff relevant for netCDF and CF conventions?

Have all the building blocks to enhance netCDF(-CF)!



Linked Data

JSON-LD

CSV-on-the-web

Vocabularies as Linked Data

General approach
Principles
Tools

Patterns for linkifying common formats

Content for annotating metadata
Tools to create, manage, publish vocabs

'*linkify*' netCDF!

5 stars of Linked Open Data

★	make your stuff available on the web (whatever format)
★★	make it available as structured data (e.g. excel instead of image scan of a table)
★★★	non-proprietary format (e.g. csv instead of excel)
★★★★	use URLs to identify things, so that people can point at your stuff
★★★★★	link your data to other people's data to provide context

netCDF(-CF) online

netCDF(-CF) online
+ additional context - links to standardised vocab URLs e.g. NERC vocabs, QUDT, Observable Properties vocabs, CF standard name URLs online, DBPedia URLs

http://5stardata.info
http://inkdroid.org/2010/06/04/the-5-stars-of-open-linked-data/

Benefits of linkifying netCDF(-CF)

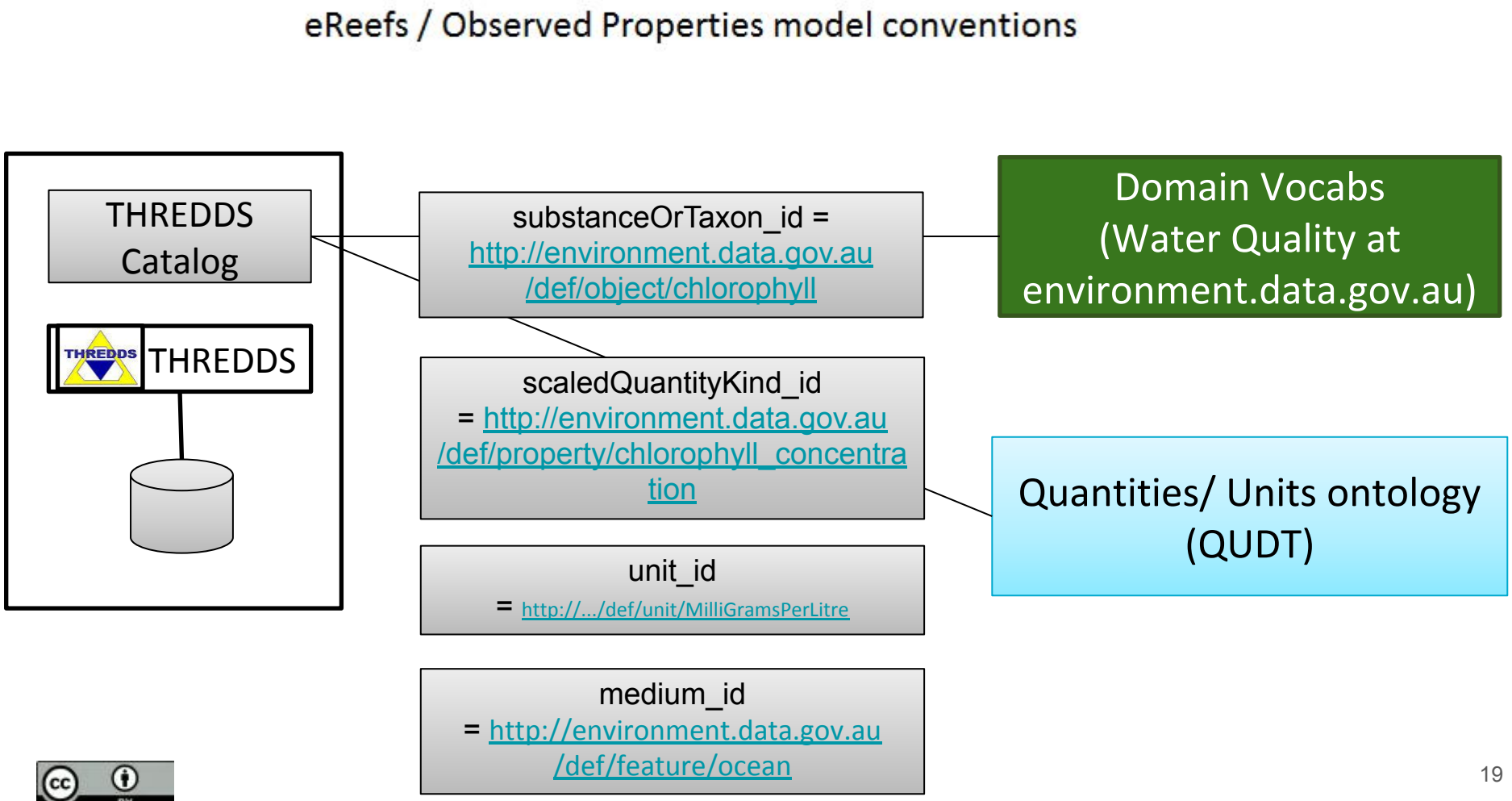
1. Improve discoverability and reduce ambiguity
   - link to vocabularies to add context
   - easier to support community profiles and validation
2. Improve data integration
3. Potential to translate netCDF to other formats
4. Potentially ease metadata generation
5. Easier to build applications

Increased discoverability -> More usage -> Greater impact

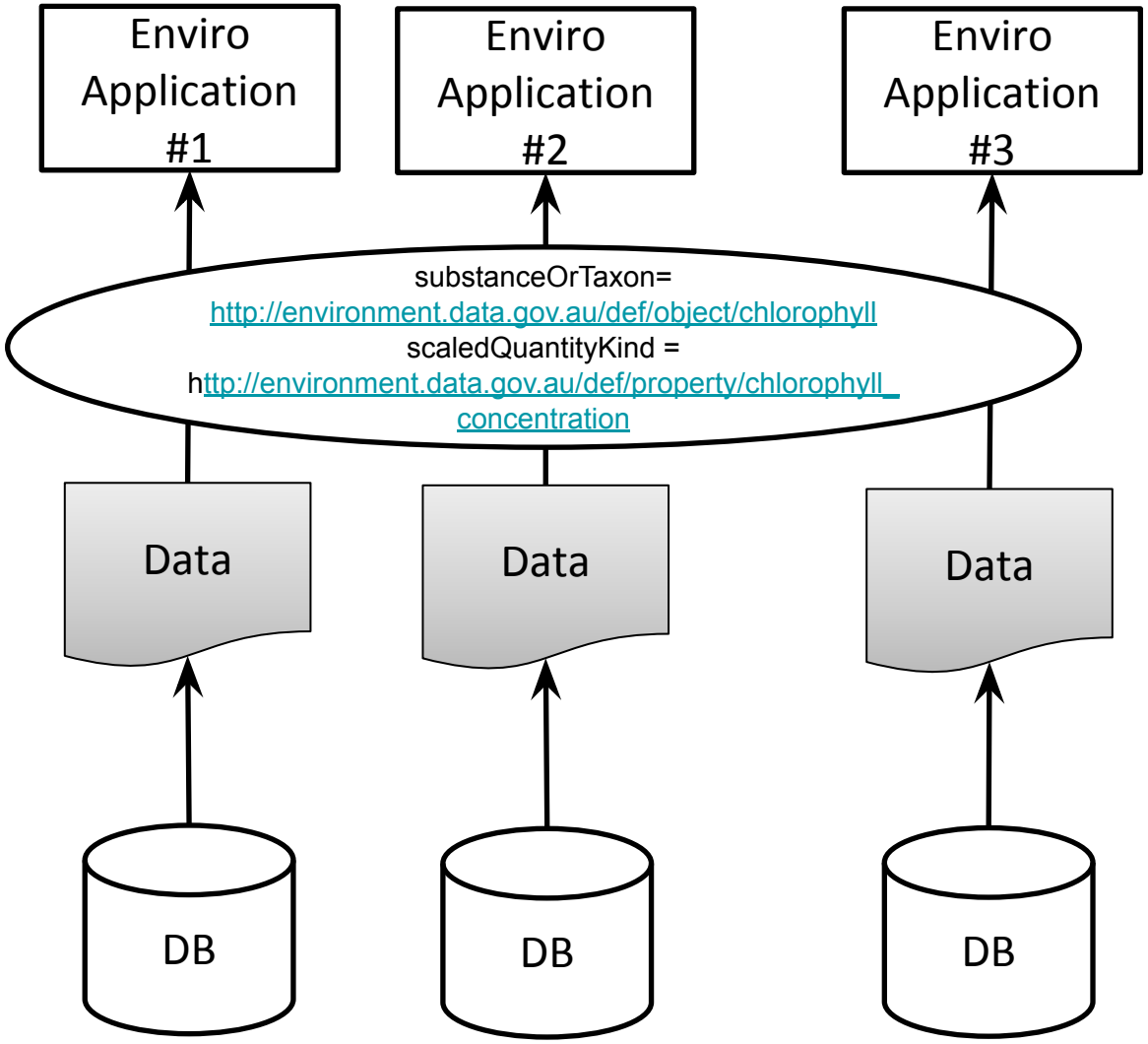
Current examples / thought exercises

1. Injecting vocabulary URIs in netCDF headers using special attr
    - eReefs/Observable property model convention
    - SeaDataNet
    - ‘Smuggling’ semantics into flags
2. netCDF-LD
3. Binary-array-LD (BALD)

#1: Injecting vocabulary URIs in netCDF headers using special attributes



Allows for harmonised access to those binding to conventions



Yu, J., Simons, B. A., Car, N. J., & Cox, S. J. (2014). Enhancing water quality data service discovery and access using standard vocabularies. Hydroinformatics conference. New York.

Others examples of injecting vocab URIs

1. SeaDataNet CF Profile

- Specifies minimum info content as attributes: `sdn_parameter_urn , sdn_parameter_name, sdn_uom_urn, sdn_uom_name`
- Binding to NERC P01 (parameters), P06 (units) vocabulary collections

2. netCDF-U - uncertainty URIs (Bigagli & Nativi 2013)

- Use of "`ref`" attribute for uncertainty concept URI, further references using ancillary_variables

3. Use of flags to encode URIs for categorical data

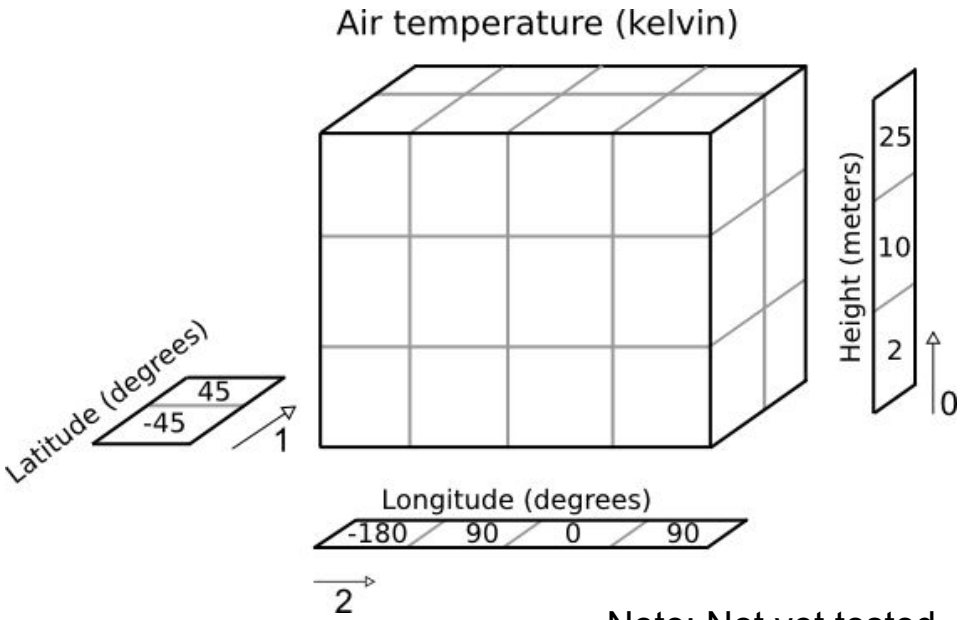
- Use of `flag_namespace` attribute to give vocab URI prefix to values in `flag_meanings`

It's already happening out in the community!
Should we co-ordinate how we do this?

#2: netCDF-LD

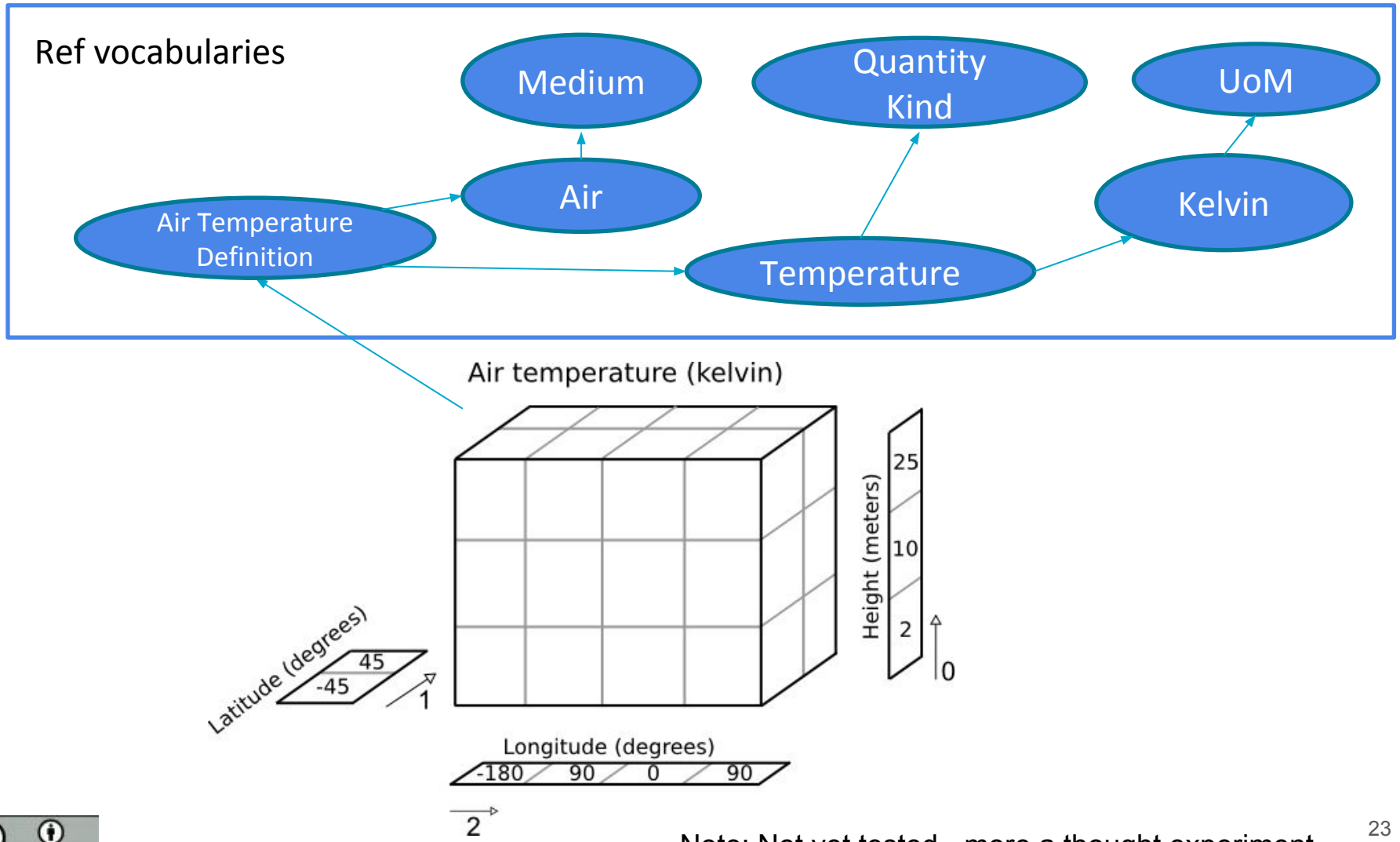
… Apply JSON-LD pattern

‘Context’ boilerplates



Note: Not yet tested - more a thought experiment...

netCDF-LD: Linkifying netCDF



Note: Not yet tested - more a thought experiment...

netCDF-LD: Global Attributes

```
:context-id = "http://foo.bar/baz";
:context-a = "http://www.w3.org/1999/02
                /22-rdf-syntax-ns#type";
:context-ref = "http://www.w3.org/1999/02
                  /22-rdf-syntax-ns#resource";
```

```
:context-datatype = "http://www.w3.org/1999/02
                       /22-rdf-syntax-ns#datatype";
```

```
:context-vocab = "http://def.seegrid.csiro.au
              /isotc211/iso19156/2011/observation#";
```


Note: Not yet tested - more a thought experiment...

Assigning URIs to variable level attributes

```
z:units = "meters";
z:units_ref = "http://qudt.org/vocab/unit#Meter";
z:a =
"http://environment.data.gov.au/def/op#quantityKind";
z:dcPartOf = "http://foo.bar/linked_netCDF_example";
z:valid_range = 0., 5000.;
```


Note: Not yet tested - more a thought experiment...

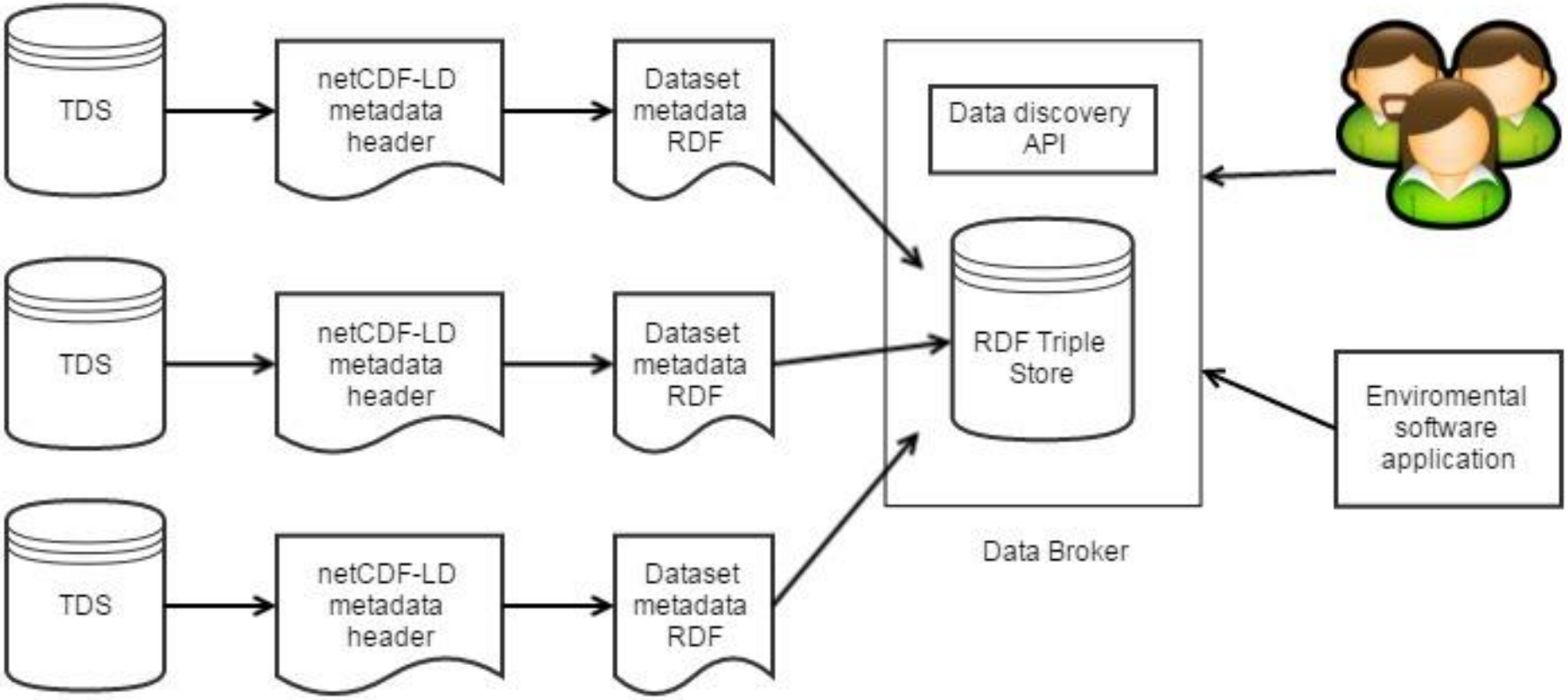
netCDF-LD to RDF

```
@prefix unit: <http://qudt.org/vocab/unit#> .
@prefix qudt: <http://qudt.org/1.1/schema/qudt#> .
@prefix op: <http://environment.data.gov.au/def/op#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

_:z qudt:unit unit:Meter;
   a op:ScaledQuantityKind;
   dcterms:isPartOf <http://foo.bar/linked_netCDF_example>.
```


Note: Not yet tested - more a thought experiment...

Use of netCDF-LD to support Data Discovery



Yu, J., Car, N. J., Leadbetter, A., Simons, B. A., & Cox, S. J. (2015). Towards linked data conventions for delivery of environmental data using netCDF. In *Environmental Software Systems. Infrastructures, Services and Applications* (pp. 102-112). Springer International Publishing. https://dx.doi.org/10.1007/978-3-319-15994-2_9

#3: BALD (Binary Array LD)

Linked Data Conventions for netCDF, HDF, …

https://github.com/binary-array-ld/bald

http://binary-array-ld.net/latest

http://binary-array-ld.net/latest?classView=true

https://github.com/binary-array-ld/bald/issues

#3: BALD (Binary Array LD)

#2 prefix identification

#3 prefix container

validation - do URIs resolve? - are array references consistent?

Aim: create an RDF graph of the metadata within a file (collection of files)

Identifying a file is an interesting question: OpenDAP presents an interesting angle on this

Discussion: Value for this community (and broader) and draft use cases

Does the community see value in a Linked Data profile/convention for netCDF? Part of netCDF-CF?

- General consensus
- Acceptable to reference external resources?

What are the use cases?

Spend some time drafting use cases

Draft use cases

Discovery

Use

- Machine readable content
- DOI for a dataset - create links for a URI

Encoding

- help data providers reference external sources
- reference features (geoms, stations, platform, instrument, sensor)

Compliance checking

- help data providers check conventions bound - e.g. practice of 1 or more conventions

cf__standard_name = cf__air_temp

ereefs__quantity = wq__

Options:

cf namespace default mixed with other conventions

standard_name = “xxx”

acdd__

Challenges

- Wary of introducing XML-ism into netCDF
  - Perhaps have default namespaces for each convention
  - People like netCDF because there's no namespace
  - not as elegant
  - alternatives for specifying LD - using '@' to prefix incl. standard_name?

Governance of prefix namespace

- falls under unidata?
- governance of other namespace

Persistence of URIs -

- injecting fragility
- already exists - references to convention documentation
- doi?

Principles

- Doesn’t break classic CF - Backwards compatible
- Prefer elegance of classic CF
- Forward looking approach

Benefits

- able to pull in content from external sources e.g. labels, features/geometries?

Draft a plan for activities to engineer prototype(s), test and validate against use cases

1. What would we need to make this work? Examples, qualified use cases from existing projects/data, endorsement?

- principles (see prev slide)
- project use cases
- endorsement - CF/ACDD/CMIP (conventions level) or netCDF (at an API level)?

2. What would an activity look like?

- github
- test BALD software on github

3. How do we organise it? Next steps and timeframes

- 6 months, propose monthly telecon in this period
- github

Participation

- contribute use cases, test cases from projects
  - e.g. features, grid specs, ship track
  - netcdf groups?
- monthly telecons
- github code and issue tracker

Thank you

UK Met Office
Mark Hedley

[insert title]

t [phone?]
e mark.hedley@metoffice.gov.uk

CSIRO Land and Water
Jonathan Yu

Research Software Engineer

t +61 3 9252 6440
e jonathan.yu@csiro.au

Linked Data principles

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.

http://linkeddata.org/
http://linkeddatabook.com

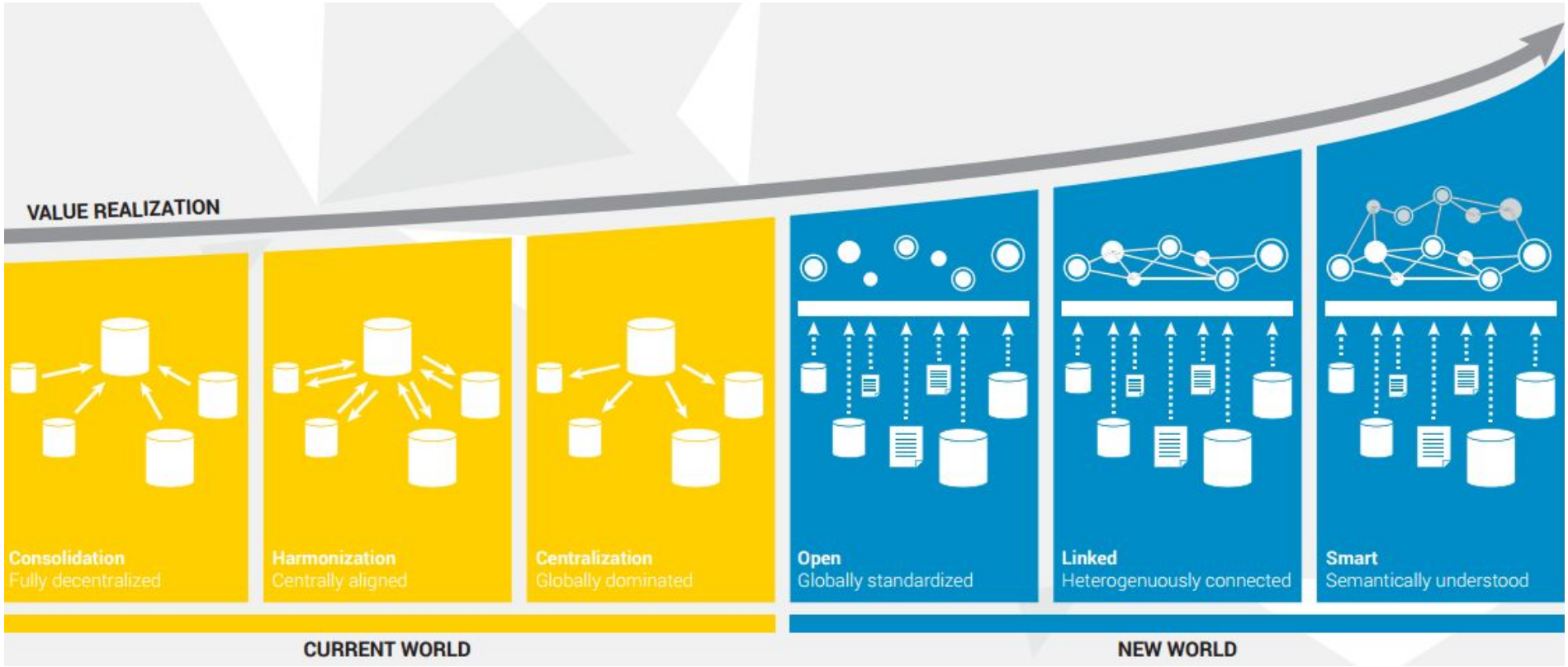
5 stars of Linked Open Data

★	make your stuff available on the web (whatever format)
★★	make it available as structured data (e.g. excel instead of image scan of a table)
★★★	non-proprietary format (e.g. csv instead of excel)
★★★★	use URLs to identify things, so that people can point at your stuff
★★★★★	link your data to other people's data to provide context

netCDF(-CF) online

netCDF(-CF) online
+ additional context - links to standardised vocab URLs e.g. NERC vocabs, QUDT, Observable Properties vocabs, CF standard name URLs online, DBPedia URLs

http://5stardata.info
http://inkdroid.org/2010/06/04/the-5-stars-of-open-linked-data/



http://www.fireflyim.com/docs/smart_enterprise_data.pdf

Break up components in standard_name into multiple attributes – ref (Yu et al. 2014)

```
float Nap_MIM(time, latitude, longitude) ;
   Nap_MIM:_FillValue = -999.f ;
   Nap_MIM:long_name = "TSS, MIM SVDC on Rrs" ;
   Nap_MIM:units = "mg/L" ;
   Nap_MIM:valid_min = 0.01209607f ;
   Nap_MIM:valid_max = 226.9626f ;
   Nap_MIM:scaledQuantityKind_id
               =
"http://environment.data.gov.au/water/quality/def/property/solids-total_suspend
ed" ;
   Nap_MIM:unit_id =
"http://environment.data.gov.au/water/quality/def/unit/MilliGramsPerLitre" ;
   Nap_MIM:substanceOrTaxon_id =
"http://environment.data.gov.au/water/quality/def/object/solids";
   Nap_MIM:medium_id =
"http://environment.data.gov.au/water/quality/def/object/ocean"
   Nap_MIM:procedure_id = "http://data.ereefs.org.au/ocean-colour/MIM_SVDC_RRS"
;
```

JSON-LD

```
{
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse":
"http://dbpedia.org/resource/Cynthia_Lennon"
}
```

JSON-LD Decorators

```
{
  "@context": "http://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

JSON-LD and Semantic Web

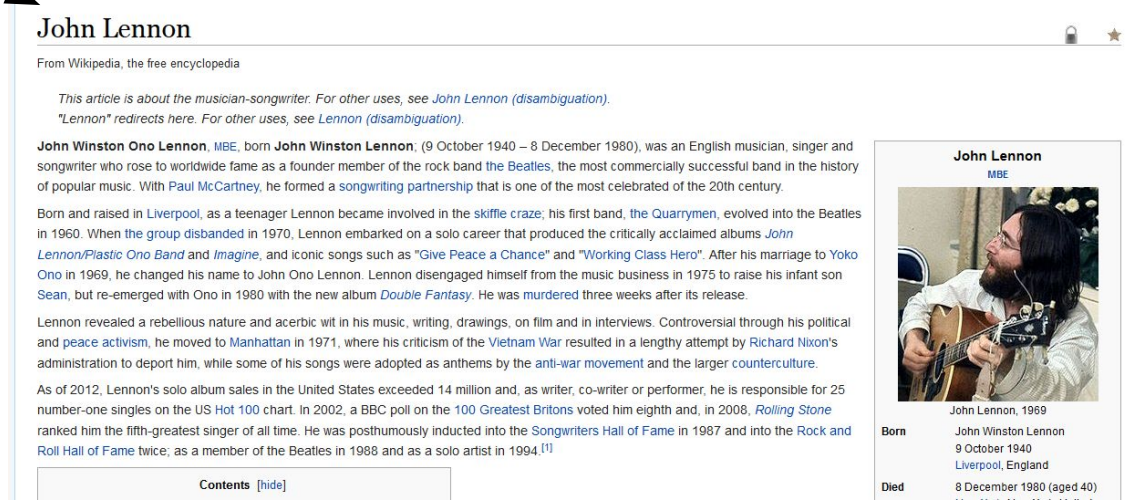
```
{
  "@context": "http://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_Lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

http://dbpedia.org/resource/Cynthia_Lennon

http://dbpedia.org/resource/John_Lennon

"John Lennon"

1940-10-09



John Lennon

From Wikipedia, the free encyclopedia

This article is about the musician-songwriter. For other uses, see John Lennon (disambiguation).
"Lennon" redirects here. For other uses, see Lennon (disambiguation).

John Winston Ono Lennon, MBE, born **John Winston Lennon**; (9 October 1940 – 8 December 1980), was an English musician, singer and songwriter who rose to worldwide fame as a founder member of the rock band the Beatles, the most commercially successful band in the history of popular music. With Paul McCartney, he formed a songwriting partnership that is one of the most celebrated of the 20th century.

Born and raised in Liverpool, as a teenager Lennon became involved in the skiffle craze; his first band, the Quarrymen, evolved into the Beatles in 1960. When the group disbanded in 1970, Lennon embarked on a solo career that produced the critically acclaimed albums *John Lennon/Plastic Ono Band* and *Imagine*, and iconic songs such as "Give Peace a Chance" and "Working Class Hero". After his marriage to Yoko Ono in 1969, he changed his name to John Ono Lennon. Lennon disengaged himself from the music business in 1975 to raise his infant son Sean, but re-emerged with Ono in 1980 with the new album *Double Fantasy*. He was murdered three weeks after its release.

Lennon revealed a rebellious nature and acerbic wit in his music, writing, drawings, on film and in interviews. Controversial through his political and peace activism, he moved to Manhattan in 1971, where his criticism of the Vietnam War resulted in a lengthy attempt by Richard Nixon's administration to deport him, while some of his songs were adopted as anthems by the anti-war movement and the larger counterculture.

As of 2012, Lennon's solo album sales in the United States exceeded 14 million and, as writer, co-writer or performer, he is responsible for 25 number-one singles on the US Hot 100 chart. In 2002, a BBC poll on the 100 Greatest Britons voted him eighth and, in 2008, *Rolling Stone* ranked him the fifth-greatest singer of all time. He was posthumously inducted into the Songwriters Hall of Fame in 1987 and into the Rock and Roll Hall of Fame twice; as a member of the Beatles in 1988 and as a solo artist in 1994.[1]

Contents [hide]

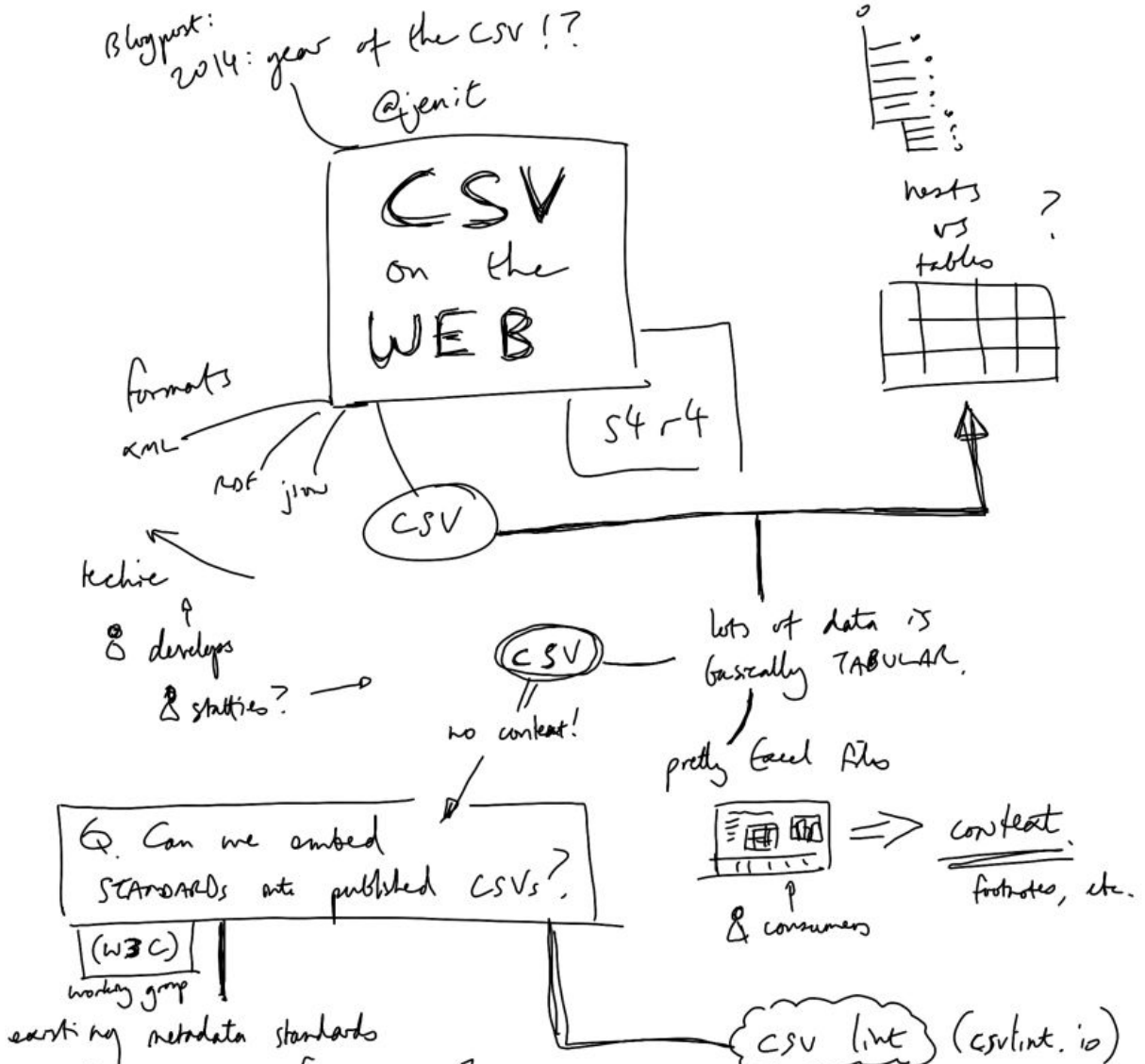
John Lennon
MBE

John Lennon, 1969

Born	John Winston Lennon 9 October 1940 Liverpool, England
Died	8 December 1980 (aged 40)

http://json-ld.org/

CSV-on-the-web

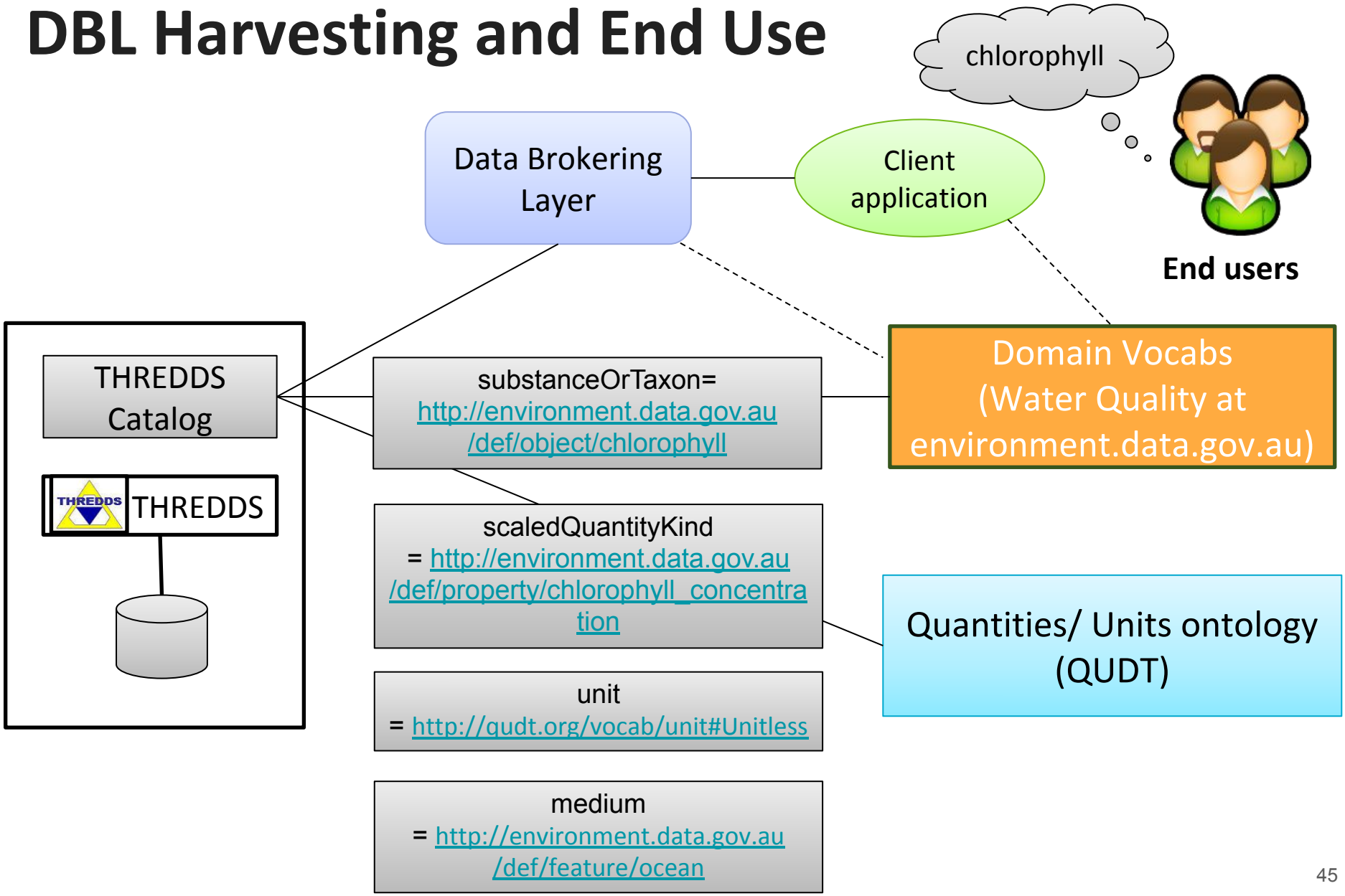


Linked data for CSV tabular data

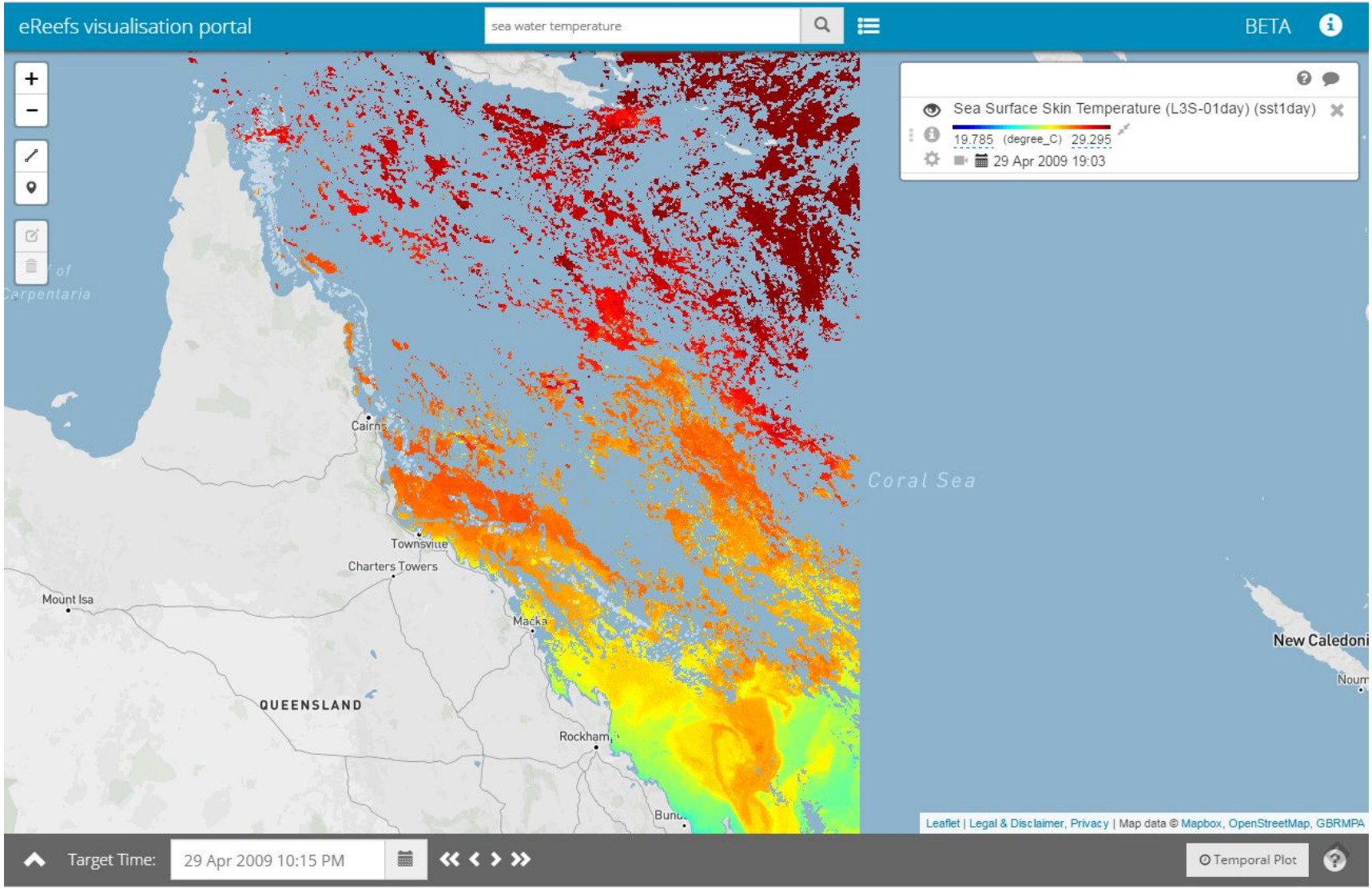
Add context to tables via metadata file

Use cases: documentation, validation, transformation (e.g. RDF, JSON, XML…), annotate semantics, enhanced discovery

DBL Harvesting and End Use



eReefs visualisation portal



http://portal.ereefs.info/

5 results found for

sea water temperature

http://environment.data.gov.au/def/property/sea_water_temperature

Sea Surface Skin Temperature (L3S-01day) - "sst1day"

SaltWater

sea water temperature | **BOM ReefTemp2 daily SST** | **"sst1day"** | OceanRegion | SeaWater

Data label: Sea Surface Skin Temperature (L3S-01day)

Data description: Sea Surface Skin Temperature (L3S-01day)

Dataset: BOM ReefTemp2 daily SST

Data owner: Bureau of Metereology Climate Water Division

Data node: Bureau of Meteorology node for eReefs

Service: Bureau of Meteorology eReefs Marine Water Quality Dashboard Services

Quantity: sea water temperature

Feature: OceanRegion

Medium: SeaWater

Date range: 19 Dec 2001 to 16 Mar 2016

Available web services:

OPENDAP	http://ereeftds.bom.gov.au/ereefs/tds/dodsC/ereefs/RT2_gridAgg_P1D_SST
WMS	http://ereeftds.bom.gov.au/ereefs/tds/wms/ereefs/RT2_gridAgg_P1D_SST
WCS	http://ereeftds.bom.gov.au/ereefs/tds/wcs/ereefs/RT2_gridAgg_P1D_SST
NetcdfSubset	http://ereeftds.bom.gov.au/ereefs/tds/ncss/grid/ereefs/RT2_gridAgg_P1D_SST

Full DBL result for this data product

Summary of approach #1: injecting vocab URIs

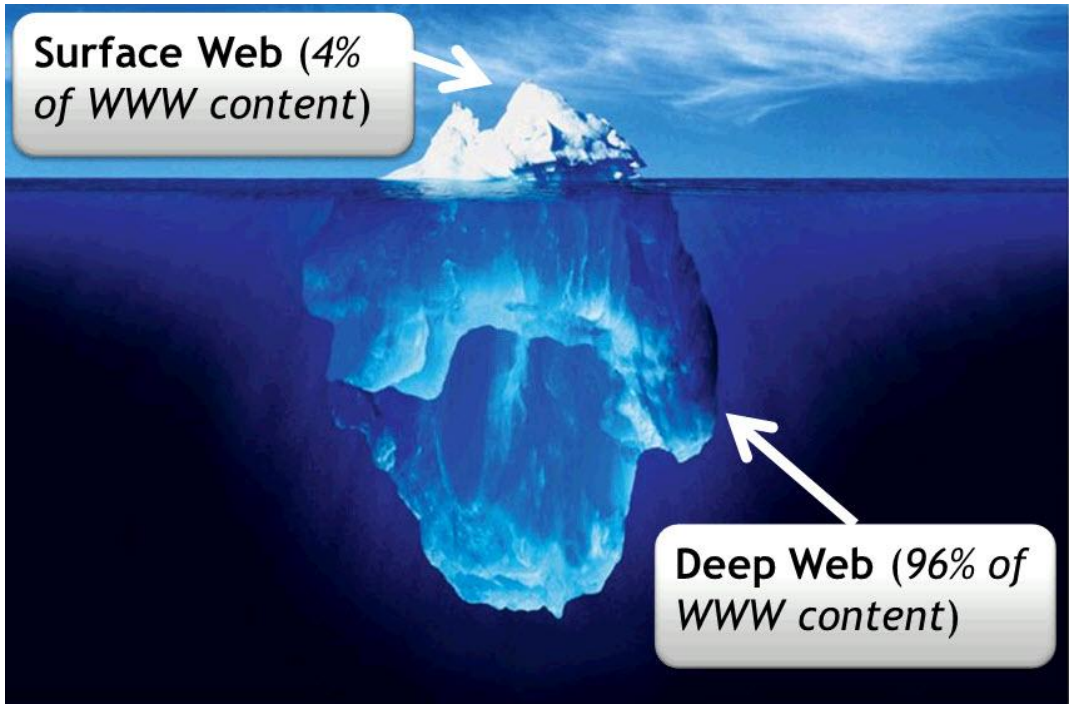
Various communities are developing approaches to add context and semantics to complement netCDF-CF.

Clearly, there are use cases for adding more semantics to current netCDF-CF metadata specifications.

Approaches are currently fragmented.

Would benefit from agreement and common approaches to profile.

Deep (or invisible) web



400-500x more public information than the Surface Web

1000-2000x greater quality than Surface Web

95% Deep Web is publicly accessible

Deep Web tend to be narrower, with deeper content

netCDF (scientific) data part of this Deep Web?