How does a data archive remain relevant in a rapidly evolving landscape: the case of the 4TU.Centre for Research Data

**Maria Cruz**, Jasmin Böhmer, Egbert Gramsbergen, Marta Teperek, Madeleine de Smaele, Alastair Dunning
TU Delft Library / 4TU.Centre for Research Data, The Netherlands

**IDCC 2018, Barcelona, 20 February 2018**
Delft (just about above sea level)
Delft University of Technology

176 years old – largest and oldest of the four Dutch technical universities.

World-class research with focus on science, engineering and design.
4TU.Centre for Research Data

Hosted by TU Delft Library
4TU.Centre for Research Data

- 2008: Started by 3 (of the 4) Dutch technical universities as a national service for archiving science & engineering data.
- 2010: Up and running as a fully operational data archive.
- Front offices in Delft, Eindhoven and Twente.
- Researchers from other (national and international) institutions can upload data.
4TU.Centre for Research Data

- Certified and trusted repository for engineering and applied sciences.
- At least 15 years long-term curation and access to open research data.
- Each dataset assigned a DOI and released with a license.
- Metadata checked and improved; documentation validated before publication.
A lot has changed since 2008...
47 repositories run by institutions in the Netherlands alone
How can 4TU.Research Data continue to be relevant in this rapidly evolving landscape?
Repositories need to have a subject or format focus to remain relevant and be successful in the long term.

Subject-based repositories often provide specific services and tools that generic repositories don’t.
What distinguishes 4TU.Research Data from other generic data repositories?

What subject specific services and tools does it already provide?
4TU.Research Data

90% of data in netCDF

• 6519 out of a total of 7548 datasets.*
• 30.3 TB out of 32.6 TB.

*As of 18 February 2018.
Two questions:

What is netCDF and what is interesting about it?

How did the archive end up with 90% netCDF data?
NetCDF – a brief introduction

- Created and maintained by Unidata, part of the University Corporation for Atmospheric Research (USA).
NetCDF – a brief introduction

“NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.”

https://www.unidata.ucar.edu/software/netcdf/
Gridded Multidimensional Data

4D data, such as temperature over an area varying with time and altitude, is stored as a series of two-dimensional arrays.

http://pro.arcgis.com
How did the archive end up with 90% netCDF data?
The simple answer is

WATER
NetCDF is extensively used in climate, ocean, and atmospheric sciences.
Water management
Coastal protection
Sea level rise
Climate change
Rain
Weather
What’s interesting about netCDF?
Self-Describing Data Format

A netCDF file includes information about the data it contains.
Information about variables, including units, coordinate systems, etc.

Community-defined conventions (defining metadata) & provenance information

R1. meta(data)

R1.3. (meta)data meet domain-relevant community standards.

R1.2. (meta)data are associated w/ their provenance.

R1. meta(data) have a plurality of accurate and relevant attributes.

```c
metood simple {
  dimensions:
    latitude = 2;
    longitude = 2;
    time = UNLIMITED; // (5 currently);
  variables:
    double time(time);
      time:standard_name = "time";
    time:units = "minutes since 1994-01-01 00:00:00";
    time:long_name = "time";
    float latitude(latitude);
    latitude:standard_name = "latitude";
    latitude:units = "degrees_north";
    latitude:point_spacing = "even";
    latitude:long_name = "latitude";
    float longitude(longitude);
    longitude:standard_name = "longitude";
    longitude:units = "degrees_east";
    longitude:point_spacing = "even";
    longitude:long_name = "Longitude";
    float temp(temp, latitude, longitude);
      temp:standard_name = "surface_temperature";
    temp:units = "deg C";
    temp:FillValue = 2.e+020f;
    temp:valid_min = -50.0f;
    temp:valid_max = 60.0f;
    temp:comment = "This parameter may be erroneous.";

  // global attributes:
  :Conventions = "CF-1.0";
  :institute = "The British Atmospheric Data Centre.";
  :source = "Model developed in conjunction with IFLAP2."
  :history = "10 Sep 2002 - Created by hand.in",
  "18 Mar 2003 - Modified by feet.\n"
  :title = "Model output from imaginary model (temperature)"
  :comment = "Not very useful data."
  :references = "A great report somewhere!"

  data:
    time = 0.5, 1.5, 2.5, 3.5, 4.5;
    latitude = 54.2, 54.4, 54.6;
    longitude = 2.0, 2.5;
    temp = 34.5, 32.2, 23.7, 19.6, 35.0, 29.2, 24.4, 5.6, 7.2, 0.1
      18.6, 15.2, 13.1, 4.6, 3.7, 0.2, 9.7, 34.2, 26.7, 28.7;
    2.1, 3.4, 5.6, 7.8, 9.0, 10.2, 11.2, 11.6, 11.7, 11.8;
```

http://cedadocs.ceda.ac.uk/72/1/nc_example.html
NetCDF files are scalable

A small subset of a large dataset may be accessed efficiently (without the need to download the whole file).
OPeNDAP Protocol

Access enhanced by serving netCDF data via OPeNDAP protocol.

Make selections of the data through a web form and export to different formats.

Inspect the structure and internal metadata of netCDF data.
We want to do more with netCDF

• Currently exploring options for providing further services, both technical and training, related to netCDF data.
• What advantages could accrue from becoming a centre of expertise and a recognised source of netCDF data, particularly in hydrology, weather, climate and atmospheric sciences?
Possible Options

- Visualisation services, data processing and data mining.
- Scalable Dynamic Data Citation (per RDA recommendation).
- Training, advice and guidance.
- More outreach and international collaboration.
Progress so far

- Conducting interviews with TU Delft researchers who have deposited data in the 4TU archive.
- Making contact with other relevant organisations (e.g. Royal Netherlands Meteorological Institute).

Open Working Blog for updates: [https://openworking.wordpress.com/category/netcdf/](https://openworking.wordpress.com/category/netcdf/)
Future Plans

• Interview researchers outside TU Delft and not just data depositors but also data users.
• Organise requirements gathering workshop.
• Publish report with our findings.
• Collaborate with other universities and organisations.
Want to collaborate?

- Interested in netCDF data?
- Have or know of netCDF datasets that need to be archived?

M.J.MarquesdeBarrosCruz@tudelft.nl
Questions, suggestions, feedback, want to collaborate?

M.J.MarquesdeBarrosCruz@tudelft.nl

- Pre-print of the practice paper: 10.17605/OSF.IO/JGRKB
- Presentation: 10.5281/zenodo.1175238