IJDC | *General Article*

From Passive to Active, From Generic to Focussed: How Can an Institutional Data Archive Remain Relevant in a Rapidly Evolving Landscape?

Maria J. Cruz
TU Delft Library

Jasmin Böhmer
TU Delft Library

Egbert Gramsbergen
TU Delft Library

Marta Teperek
TU Delft Library

Madeleine de Smaele
TU Delft Library

Alastair Dunning
TU Delft Library

Abstract

Founded in 2008 as an initiative of the libraries of three of the four technical universities in the Netherlands, the 4TU.Centre for Research Data (4TU.Research Data) has provided a fully operational, cross-institutional, long-term archive since 2010, storing data from all subjects in applied sciences and engineering. Presently, over 90% of the data in the archive is geoscientific data coded in netCDF (Network Common Data Form) – a data format and data model that, although generic, is mostly used in climate, ocean and atmospheric sciences. In this practice paper, we explore the question of how 4TU.Research Data can stay relevant and forward-looking in a rapidly evolving research data management landscape. In particular, we describe the motivation behind this question and how we propose to address it.

Received 05 February 2018 ~ *Accepted* 05 February 2018

Correspondence should be addressed to Maria Cruz, University Library, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands. Email: m.j.marquesdebarroscruz@vu.nl

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v13i1.613

Introduction

From Generic to Focused – How Does Our Project Fit in the Context of Generic vs. Domain-specific Archives?

A recent study (Cox et al., 2017) covering higher education libraries in Australia, Canada, Germany, Ireland, the Netherlands, New Zealand, and the United Kingdom (UK), reports that academic libraries currently focus mainly on advisory services, such as data management planning support and training, with limited provision of technical services, such as carrying out long-term preservation of data. This study also reports that over 50% of the libraries surveyed in Australia, the Netherlands, and the UK, identified "running a data repository" as one of their top strategic priorities.

When institutions build their own data repositories, these tend to be generic in nature, suitable for dealing with the wide range of data formats and disciplines represented by the modern university. However, as reported by Assante et al. (2016), generalist data repositories are unable to deal with unusual datasets and their different usage, and they lack a "designated community". At the same time, researchers tend to prefer – or are often recommended – to use subject-based repositories (Husen et al., 2017), which provide specific tools for the types of data they create and allow their data to connect with other practitioners in their field.

In the light of this context and to stay relevant in an age of growing and evolving capacity in research data services, we have embarked on a project that aims to reinforce the services provided by 4TU.Research Data. Hosted by the Library of the Delft University of Technology (TU Delft) and financed by three of the four technical universities in the Netherlands (TU Delft, Eindhoven University of Technology and the University of Twente), 4TU.Research Data supports research data management and long-term archiving of general technical and scientific research data[1]. In our project, we are exploring options for providing and expanding services related to netCDF data – 4TU.Research Data's dominant data type. This is an area where 4TU.Research Data already provides specific technical services and around which a community and a centre of expertise could be built.

NetCDF Data and the FAIR Principles

NetCDF[2] is a data format and data model that is especially suited for multidimensional, gridded, numeric data. It is very interesting from an archiving point of view, because netCDF files include information about the data they contain in a fully-machine readable way; in particular, netCDF files combine data and metadata in one single data container. The metadata includes general, supporting and provenance information, as well as detailed information about variables, scales, units, etc. Conventions on how to apply and fill in these specific and global attributes exist[3]; the 'NetCDF Climate and

1 Data archive of 4TU.Centre for Research Data: http://data.4tu.nl/
2 Network Common Data Form (NetCDF): https://www.unidata.ucar.edu/software/netcdf/
3 NetCDF Documentation: http://www.unidata.ucar.edu/software/netcdf/docs/attribute_conventions.html

Forecast (CF) Metadata Conventions'[4] is one example of a set of community-driven conventions, used for the description of Earth Sciences data.

From Active to Passive – NetCDF Data as an Example of Putting the FAIR Principles into Practice

In July 2016, the European Commission issued new guidelines[5] on FAIR data management in Horizon 2020[6], the Commission's eighth framework programme funding research, technological development, and innovation. The guidelines are intended to help Horizon 2020 grantees make their research data findable, accessible, interoperable and reusable (FAIR) – a requirement that has been extended to all projects funded by Horizon 2020 after a successful pilot programme in 2014-2016, which included only selected areas of Horizon 2020. From 2017, research data in Horizon 2020 "is open by default, with possibilities to opt out"[7].

In relation to the European Commission's guidelines on FAIR data management, the netCDF format appears to have beneficial features regarding interoperability and reusability. As required by the FAIR principles (Wilkinson et al., 2016) in what concerns interoperability and reusability, netCDF files include vocabulary conventions, qualified references, and domain-relevant community standards. In contrast to many other data formats, this information is included in one single container – the data file itself. Storing these 'enhanced' files in an archive, such as the 4TU.Research Data archive – which provides a DOI, standardised metadata and a license – secures and amplifies the findability, accessibility, and also the reusability of netCDF datasets.

NetCDF Data at 4TU.Research Data

As of February 2018, 4TU.Research Data stored 7,543 datasets, corresponding to 29.6 tebibytes[8] (TiB) of data (see Table 1). Of these datasets, 6,519 were coded in netCDF, corresponding to 27.6 TiB of data, or just over 90% of the total data in volume. The vast majority of netCDF datasets in 4TU.Research Data originate from TU Delft, in particular from the Faculty of Civil Engineering and Geosciences – consistent with the wide use of netCDF among some fields in engineering and geoscience, namely hydraulic engineering and weather and climate research[9].

4 CF Conventions and Metadata: http://cfconventions.org/
5 Guidelines on FAIR Data Management in Horizon 2020: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
6 Horizon 2020: https://ec.europa.eu/programmes/horizon2020/
7 Open Research Data in Horizon 2020: https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf
8 A tebibyte (TiB) is a unit of information used to quantify computer memory or storage capacity. 1TiB = 1.099511627776 TB.
9 Where is netCDF used? https://www.unidata.ucar.edu/software/netcdf/usage.html

Table 1. Data stored in 4TU.Research Data as of February 2018 by the institution of the data depositor. Other institutions contributing netCDF data to 4TU.Research Data include Wageningen University and Research in the Netherlands, the Finnish Meteorological Institute, and Warsaw University of Life Sciences in Poland.

Institution	All Data		NetCDF Data	
	Number of Datasets	Size in TiB	Number of Datasets	Size in TiB
TU Delft	7,269	26.6	6,467	24.8
TU Eindhoven	118	2.72	24	2.65
University of Twente	41	0.077	0	0
Other	115	0.21	28	0.144
Total	7,543	29.61	6,519	27.59

In terms of volume, most of the netCDF data stored in 4TU.Research Data comes from one single experiment – the IRCTR Drizzle Radar (IDRA) installed on top of the Dutch meteorological observatory Cabauw Experimental Site for Atmospheric Research in the Netherlands (Otto and Russchenberg, 2014). The IDRA project, from TU Delft's Faculty of Civil Engineering and Geosciences, contributed with 2,325 datasets to date, corresponding to 24.7 TiB of data. IDRA data is a growing time series of datasets, updated every few months from 2009 onwards – and part of the reason 4TU.Research Data ended up with over 90% netCDF data in volume.

NetCDF Services

For all netCDF datasets, besides the usual download, 4TU.Research Data has offered OPeNDAP (Open-source Project for a Network Data Access Protocol) access since 2011. This is a protocol that "provides a discipline-neutral means of requesting and providing data across the World Wide Web"[10]. Some of the benefits of using this protocol include: viewing internal metadata hidden in the data files without having to download the files; accessing slices and subsamples of datasets without having to download the full datasets; access to data with APIs (Application Programming Interfaces) for Java, Python, MATLAB and other programming languages commonly used by researchers.

For large series of datasets, such as the IDRA collection[11], 4TU.Research Data offers the option of making custom agreements about data ingestion, data aggregation and metadata enrichment. Data conversion services have been also occasionally provided. For example, in 2011, after 4TU.Research Data fully embraced the netCDF format, data from the DARELUX (Data Archiving River Environment LUXemburg) project[12] were converted first to xml and then to netCDF. Conversion to xml ensured that metadata was easy to add to the files and conversion to netCDF increased the ways the data could be interacted with (via the OPeNDAP protocol)[13].

10 About OPeNDAP: https://www.opendap.org/about
11 Atmospheric observations – IDRA, Cabauw: https://data.4tu.nl/repository/collection:cabauw
12 Darelux – River Environment Luxemburg: https://data.4tu.nl/repository/collection:darelux
13 Researchers about 4TU.ResearchData: http://researchdata.4tu.nl/fileadmin/editor_upload/Brochure/Brochure__3TU.Datacentrum_2014.pdf

For the Sand Motor project[14] (Stive et al., 2013), 4TU.Research Data collaborated with Deltares[15] to create an online environment called OpenEarth DataLab[16] – a single place online where active research data could be stored, shared, edited, processed and visualised. At the end of the project and upon closure of the DataLab, selected netCDF data were transferred to 4TU.Research Data's OPeNDAP server (Rijkswaterstaat, 2017).

From Passive to Active, From Generic to Focused

We focus on the netCDF format, where 4TU.Research Data has built considerable experience and expertise, to explore the question of how an institutional and generic archive can remain relevant in a constantly and rapidly evolving research data management landscape.

Driver to Reinforce the NetCDF Services Provided by 4TU.Research Data

A review of 4TU.Research Data in the context of the FAIR principles (Dunning, 2017) provides one of the drivers to reinforce the archive's netCDF services. This review focused on the metadata that describes each dataset rather than the data sitting within each dataset. According to this analysis, 4TU.Research Data is strong in the areas of findability and accessibility, but less so when it comes to interoperability and re-usability. The data are openly available and easy to find and access, but active data re-use is not yet a reality[17].

For data to be re-usable, the FAIR principles require that "(meta)data meet domain-relevant community standards"[18]. As noted by Dunning (2017), it is difficult for 4TU.Research Data to have subject-specific metadata when it covers many different subjects. However, focusing on a specific data format, such as netCDF, which is mostly used by a few very specific communities, could help overcome this problem. To be re-usable, the FAIR principles also require that "(meta)data are associated with their provenance." This is another principle that 4TU.Research Data struggles to meet, but where netCDF data could be of help, because netCDF files are self-describing and include information about provenance.

What Services and Improvements Could We Offer in the Future?

We are considering how to implement the Recommendations of the Research Data Alliance (RDA) Working Group on Dynamic Data Citation (WG-DC)[19]. The WG-DC recommends that persistent identifiers are generated for every query that results in the creation of a subset dataset. This makes the exact queries citable (for example, in scientific publications) and allows for the exact same subset datasets to be downloaded and re-used by others. This feature seems particularly useful for 4TU.Research Data,

14 The Sand Motor: http://www.dezandmotor.nl/en/
15 Deltares: https://www.deltares.nl/en/
16 OpenEarth DataLab: https://publicwiki.deltares.nl/display/OET/OpenEarth+DataLab
17 From Passive to Active – The Future of 4TU.Centre for Research Data: https://openworking.wordpress.com/2017/10/15/from-passive-to-active-the-future-of-4tu-centre-for-research-data/
18 The FAIR Principles: https://www.force11.org/group/fairgroup/fairprinciples
19 RDA Data Citation WG: https://www.rd-alliance.org/groups/data-citation-wg.html

given that users can already make subset selection of datasets by using the OPeNDAP protocol. The ability to assign persistent identifiers to these queries provides an attractive possibility for making the references to datasets more accurate and research results more easily reproducible. Similar implementation of the WG-DC recommendations at the Climate Change Center Austria, which is another netCDF data repository, proved to be successful[20].

Other services we are evaluating include further technical services, such as data visualization, data processing and data mining, as well as providing training and guidance. Conversion of files to netCDF as an 'afterthought' at the end of a project is not practical and scalable. It would be much better to train early career researchers to produce netCDF files so they can enjoy the benefits of the format from an earlier stage of the research cycle. 4TU.Research Data could also provide templates that ease the production of netCDF files and the inclusion of appropriate metadata. Developing and providing some of these services may require stronger community engagement and the building of a user community.

Community Engagement – Progress So Far and Plans for the Future

To assess what expanded netCDF services 4TU.Research Data could usefully and efficiently provide, and to ensure that any new and current netCDF services continue to be relevant and evolve simultaneously with the needs of the research community, we have initiated and plan to maintain active engagement with data depositors and re-users. So far, we have started interviewing TU Delft researchers, mainly data depositors, who produce and use netCDF files. We will continue these interviews and will also contact researchers at other institutions, starting with those who have deposited data in 4TU.Research Data. Based on the results of these interviews, we will organize a user requirements gathering workshop later this year.

The interviews we have conducted so far focused on the researchers' use of netCDF, data management and archiving practices, choice of archive and training needs. We contacted all TU Delft researchers who have deposited netCDF datasets in 4TU.Research Data. We are interested to know why these researchers use netCDF, whether it is the standard in their community, and whether they adhere to community standards (if any) when it comes to metadata. We are also interested in the researchers' training needs and to know if they have received any previous formal training in using and producing netCDF files. Because we are currently exclusively interviewing researchers who deposited data in 4TU.Research Data, we are interested to know why they chose the archive. Were there other suitable choices? Did the researchers consider subject-specific archives? Were any of the services on offer key to the decision of using 4TU.Resarch Data? What services did they value; what services were not important? What services would be useful to have? Did the researchers consider the needs of any potential data re-users? The results of these interviews have been published in October 2018 (Cruz and Gramsbergen, 2018).

20 'Implementing the RDA data citation recommendations by the Climate Change Centre Austria (CCCA) for a repository of NetCDF files' webinar: https://www.rd-alliance.org/implementing%C2%A0-rda-data-citation-recommendations-climate-change-centre-austria-ccca-repository-netcdf

References

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal 15*(6). doi:10.5334/dsj-2016-006

Cox, A.M., Kennan, M.A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology 68*: 2182–2200. doi:10.1002/asi.23781

Cruz, M. & Gramsbergen, E. (2018). NetCDF at the 4TU.Centre for Research Data. *Zenodo*. doi:10.5281/zenodo.1465950

Dunning, A. (2017). FAIR principles: Review in context of 4TU.ResearchData. Retrieved from https://docs.google.com/document/d/1JmLJXMv-1mMHpkQ082lM0p_7iM4z8k2sgoX73VecjLg/pub

Husen, S.E., de Wilde, Z.G., de Waard, A., Cousijn, H. (2017). Recommended versus certified repositories: Mind the gap. *Data Science Journal 16*, p.42. doi:10.5334/dsj-2017-042

Otto, T. & Russchenberg, H.W.J. (2014). High-resolution polarimetric X-band weather radar observations at the Cabauw Experimental Site for Atmospheric Research. *Geoscience Data Journal 1*, p.7. doi:10.1002/gdj3.5

Rijkswaterstaat; Provincie Zuid-Holland; EcoShape. (2017). Zandmotor data. TU Delft. Dataset. doi:10.4121/collection:zandmotor

Stive, M.J.F., de Schipper, M.A., Luijendijk, A.P., Aarninkhof, S.G.J., van Gelder-Maas, C., van Thiel de Vries, J.S.M., et al. (2013). A new alternative to saving our beaches from sea-level rise: The sand engine. *Journal of Coastal Research*, *29*(5): 1001-1008. doi:10.2112/JCOASTRES-D-13-00070.1

Wilkinson, M.A., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data. 3*, 160018. doi:10.1038/sdata.2016.18