Parameter Vocabularies in the NERC Data Grid (NDG) Project

Roy Lowry[3], Ray Cramer[3], Marta Gutierrez[2], Michael Hughes[3], Kerstin Kleese van Dam[1], Siva Kondapalli[3], Susan Latham[2], Bryan Lawrence[2], Kevin O'Neill[1], Andrew Woolf[1]

[1]CCLRC e-Science Centre, [2]British Atmospheric Data Centre, [3]British Oceanographic Data Centre

NDG aim: The seamless integration of distributed sources of scientific data without the need for human intervention. Current focus is to provide integrated access to the extensive data holdings of the British Atmospheric Data Centre (BADC) and the British Oceanographic Data Centre (BODC).

Requirements:

Parameter usage vocabulary: Controlled metadata vocabulary to provide detailed and unambiguous labels for source data.

Parameter discovery vocabulary: Broad terms arranged into a hierarchy of increasing specificity to focus a user search for a specific data point.

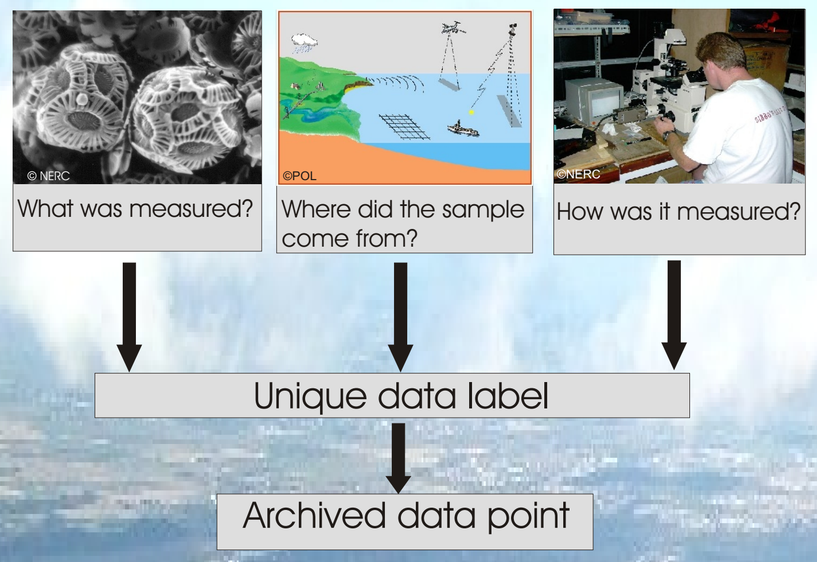

Fig 1. The components of a BODC data definition. This provides a consistent, unambiguous and detailed Parameter Usage Vocabulary

NDG 'Use' metadata

All BODC data are labelled using keys defined in BODC's 17,000+ term "Parameter Usage Vocabulary" (Fig 1).

BADC mostly use non-standardised text strings to label data but also use some "Standard Name List" terms from the Climate and Forecast (CF) content standard.

For interoperability, NDG aims to unite the two centres' data holdings via the GML-based Climate Science Markup Language (CSML).

Challenges:

Creating a phenomenon dictionary encompassing the two data centres.

Constructing maps between data and dictionary entries to allow tools to automatically generate CSML records.

NDG 'Discovery' metadata

Rules for generating discovery metadata:

Discovery vocabulary terms can be organised into a hierarchy of increasing specificity (Fig 2).

Parameters can be described using as many vocabularies as required (Fig 3).

NDG approach:

Adoption of the Global Change Master Directory (GCMD) parameter discovery vocabulary.

Linkage of the GCMD vocabulary with the BODC Parameter Discovery Vocabulary.

Linkage, by proxy, to the BODC Parameter Usage Vocabulary and eventually the CF Standard Names.

This will permit automatic generation of discovery metadata from datasets.

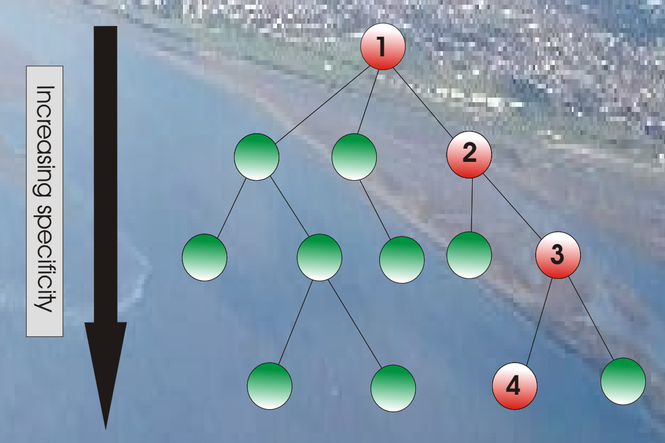

Fig 2. Navigation through a parameter discovery vocabulary. Each parent node represents a broader term linked to more specific child nodes. A search starting at point 1 requires only a few steps to get to point 4

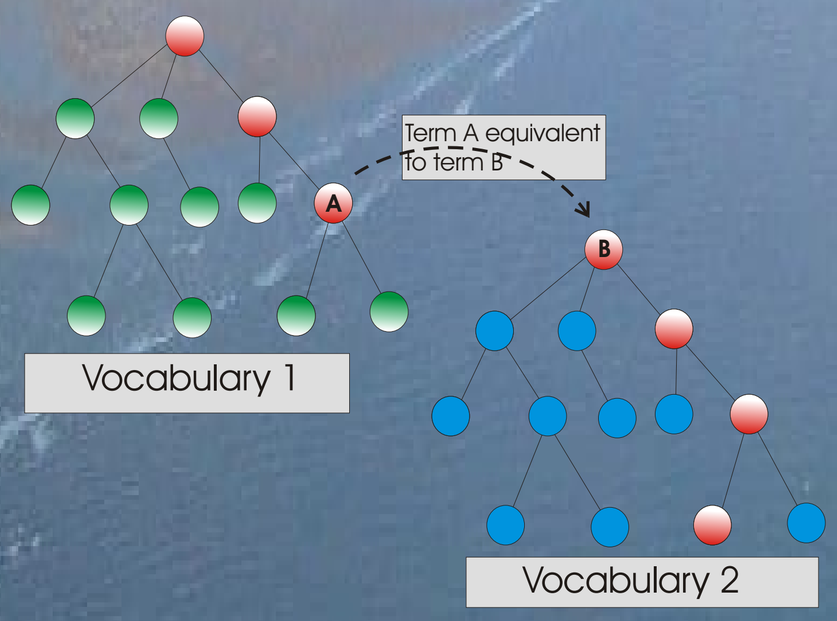

Fig 3. Linkage of two discovery vocabularies via equivalent term mapping allows the user to travel from one set of definitions to another

Issues

Synonyms: Data users and creators use many different words to describe the same things. For successful interoperability, a system must cope with this. For example, how is a computer or a non-expert user to know that "PCB28" is the same as "2,4,4-trichlorophenyl"?

Semantic web technologies such as Web Ontology Language (OWL) and the Simple Knowledge Organisation System (SKOS) have great potential in this area, but this is yet to be exploited by the NDG project team.

We invite input from members of the e-Science community who have experience in this area.

For more details, see the paper associated with this poster.

Contact: Roy Lowry, rkl@bodc.ac.uk; **Website:** http://ndg.nerc.ac.uk/

NATURAL ENVIRONMENT RESEARCH COUNCIL

British Oceanographic Data Centre
NATURAL ENVIRONMENT RESEARCH COUNCIL

BRITISH ATMOSPHERIC DATA CENTRE

CCLRC