



www.bodc.ac.uk

The importance of a data legacy

KATIE GOWERS, BRITISH OCEANOGRAPHIC DATA CENTRE
BANGOR UNIVERSITY OPEN ACCESS WEEK, OCTOBER 2015



**National
Oceanography Centre**
NATURAL ENVIRONMENT RESEARCH COUNCIL

noc.ac.uk

NERC SCIENCE OF THE
ENVIRONMENT



**“An experiment is a question
which science poses to Nature,
and a measurement is the
recording of Nature’s answer.”**

Max Planck
Theoretical Physicist



National
Oceanography Centre
NATURAL ENVIRONMENT RESEARCH COUNCIL

noc.ac.uk

NERC SCIENCE OF THE
ENVIRONMENT



Scientific Data



- Testing a hypothesis
- Increasing our understanding of a system/ process/ organism
- Trends and patterns over time and/ or space
- Validation of models
- Reproducibility



Marine data are ~~valuable~~ priceless



Collecting marine data is expensive



It's often time dependent (we can't collect again)

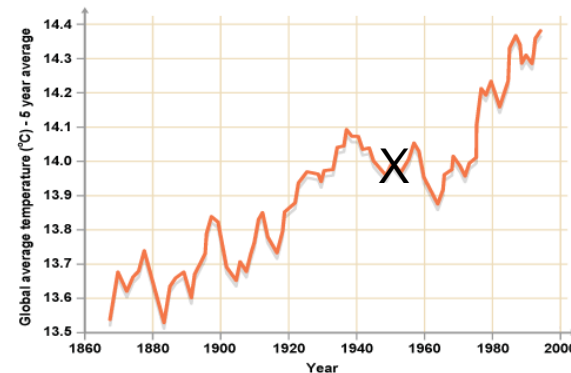
To fully understand the system we need data that have:

Global and local coverage



AND

Discrete samples and decadal trends





Why do we need good
data management?





Why do we need good data management?

Data Sharing and Management Snafu in 3 Short Acts

NYU Health Sciences Library on YouTube

<https://www.youtube.com/watch?v=N2zK3sAtr-4>

Data Security

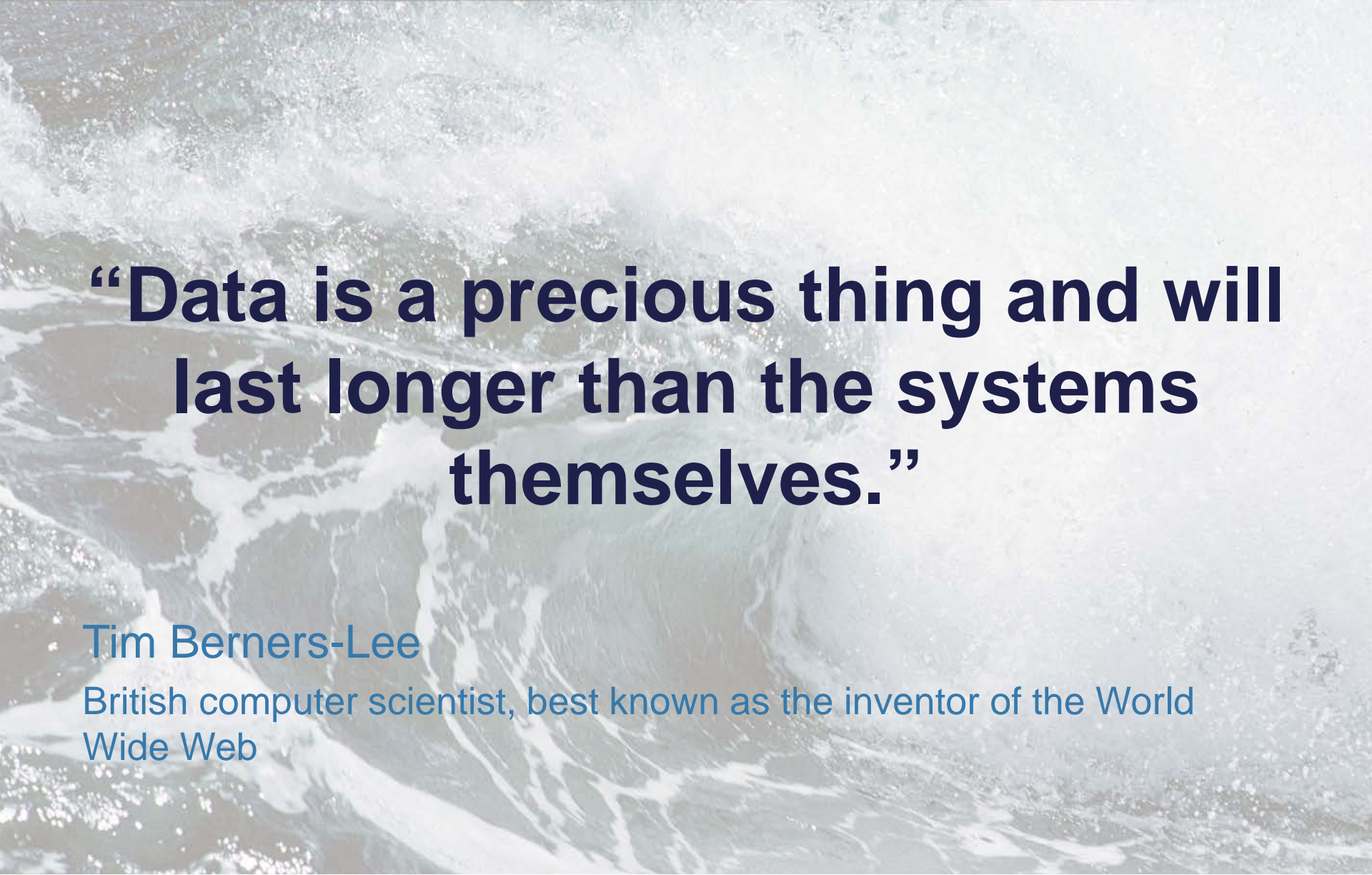


- Secure data storage and backups
- Software updates, anti-virus software and firewalls
- Write access to directories (concept of least privilege)
- Confidentiality

Data Security



- Data versioning
- Interoperability (file formats)
- Develop clear procedures
- Sound programming
- Processing notes



“Data is a precious thing and will last longer than the systems themselves.”

Tim Berners-Lee

British computer scientist, best known as the inventor of the World Wide Web

What?

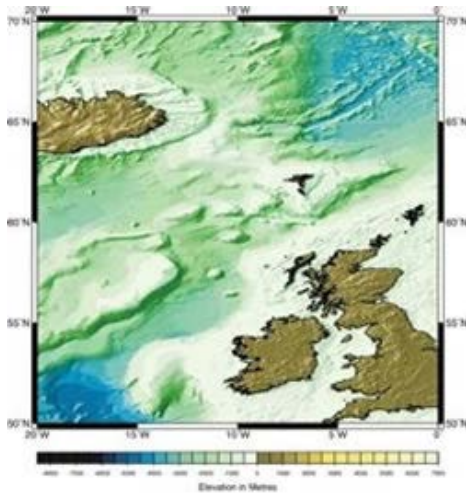
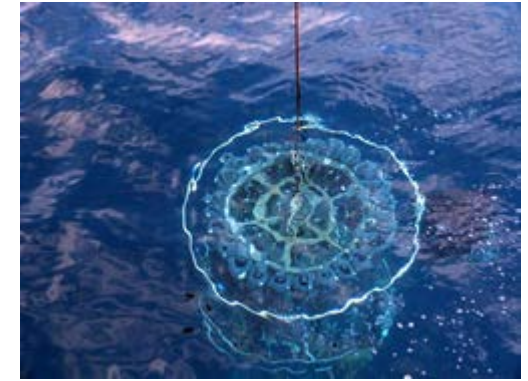
When?

Where?

How?

Why?

Who?



Without a context environmental data are just numbers ...

Metadata



- Unambiguous parameter titles
- Full Units
- What instruments and methodology followed?
- What quality control have you done? Are the data of a good quality?
- Who is responsible for the project and the dataset?
- What restrictions should be applied?



Quality Control

BODC do not delete outliers – they are flagged instead.

Each parameter has its own flag column which is marked with an appropriate flag for the cycles affected.

Values that are impossible (a concentration of $-100 \mu\text{mol/l}$) are set to an appropriate absent data value.

All data submissions are archived (so it's always possible to get back to what was sent to us).

By retaining the values an end-user can decide what to include (or not include) in their analysis.

The authors of 516 biological studies published between 1991 and 2011 were emailed and the raw data requested.

> 90 % of the oldest data (from papers written more than 20 years ago) were inaccessible.

In total, even including papers published as recently as 2011, they were only able get hold of the data (from the authors) for 23 % of the papers.

Vines, T.H. et al, 2014, The Availability of Research Data Declines Rapidly with Article Age, *Current Biology*, 24 (1), 94-97.

Any questions?

“Things get done only if the data we gather can inform and inspire those in a position to make [a] difference.”

Mike Schmoker
Author