NCAS-Obs Data Standards: Improving Management, Documentation, Processing, Discovery and Usage.

G. Vaughan[1], B. Brooks[2], A. Stephens[3]. G. Parton[3], W. Garland[3], J. Groves[4,] D. Walker[4], D. Sproson[5], C. Walden[6] & D. Hooper[6]

1- NCAS @ University of Manchester, 2 - AMF, 3 - STFC CEDA, 4 - NCAS-IT, 5 - FAAM, 6 - NFARR

It's Complicated!!

Science is complicated. Data management is complicated.

Traditionally scientists tend to focus more on their findings rather than generating re-usable and well-described products. In the modern research environment there are new drivers to document, disseminate and publish both code and data so that it can be cited in the literature. Furthermore, the need to demonstrate the "impact" of science necessitates a new approach to managing data. Whilst the requirement is clear, there is a need for more work within our community to provide clear guidance to scientists to help them produce outputs that can be easily catalogued, discovered, analysed and are fit for re-use across a range of disciplines.

The logical solution to this problem is to provide, and enforce, detailed data standards, comprised of:

- Agreed data formats: ideally binary (efficient for storage) and supported by open source software packages.
- Agreed metadata standards and structures: rules on how to structure spatial, temporal and other data types.
- Agreed file/directory naming conventions: supporting both machine- and human-readable patterns.
- Controlled Vocabularies: agreed terminology that defines scientific and project-specific names to ensure consistency and avoid ambiguity.

The NCAS-Obs approach to data standards

In the NCAS-Obs project we have taken a rigorous and pragmatic approach that engages with scientists, provides clear templates for how the data files should look and allows automated checking of the format, structure, naming-conventions and adherence to controlled vocabularies. Figure 1 shows how this approach is embodied by a cycle of continuous improvement.

The **advantages for the data provider** are:

1. Complex standards/formats are simplified into logical templates
2. All data generation, and metadata tagging, can be automated to comply with the standards
3. Compliance-checking tools are available to check data files
4. Workflows can be standardised to reduce errors/inconsistencies
5. Data products are formatted consistently (in CF-netCDF)
6. Tools, such as CIS, can be used to analyse/visualise the products

The standardisation has the resulting **advantages for the end-user**:

1. CF-standard parameter definitions enable confidence in combining with other data sources
2. Consistency of format and metadata
3. Consistency between different products - reducing time required to work with new data products
4. Discoverability and detailed meaningful metadata
5. Tools are available for analysis and visualisation
6. Reliable, documented products and procedures

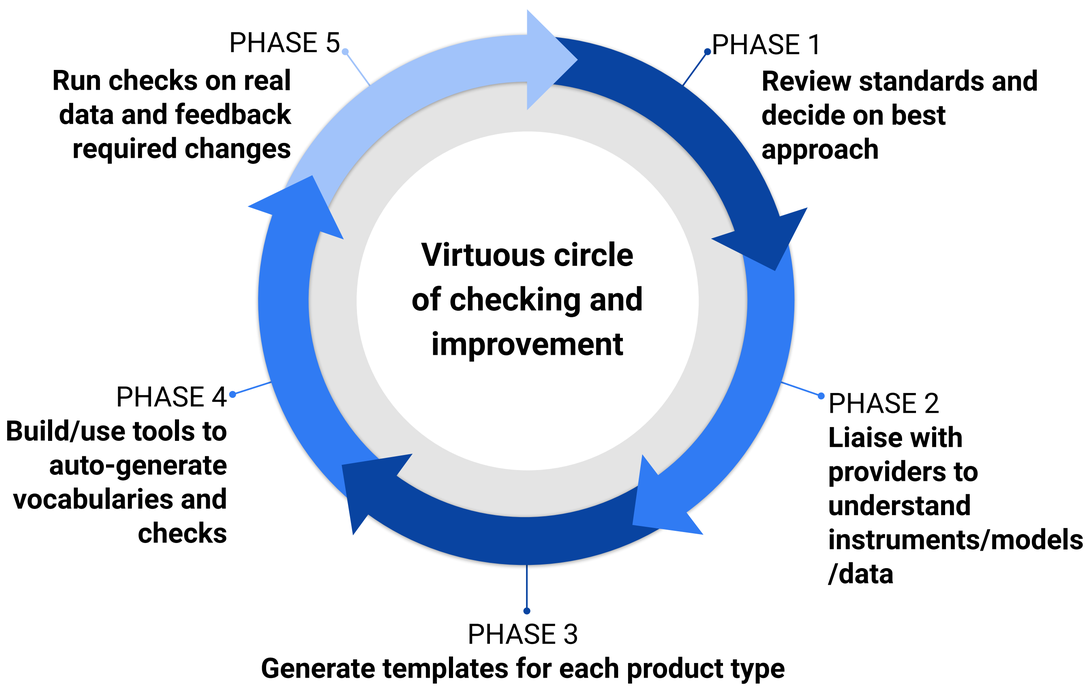

Figure 1. Cycle of creating data standards, development, testing, feedback and improvement.

Conclusions

Investment in data standards and protocols involves significant effort on the side of both data managers and scientists. Expertise is required on both sides to understand the domain, the outputs and the appropriate standards to adopt. However, the initial investment of time pays much greater dividends later through:

- Increased automation and robustness of product-generation processes.
- A reduction in the number of iterations required before data can be submitted to a data centre/repository.
- Ease of archival, scanning, conversion, intercomparison, analysis and visualisation of data.

These savings are important to the scientist who provides the data because the outputs of her research become more accessible which in turn leads to greater usage, citation and collaboration opportunities.

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT