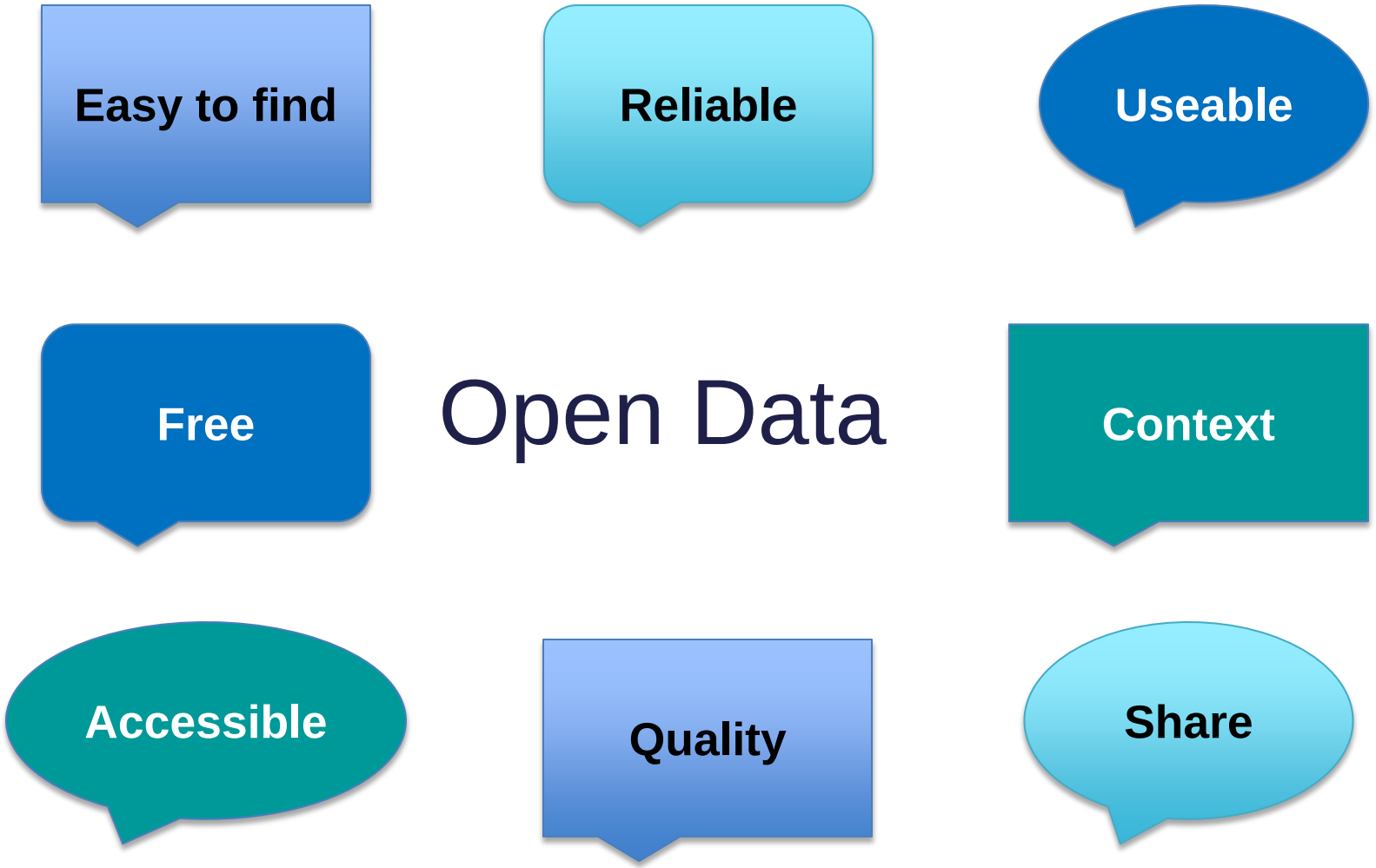
www.bodc.ac.uk

A Changing World of Data

KATIE GOWERS, BRITISH OCEANOGRAPHIC DATA CENTRE

BANGOR UNIVERSITY OPEN ACCESS WEEK, OCTOBER 2015


Easy to find
Reliable
Useable
Free
Open Data
Context
Accessible
Quality
Share

Open Data

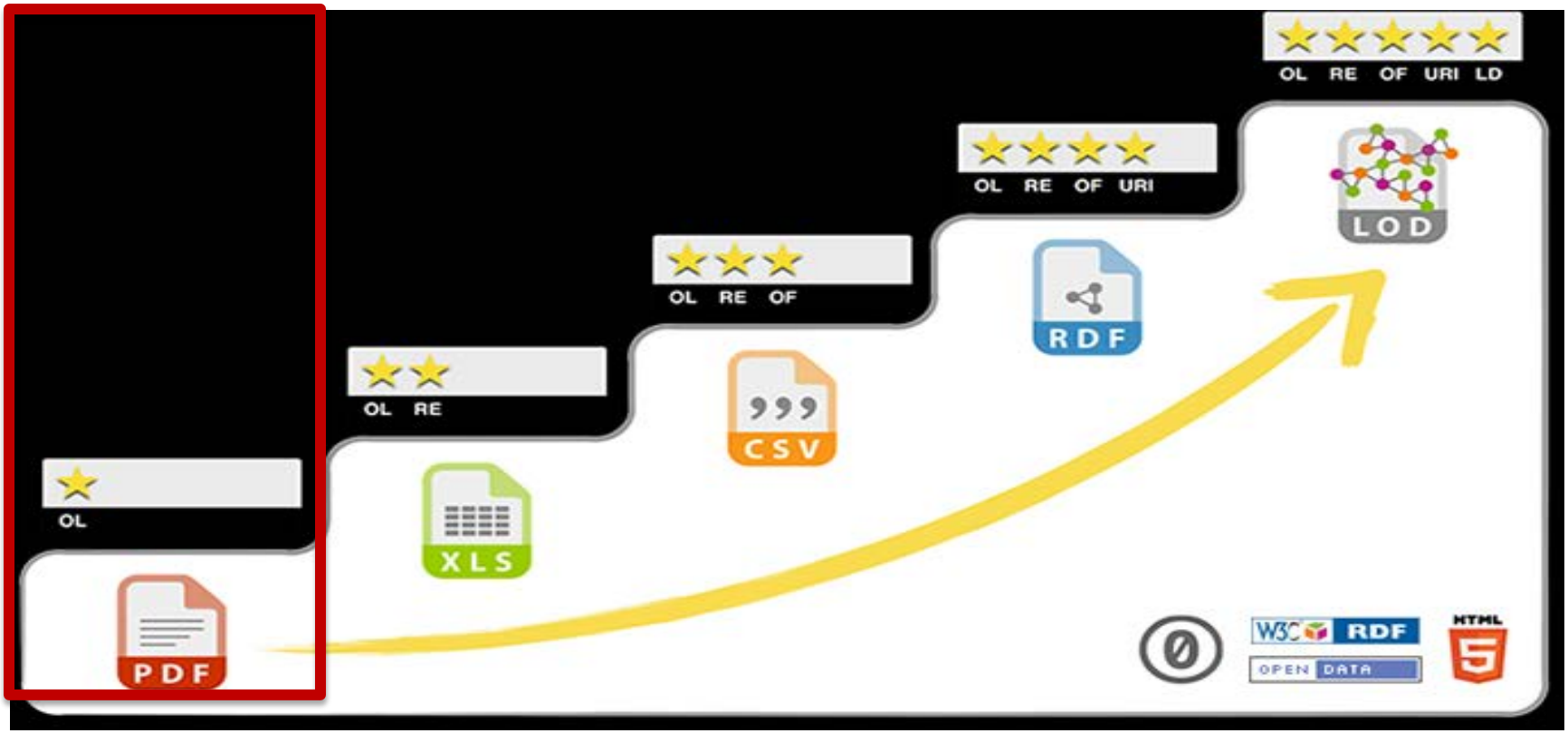
The Open Data Institute states that good open data:

- can be linked to, so that it can be easily shared and talked about
- is available in a standard, structured format, so that it can be easily processed
- has guaranteed availability and consistency over time, so that others can rely on it
- is traceable, through any processing, right back to where it originates, so others can work out whether to trust it

http://theodi.org/guides/what-open-data

★★★★★ 5 Star Open Data ★★★★★

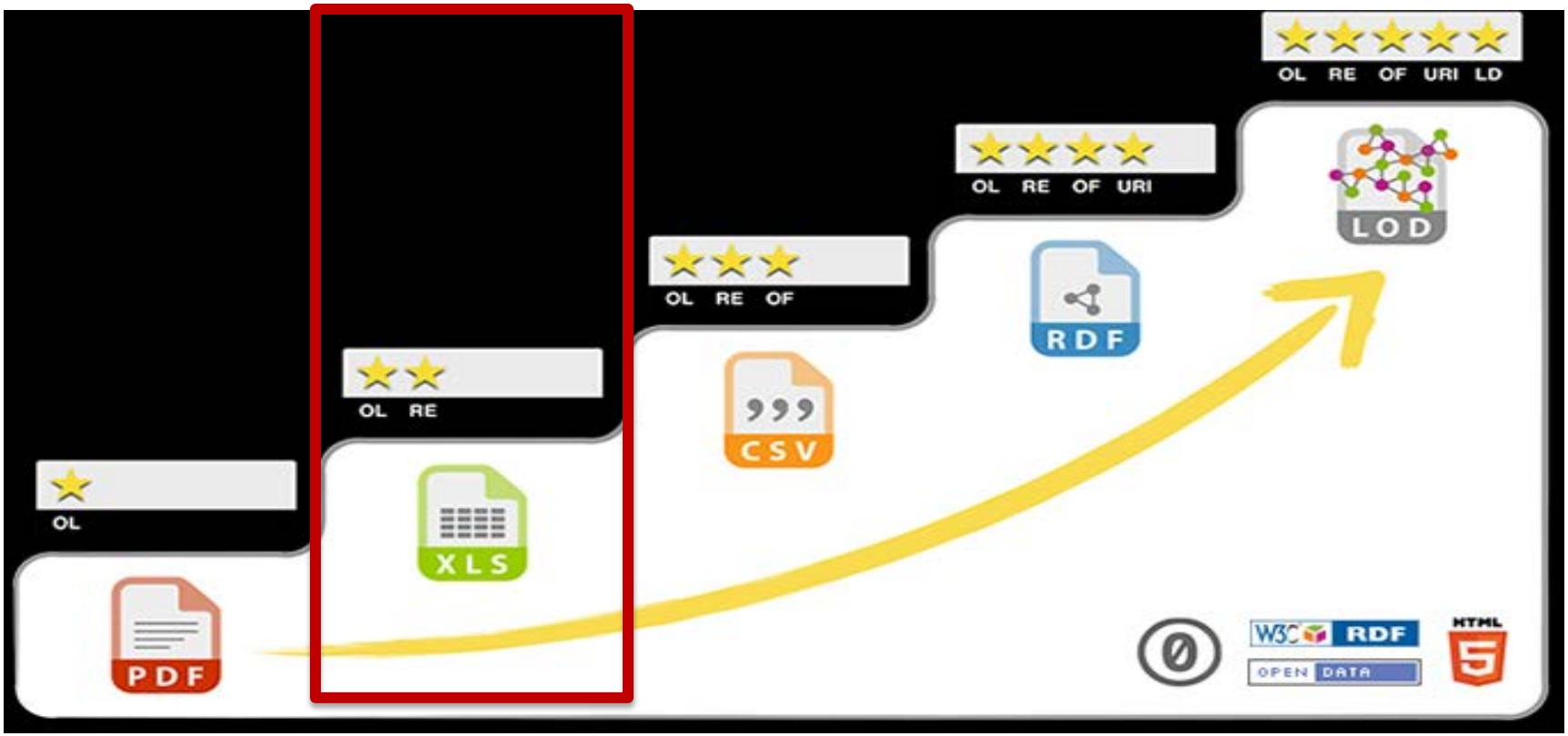
- Open data comes in many formats and is deployed in different ways
- Tim Berners-Lee (the inventor of the Web and Linked Data initiator) came up with a system to describe the accessibility of open data.
- This system describes the ways that you can improve the effectiveness of your open data
- There are many tools available to enable you to make your open data 5 star data



http://5stardata.info/en/

1 star

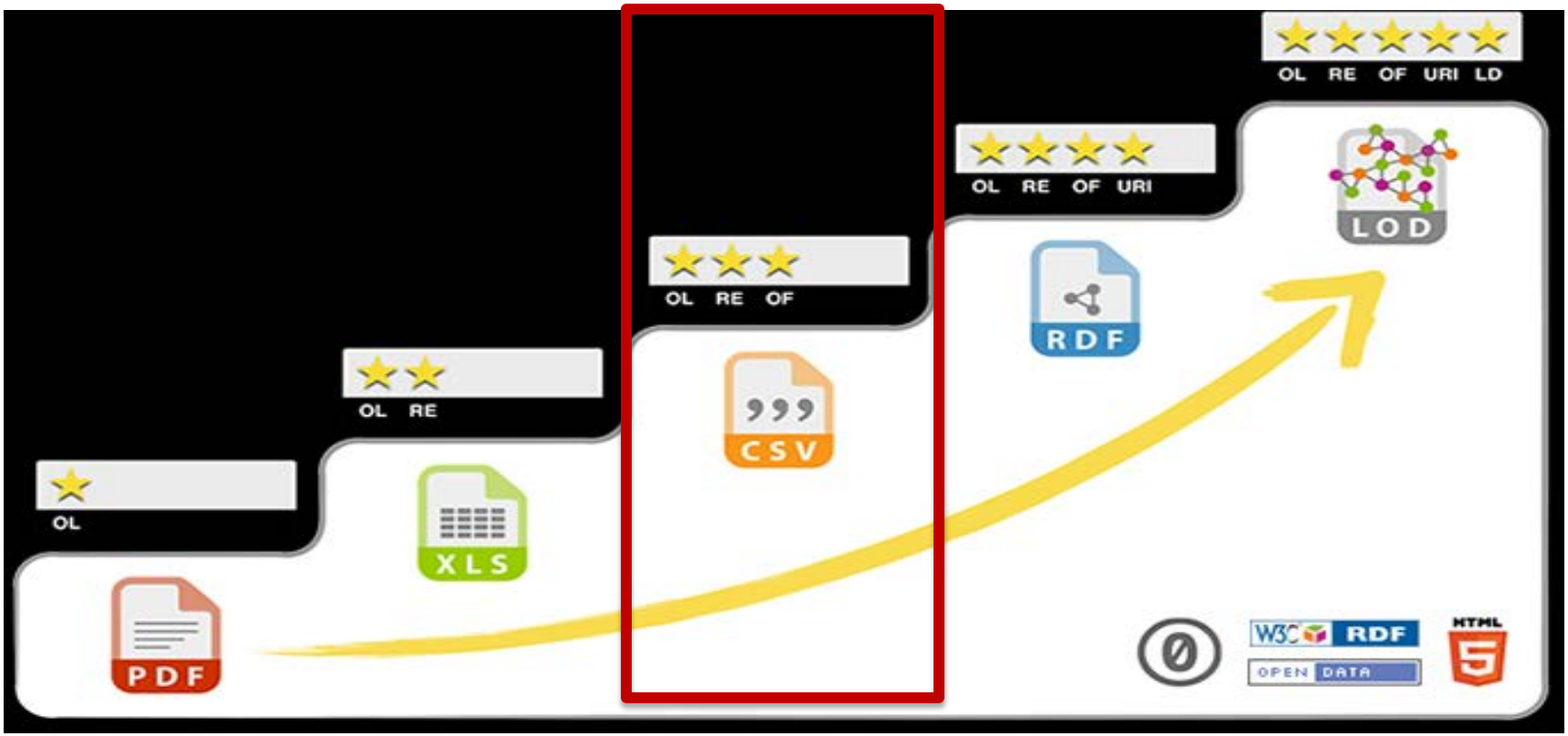
Data available on the web (in any format) under an open license



http://5stardata.info/en/

2 stars

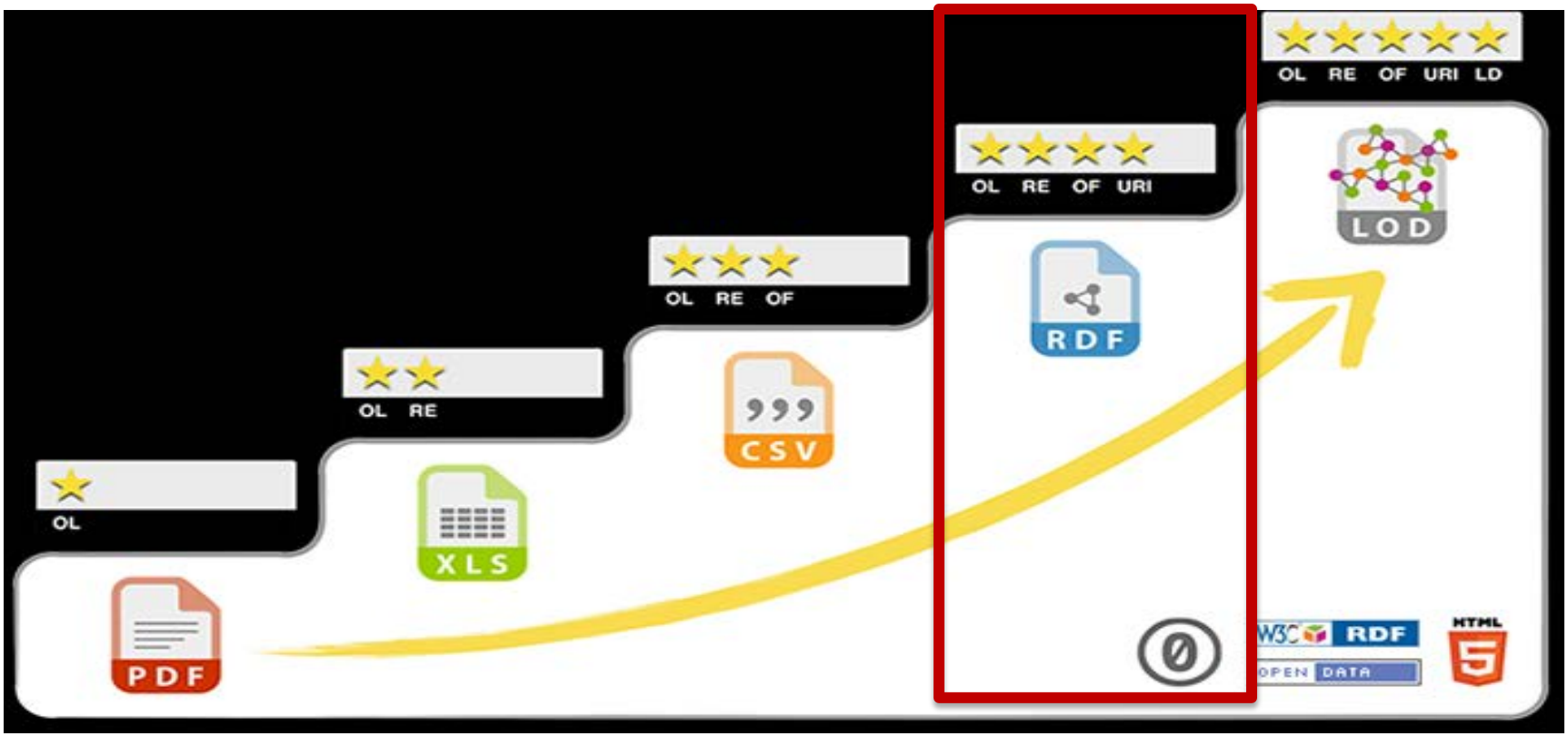
Make it available as structured data (e.g., Excel instead of image scan of a table)



http://5stardata.info/en/

3 stars

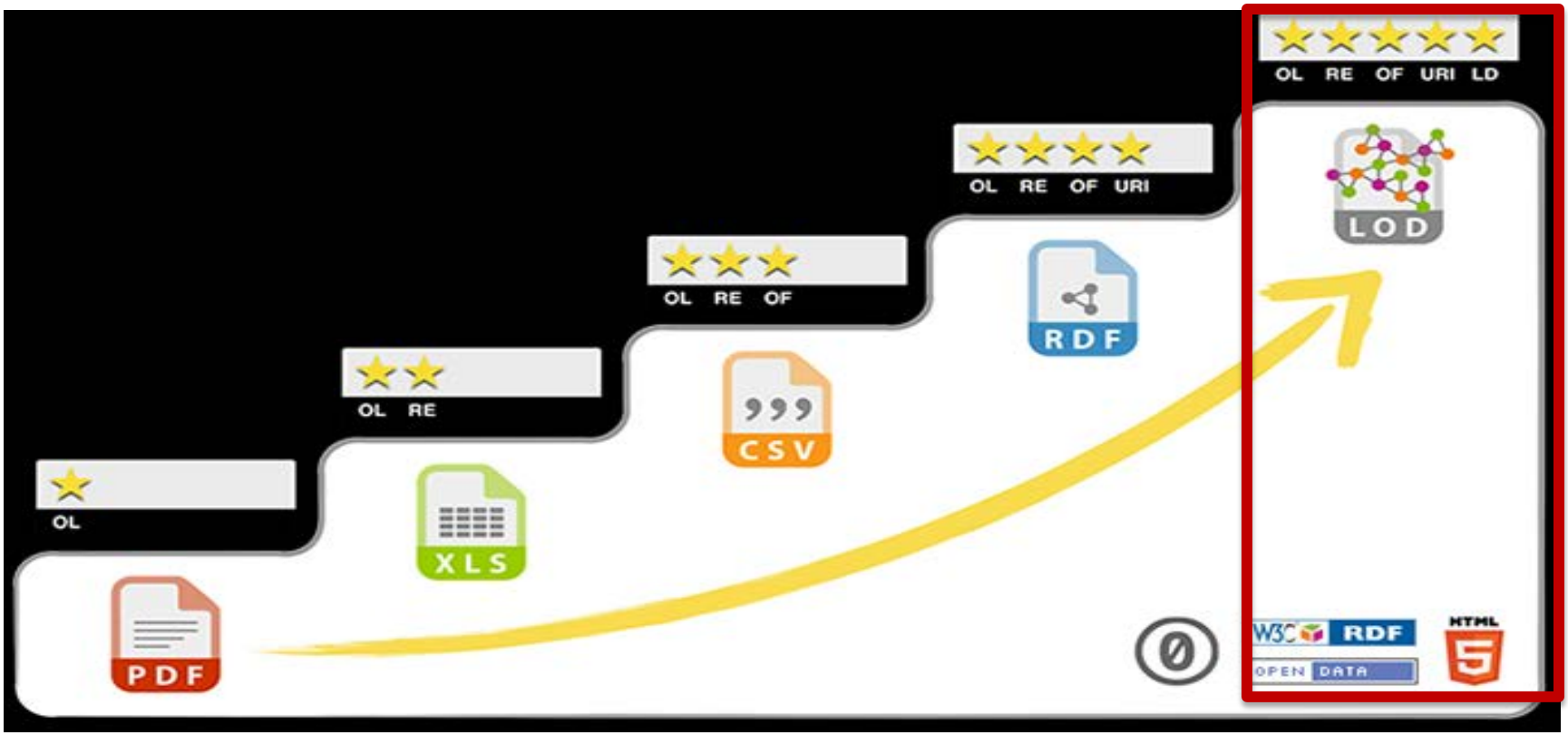
Make it available in a non-proprietary open format (e.g., CSV as well as of Excel)



http://5stardata.info/en/

4 stars

Use URIs to denote things, so that people can point at your stuff



http://5stardata.info/en/

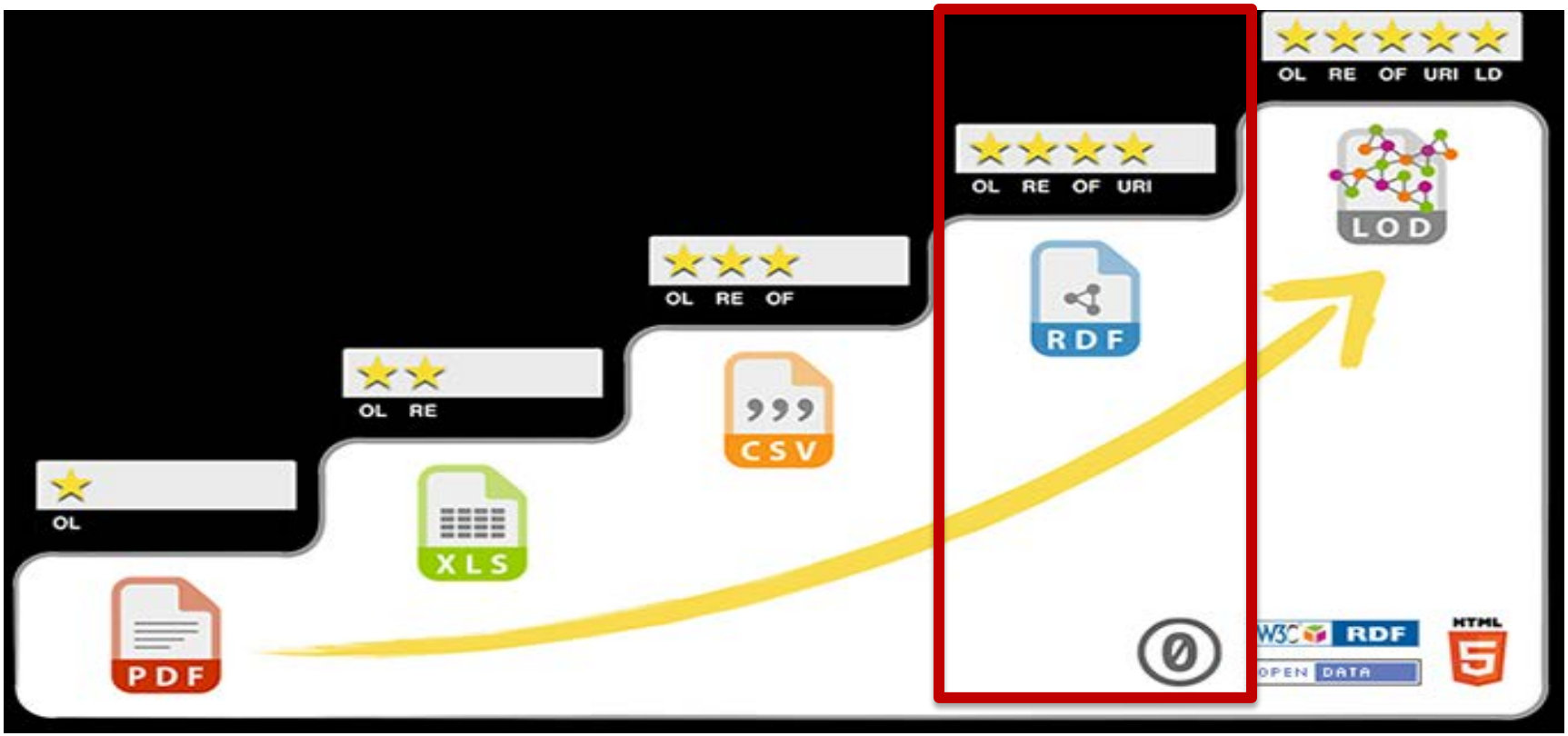
5 stars

Link your data to other data to provide context

“The goal is to turn data into information, and information into insight.”

Carly Fiorina,

Former President of Hewlett-Packard Co.



http://5stardata.info/en/

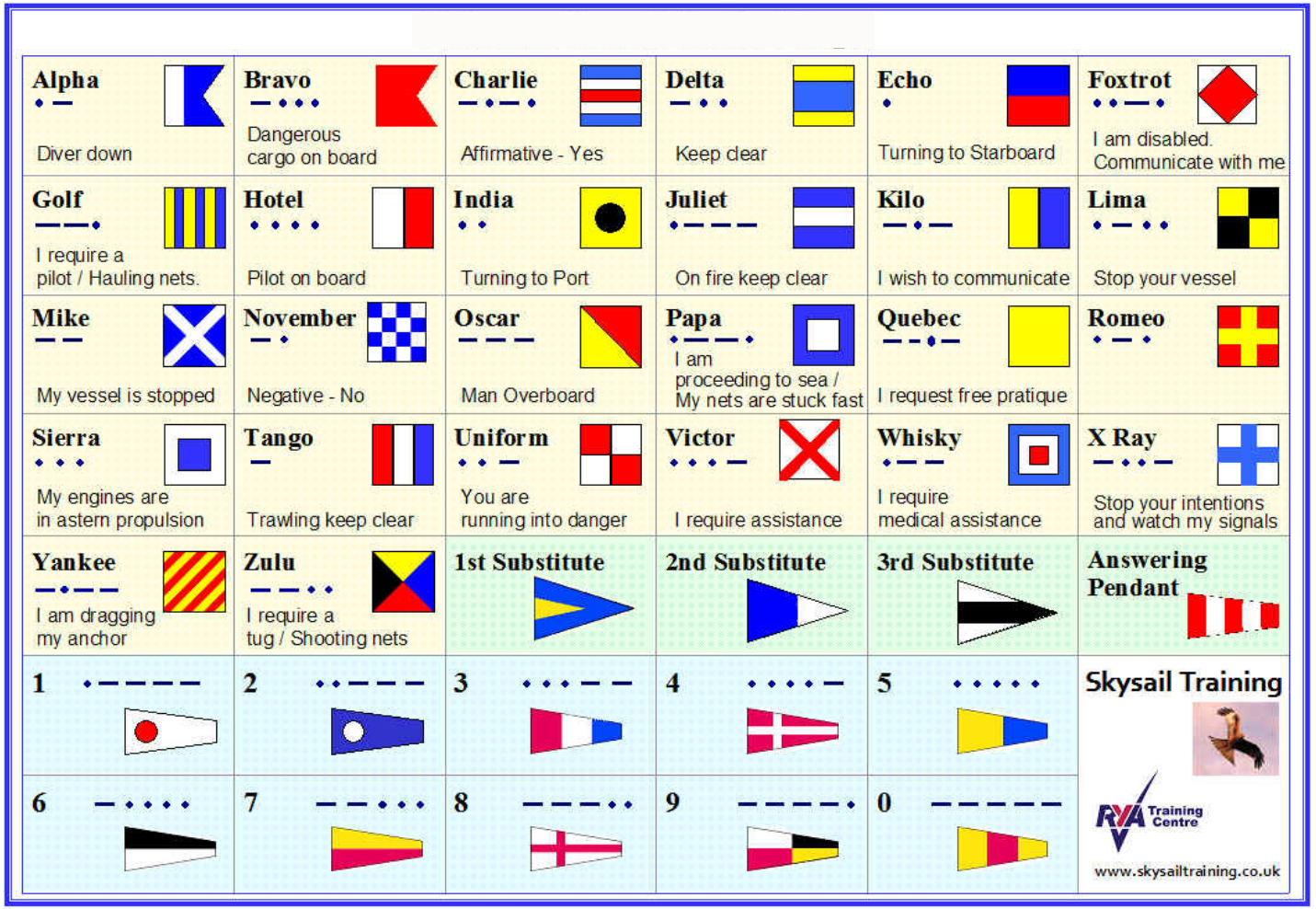
4 stars

Use URIs to denote things, so that people can point at your stuff

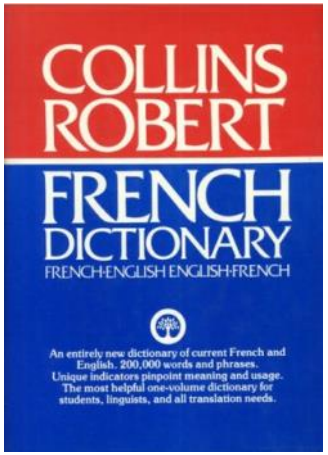
Uniform resource indicator (URI)

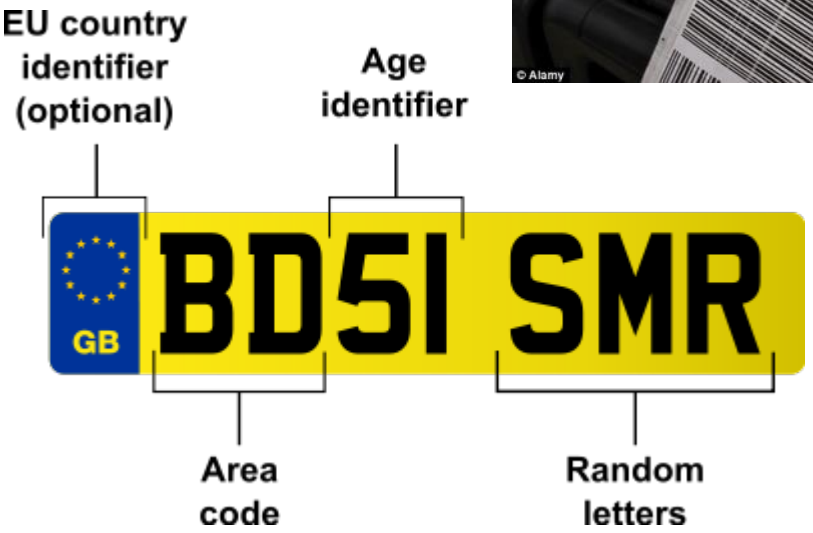
- A URI is an identifier for a specific piece of content on the World Wide Web
- It could relate to a page of text, a video or sound clip, a still or animated image, or a program
- A URI is always unique and always used to identify the same object
- We often use Uniform Resource Locators (URL) to describe the location of that object (e.g. http://vocab.nerc.ac.uk/collection/P01/current/ASLTZZ01/)
- Open data is more useful if associated with URI's that aid the definition of terms used

International Maritime Signal Flags



Standardisation everywhere





+44 (0)20 7323 8000

Country Code | City Code | Phone Number

Our cruise

Who?

PSO
Principal investigators
Captain
Crew
Technicians
Students

Etc!

Our cruise

Who?

PSO is Jane Bloggs from University of Bangor

In the cruise metadata she is referred to as:

Professor Jane Bloggs
Professor J. Bloggs
Dr Jane Bloggs (oops someone forgot her promotion!)
J. Bloggs
Bloggs, Jane

Etc!

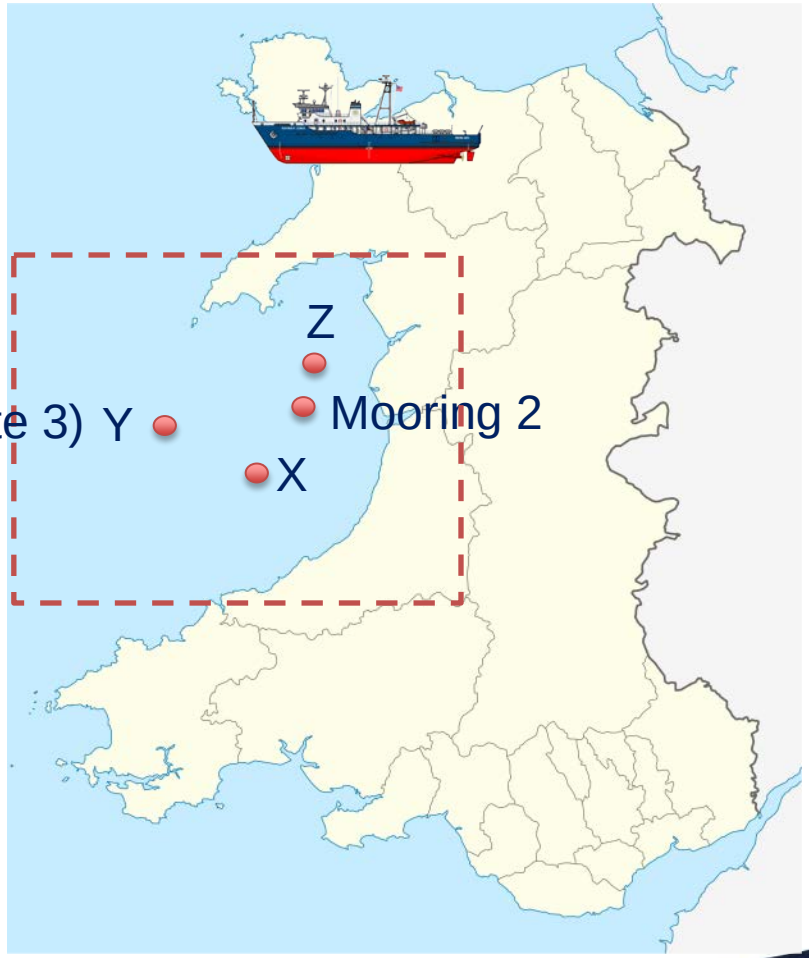
Our cruise

Where?

There are also four mooring sites:

X
Y
Z and
Mooring 2

Mooring 2 and Y have been visited every year for 10 years but sometimes site Y is referred to as Site 3.

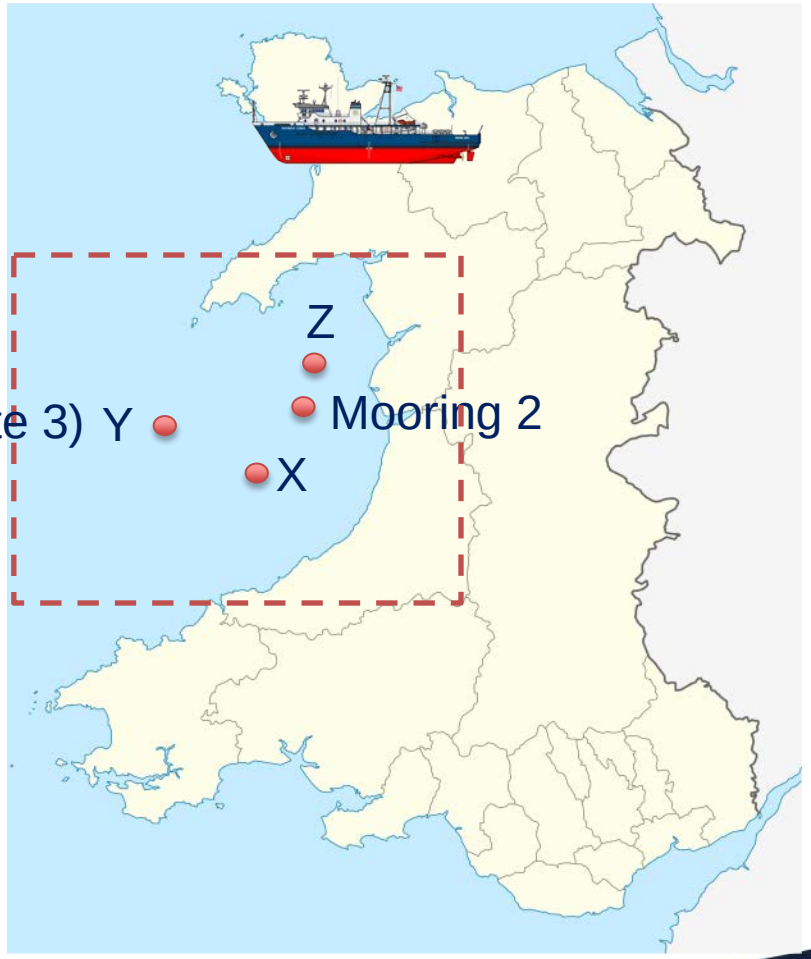


Our cruise

Where?

Sampling area is referred to as:

Irish Sea
North East Atlantic
Cardigan Bay
Wales
Etc.



Free Text

Using free text fields opens you to many problems including:

- Same concepts tagged up in so many different ways
- Difficult to map one concept to another
- Open to human error

Without reading through every piece of documentation how are you going to find all datasets that Professor Jane Bloggs was PSO for?

Vocabs – an example

Instead you might want to make a vocabulary of people with a consistent format for names.

Each record is unique and has an id associated:

Personid 490 = Professor Jane Bloggs, University of Bangor

Each dataset that she is associated with can be assigned Personid 490 and so it's easy to find them all with a simple query.

It's important that this list is well managed, the information is checked/ correct and duplicates aren't allowed to creep in!

Change is all around us

In real life things change:

Jane Bloggs gets married and changes her name to Jane Jones.

She then moves to the University of Liverpool

We could update the record with the changes

- Lose all previous information
- Lose links to Bangor for earlier records

Make new records

- How are we going to work out which record to use when?

Flexible Vocabs

- We make new entries so now we have several entries for Jane Bloggs
- One (of several) possible solutions is to split title and organisation from the name so we can have controlled vocabularies for each.

Personid	Title	Name	Organisation
490.1	Dr	Jane Bloggs	University of Bangor
490.2	Professor	Jane Bloggs	University of Bangor
490.3	Professor	Jane Jones	University of Bangor
490.4	Professor	Jane Jones	University of Liverpool

- Personid now tracks evolution and datasets can be linked to personid correct at the time of collection.

Vocabs gone bad

Not using the vocab in the way designed.

A project involved several ships being deployed to collect data. They decided to make one record in cruise and ship vocabs to cover everything.

Cruise
D354
JC048
Several

Ship
RRS Discovery
RRS James Cook
Various

This breaks the data model.

Web pages are built from the metadata:

The CTD data were collected during cruise D354 on RRS Discovery
The CTD data were collected during cruise Several on Various

The last step

You publish this Vocab **(or use an existing published one)** and assign a URI to each personid

Now you can tag your datasets with a URI which always points to the responsible person.

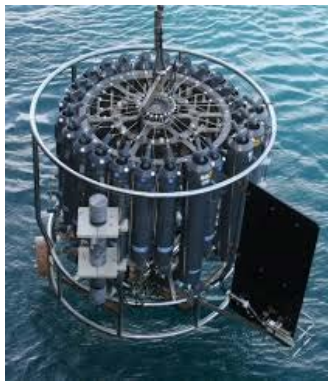
People have different roles so you also have a vocabulary of roles as well e.g. PSO, PI, Technician

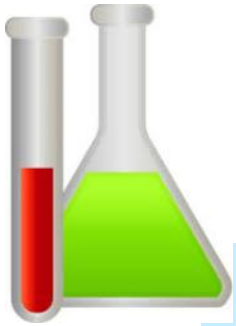
You can now find datasets Jane Bloggs was associated with or only datasets that Jane Bloggs was PSO for …

Our cruise

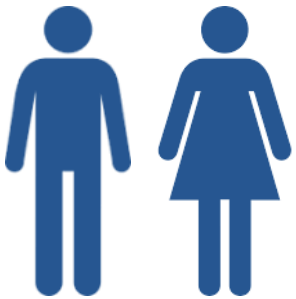
Each dataset can be linked to a unique URIs for:

- People and their roles
- The Sea Area
- The parameters collected
- The ship
- Instruments used etc.





The mooring sites can have a unique record for their name(s), locations etc. Whether locally referred to as Site 3 or Y all datasets collected at this site can be identified because linked to one Fixed Station.



Vocabs - housekeeping

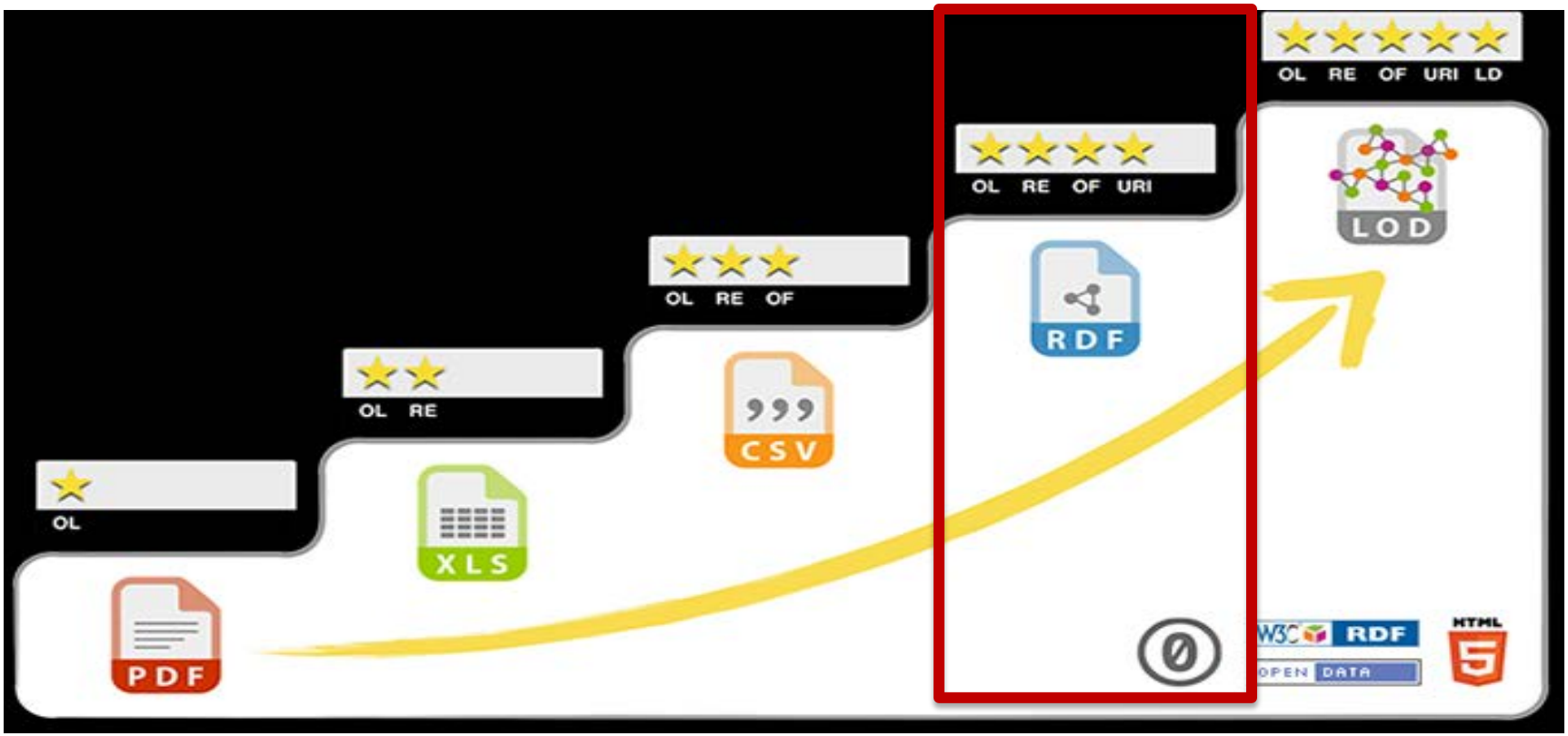
Once a vocab record and URI are created the record shouldn't be deleted

Instead you need a method to manage evolution or deprecation and replacement of terms

Vocabs need to be well managed and controlled

It removes ambiguity and variation in how information is stored

They are really powerful for grouping datasets (we will come back to this!)



http://5stardata.info/en/

4 stars

Use URIs to denote things, so that people can point at your stuff

Spoiler Alert

Later I am going to talk about existing Vocabularies managed by BODC which you can search and use.

The good news is if you want to use vocabularies you don't need to start from scratch!

There is also an existing vocab for people:

http://orcid.org/

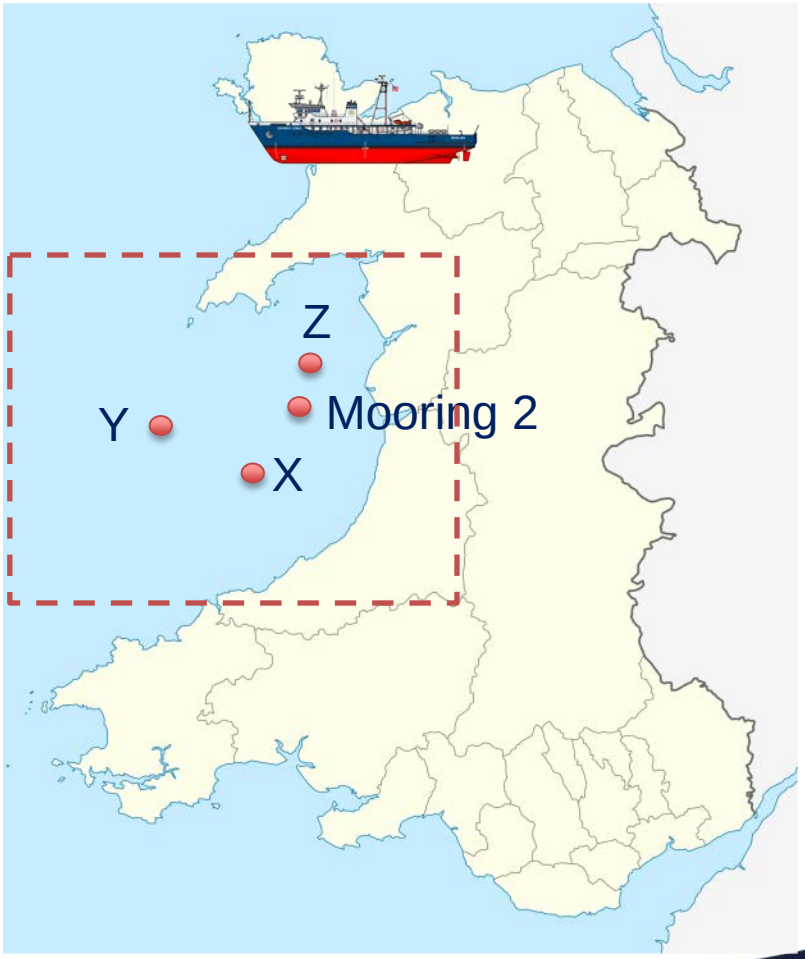
Linked Data

Where?

Sampling area is referred to as:

Irish Sea
North East Atlantic
Cardigan Bay
Wales
Etc.

We link our data to a record for Cardigan Bay



Linked Data

This will help us to find all data from Cardigan Bay but what if we are looking for any data in the Irish Sea?

We could link our data to all the variations of Sea Area:

Or we could create a hierarchy of terms in our vocab so that:

If someone is looking for data for the North Atlantic Ocean they don't need to search all of the local sea areas.

Cardigan Bay

Is a child record of

Irish Sea

which is a child record of

North Atlantic

which is a child record of

Atlantic Ocean

which is a child record of

Worldwide

If someone is only interested in data from Cardigan Bay they don't need to filter out datasets collected across the North Atlantic.

Linked Data

You can take linked data one step further:

Jane Bloggs in Bangor collects temperature data from the Irish Sea.

She delivers this data to BODC who tag it up with URI's from Vocabs and make it available under an open data license.

BODC is involved in SeaDataNet (a pan-European, marine data management infrastructure used by the oceanographic community in Europe).

We will come back to this!

Linked Data

Someone in Germany searches for all temperature data – they get Jane Bloggs data + all other open data collected by other organisations.

Jane Bloggs in Bangor wants all data collected in the Irish Sea. She gets data from organisations across Europe (and the world).

We will come back to this!

Big Data

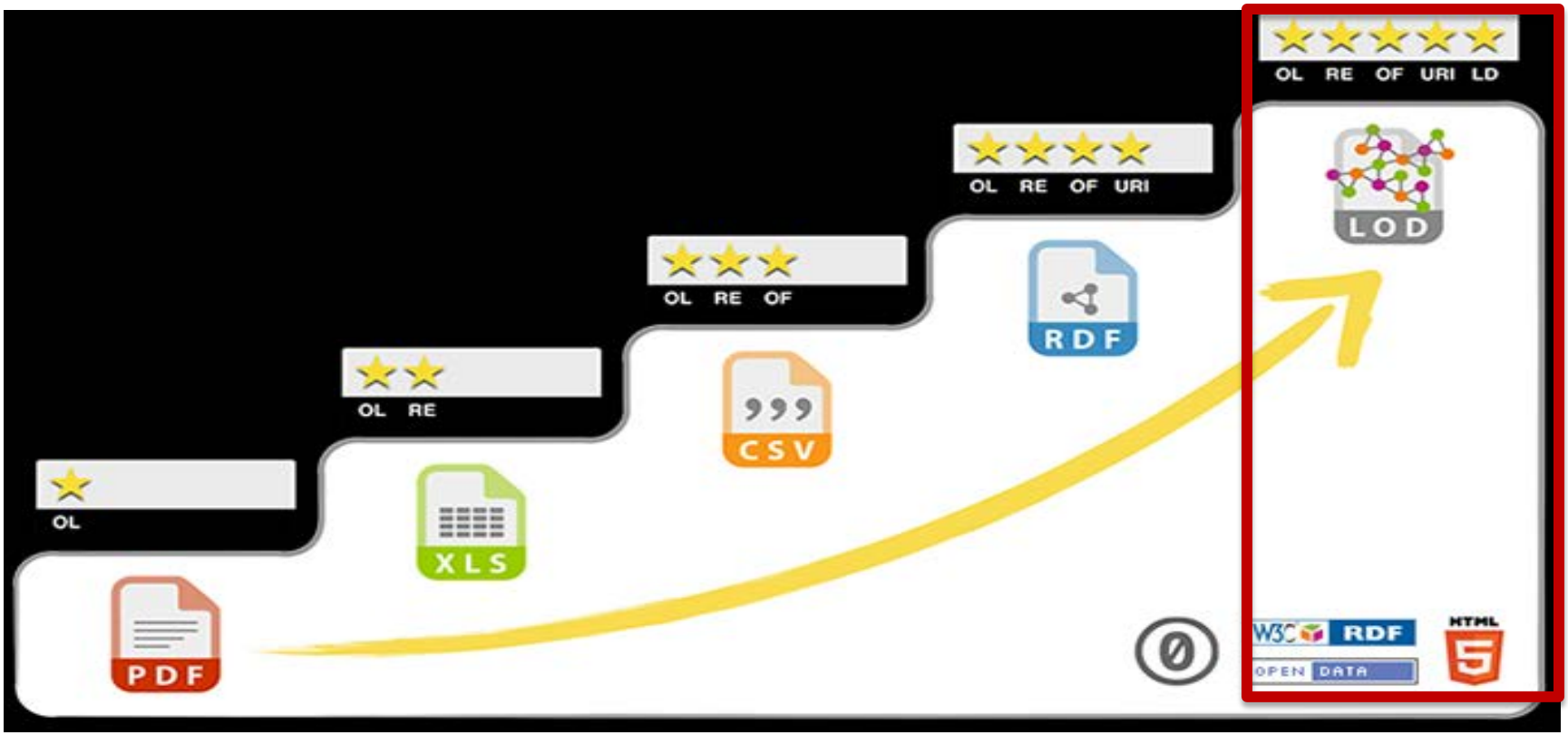
Linked data is becoming more common across all scientific disciplines. It is also growing in strength across the world.

The vocabularies of terms used to describe data are being mapped so that language and terminology is no longer a barrier.

These are the conditions where aggregating Big Data products across country borders and disciplines become a real possibility.

These data products can be designed to explore real world problems and make an impact.

Economics + Social science + Climate science



http://5stardata.info/en/

5 stars

Link your data to other data to provide context

“We are drowning in information and starving for knowledge.”

Rutherford D. Rogers

A Changing World of Data

Linked data is opening up the possibilities for Big Data across the science community and the World.

The volumes of data we as a community are collecting is increasing particularly with the rise of automated platforms (e.g. gliders and floats).

This makes data management more challenging.

It also makes the possibilities for research endless!

Summary

Open data is a blossoming concept in science and particularly the environmental science community.

It has the support of the government, funding bodies and journals.

Effective open data relies on user friendly file formats, effectively cataloguing metadata with URIs and linked data.

Data management and the use of data in the oceanographic community is changing.

Later we are going to look at the tools already available to support this …

Any questions?

"Torture the data, and it will confess to anything."

Ronald Coase

Economics, Nobel Prize Laureate