BAMS In Box

Datagrams

Diagrammatic Metadata for Humans

Sara Morris and Taneil Uttal

ABSTRACT: Creation of metadata (data about data) takes many forms and has many standards, much of which are designed to provide information for computer algorithms to find, access, and distribute data rather than for how humans might ingest data information. The humans (engineers, technicians, operators, scientists, data managers) that are increasingly tasked with being the providers of standard scientific metadata by the data science community also have a critical need for a different kind of metadata: metadata that can be used in the field (often offline) that provide a detailed visual map of the pathway taken by the electronic signal from a measuring device to a finalized, quality controlled geophysical variable. Datagrams presented here have been developed to fill this requirement and are a user-friendly, information-rich, graphical format that outline, record, and detail the critical information and steps involved with origin, collection, dataflow, processing, and archiving of data. Datagrams are designed to provide critical information across engineering, maintenance, data processing, and scientific teams that might speak different languages but are all required to process and maintain the data or instrument. The essential components of datagrams developed for instruments operating at remote Arctic stations are described here, but of course the concept is applicable to any type of observing protocol in any location.

KEYWORDS: Data processing/distribution; Databases; In situ atmospheric observations; Data science

https://doi.org/10.1175/BAMS-D-21-0219.1

Corresponding author: Sara Morris, sara.morris@noaa.gov

Supplemental material: https://doi.org/10.1175/BAMS-D-21-0219.2, https://doi.org/10.1175/BAMS-D-21-0219.3

In final form 5 January 2022

AFFILIATIONS: Morris and Uttal—NOAA/Earth System Research Laboratories/Physical Sciences Laboratory, Boulder, Colorado

The life story of a data value for even the simplest quantity is often complicated. Atmospheric data are measured with a wide variety of sensors (thermometers, anemometers, barometers, gas analyzers, radiometers, spectroradiometers, radars and lidars, particle sensors) to quantify physically relevant variables (temperature, wind speed, pressure, concentrations/types of atmospheric gases, solar radiation, atmospheric irradiance over specific ranges of the electromagnetic spectrum, or the presence/extent of clouds and the size and shape of an ice crystal or droplet in a cloud). In virtually all cases, when instruments that are used to "measure" environmental quantities, the parameter of interest is not actually measured directly, but rather it is inferred from induced electrical signals (current, voltage, charge, or magnetic flux) produced by devices such as thermopiles, transducers, thermocouples, and thermistors. These measured signals must be converted to a physical quantity, usually with calibration factors that are instrument specific and the use of theoretical relationships between the actual measurement and the value of interest. Furthermore, primary physical variables (e.g., temperature, humidity, wind speeds) are often used to further calculate latent values such as turbulent heat fluxes. Datagrams are designed to document the life story of a data value from start to finish and provide a guide for humans to design, deploy, troubleshoot, repair, record, transmit, process, and archive data collected with measuring devices. Datagrams are intended as a type of visual readme file, where users can see and follow the flow the of the data value from its collection to its use scientifically (Guptill 1999; Nogueras-Iso et al. 2004; Pallickara et al. 2010).

The need for datagrams

Imagine that you have just arrived at a remote Arctic research station to start an exciting and adventurous hitch as an on-site technician: you will be responsible for the smooth operation of a dozen or more instruments that have been installed by many different research groups and institutions. The personnel that you are replacing need to leave on the same charter aircraft or vessels that you just arrived on; therefore, although tired from a long and arduous journey, you must go through rapid-fire training on the where, what, and who of each instrument suite. Also, imagine that this personnel change-over is taking place at temperatures of −40°C. Your mobile device loses battery power in minutes at those temperatures, so that you cannot take notes or pictures, and even if you know that a pencil is still the most reliable writing instrument in the Arctic, it is difficult to take notes with large mittens. You uneasily watch your predecessor depart and wonder if you know what is going on and if you will be able to keep everything operating based on your brief training, the sticky notes stuck to computers, and the bookshelf of manuals.

Imagine it is two weeks later and you have now become fairly comfortable with operations and you get an email from a scientist or engineer that is receiving the data at their warm, southerly home institution and they tell you that something is off with one of the several different types of broadband radiometers in a package of instruments designed to measure the separate incoming, outgoing, shortwave, longwave, direct, diffuse, and total components of the surface radiation energy budget. When the weather allows, you trudge out in the dark with

your headlamp to the radiation measurement platform, which is divided between a rooftop complex where some of the upward-looking sensors are mounted on a moving sun-tracking device and a nearby mast where other downward-looking sensors are mounted with booms holding sensors out over ground that should ideally stay undisturbed. The radiometers all look distressingly the same. Which one could it be that is registering the faulty measurement?

Now, imagine that you are responsible for data processing and data management and you are assigned the responsibility for processing data that are flowing from a remote site. You open the file that was created by a datalogger program and discover a "raw" data file with voltages, and there is no information on how to convert to scientific units, like W m^{-2}. Furthermore, there are many ancillary housekeeping variables and no clear documentation on what each set of values is without looking at the original logger program, which you do not have access to. The engineering technician that calibrated and deployed the instruments has retired recently. You know that it is important to know what individual instrument calibration factors and schedules are, but you cannot find the necessary numbers or dates anywhere, so you cannot do the necessary processing to accurately translate from raw voltages to a reliable W m^{-2} incoming/outgoing longwave, broadband radiometric product.

Envision yourself as a modeler and you are excited when a data portal or a data discovery mapping tool suggests a network of ground heat flux data in your region of interest that you can assimilate into a permafrost melt model. A colleague that does field work informs you that there are multiple measuring techniques and algorithms for inferring ground heat flux; interoperability of the values between sites (or over different time periods if sensors were changed) may be questionable. There goes your plan to create a near-real-time service product on permafrost melt. It is these kinds of real-life and reoccurring situations over many years that have inspired the concept of datagrams to minimize data loss and misuse from the many points of potential failure along the data transfer chain.

Anatomy of a datagram

The intended scope of a datagram is one datagram for each type of file collected, meaning that if one data file is output from an instrument platform, then one datagram would be created for that file. Similarly, datagrams can also be generated for each station, each experiment (see Fig. ES1 in the online supplemental material), each platform (see Fig. 1), or each instrument—the level of detail is up to the user and their needs and should be considered a long-term document that follows the station or experiment, meaning that if the station closes or the experiment ends, then an end date is documented. Datagrams will vary from application to application, but the primary components and information will be the same. Datagrams are intended to be public-facing documents that accompany a dataset, station, or instrument platform, but can contain expert information relevant to only the research team, so it is important to maintain an updated contact list at the top of the datagram if details need to be clarified for outside users.

To compose a datagram, the author will need to spend some time gathering the information from different sources and people to input into a datagram spreadsheet or other type of schema (see Table ES1). Note that this spreadsheet or schema will act as a historical document that will trace, for example, the instrument history of a measurement collected at a station, and if instruments are swapped out, then a new serial number and calibration coefficient can be documented here. It is therefore important to ensure that a datagram and its spreadsheet or schema remain intact and are updated in parallel to maintain a historical record of the measurements obtained. The front-facing datagram depicts the most recent configuration of a station/platform/experiment, while the spreadsheet or schema hosts the back-end historical details utilized for data processing and metadata storage. Upscaling the datagrams to cover details for an entire network of stations is doable through this spreadsheet or schema.

Ideally, in the future, a datagram could be generated utilizing the spreadsheet to build datagrams through a web interface, so that datagrams could be generated for large networks of instruments or stations. The main components of a datagram are exemplified in Fig. 1 (and Fig. ES1) and are described in the sections below.

Site panorama and contacts. The header of a datagram gives basic general information about the experiment or site. This includes the name of the experiment or station where the data are being collected, followed by the contact information for the principle investigator(s) or lead scientist(s), technician(s), research coordinator(s), and data manager(s). Logo(s) can be added here for institutions that are sponsoring and supporting the data collection. In the background of the header is a panoramic image of the landscape or environment where the data are being collected.

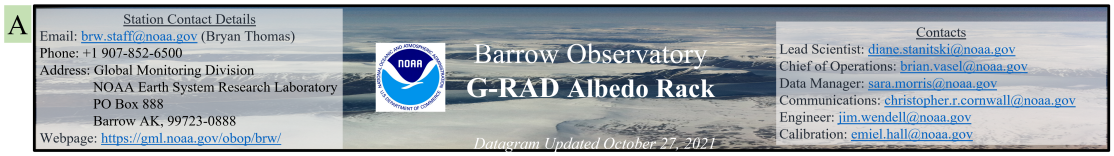
Site and facility information (photo and/or map). A detailed map indicating the location of instrument suites as they relate to the local geography or field station is imperative for supporting the quality control of the measurements in postprocessing. This map/photo also assists on-site technicians in locating instrument platforms and data acquisition systems. This map should include a legend when necessary and compass directions (specifically true north for Arctic regions). Other geographic details important to the region or measurements, like climate regime, should be included (i.e., wind direction for instrument installations, or travel routes/roads for postprocessing local pollution events).

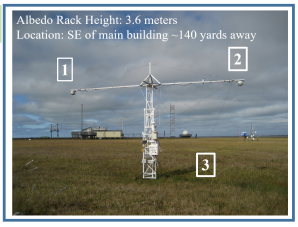
Instrument platform(s): Individual instruments annotated with reference labels. A photo and/or diagram of the instrument suite, including details of where each instrument is located on the platform, will assist any on-site technicians to know where instruments are located if they need to be serviced. This schematic identifies each instrument as it sits at the station and references a label that is detailed in the "Instruments details" section regarding instrument specifics. Additional information on instrument location, instrument power voltages, data acquisition services and hardware, and cameras, if they exist, should also be included in the schematic.

Network information. Outlining the data acquisition process is very important for troubleshooting dataflow issues if/when they occur. Some sensitive information needs to be redacted when a datagram is posted online (e.g., IP addresses, passwords, and network details), but these should be included in on-site digital versions on secure computers and hard copies being used in the field. It is also advantageous to include power details and the make/model of the acquisition equipment for troubleshooting, including links to equipment manuals. A flowchart of the data acquisition services can also be helpful to identify points of failure during data collection.

Individual instruments photos. It is important to have pictures of what each instrument looks like so that it can be identified by technicians and researchers on-site.

Instruments details. This section is the most detailed section of the datagram and includes pertinent information about the instrumentation and measurements collected. Details included are serial numbers, measurement type, calibration coefficients, calibration history, voltage output, response time, field notes, equipment/technical manuals (including links). Instrument details such as make, model, and manufacturer should also be included.





D

Data Acquisition

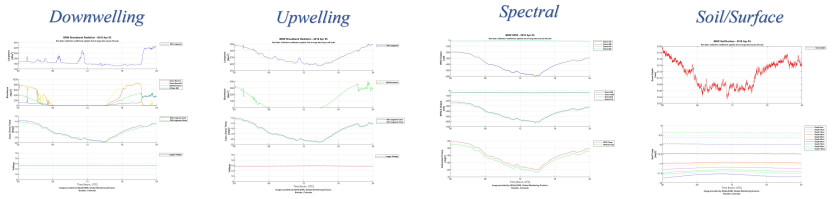
Type: CR1000 Logger
Power: 12 VDC
Comms: Moxa Nport serial to TCP/IP

IP: XXX.XXX.XX
IP: XXX.XXX.XX

F

	Measurement	Instrument Type	Manufacturer	Model	Serial Number (current)	Calibration Factor [μV/W/m²] (current)	Response Time	Sensitivity	Field Notes
1	Longwave Radiation	Precision Infrared Radiometer, "Pyrgeometer"	Eppley	PIR	32598	3.71 DCF: 3.187	1 second	~ 3 μV/Wm^{-2}	
2	Shortwave Radiation	Precision Spectral Pyranometer	Eppley	PSP	8040	6.22	1 second	~ 8 μV/Wm^{-2}	
3	Soil Thermal Conductivity	Thermistor Soil Probe	In-house	YSI 44031	n/a	n/a	1 second	n/a	Depth 5-120 cm

G On-Site Visualization Examples



Data Visualization Details
Computer producing plots located in Boulder, CO at NOAA DSRC Global Radiation and Aerosol Group

Daily Plots:
https://gml.noaa.gov/dv/iadv/graph.php?code=BRW&program=g-rad&type=dr

H Output File Details

KEY:

BATT_V: Battery Voltage
UWSW_PSP: upwelling shortwave
std: standard deviation
UWLW_PIR: upwelling longwave
fan: aspirated housing fan
AirTemp: air temperature

SnDepth: snow depth
Tempfan: air temperature aspirated housing fan
Depth [5-120]cm: thermistor depth recording
UWLW_PIRcase: case temp for upwelling longwave
UWLW_PIRdome: dome temp for upwelling longwave
Panel_T: logger panel temperature

	File Name	Start Date Time of Header	End Date Time of Header	Line ID	Day Of Year	Hour Minute																															
Previous Headers	BRWA	1/1/17 0:00	9/21/17 12:21	101	Year	DOY	hhmm	BATT_V	UWSW_PSP	UWSW_PSPstd	UWLW_PIR	UWLW_PIRstd	UWLW_PIRcase	UWLW_PIRdome	UWLWfan	UWSWfan																					
				301	Year	DOY	hhmm	AirTemp_OLD	SnDepth	Tempfan																													
				311	Year	DOY	hhmm	Depth5cm	Depth10cm	Depth15cm	Depth20cm	Depth25cm	Depth30cm	Depth45cm	Depth70cm	Depth95cm	Depth120cm																					
	gradobs.brw-alb	9/21/17 12:22	2/22/18 16:05	101	Year	DOY	hhmm	BATT_V	UWSW_PSP	UWSW_PSPstd	UWLW_PIR	UWLW_PIRstd	UWLW_PIRcase	UWLW_PIRdome	UWLWfan	UWSWfan																					
				301	Year	DOY	hhmm	AirTemp_OLD	SnDepth	Tempfan																													
				311	Year	DOY	hhmm	Depth5cm	Depth10cm	Depth15cm	Depth20cm	Depth25cm	Depth30cm	Depth45cm	Depth70cm	Depth95cm	Depth120cm																					
	gradobs.brw-alb	2/22/18 16:08	9/20/18 22:07	101	Year	DOY	hhmm	BATT_V	UWSW_PSP	UWSW_PSPstd	UWLW_PIR	UWLW_PIRstd	UWLW_PIRcase	UWLW_PIRdome	UWLWfan	UWSWfan																					
				301	Year	DOY	hhmm	SnDepth																															
				311	Year	DOY	hhmm	Depth5cm	Depth10cm	Depth15cm	Depth20cm	Depth25cm	Depth30cm	Depth45cm	Depth70cm	Depth95cm	Depth120cm																					
Current Header	gradobs.brw-alb	9/20/18 22:11		102	Year	DOY	hhmm	BATT_V	Panel_T	UWSW_PSP	UWSW_PSPstd	UWLW_PIR	UWLW_PIRstd	UWLW_PIRcase	UWLW_PIRdome	UWLWfan	UWSWfan	SnDepth	Depth5cm	Depth10cm	Depth15cm	Depth20cm	Depth 25cm	Depth 30cm	Depth4 5cm	Depth70 cm	Depth95 cm	Depth1 20cm
								V	*degC*	*mV*		*mV*		*mV*	*mV*	*RPM*	*RPM*	*mm*	*degC*	*degC*	*degC*	*degC*	*degC*	*degC*	*degC*	*degC*	*degC*	*degC*
								UNITS:																													

I Ingest Data Processing

Processing Conversions:

Shortwave Radiation (units: W/m^2)
Current Calibration Coefficient: 6.22 μV/W/m^2
SW = 1000 * UWSW_PSP / 6.22

Longwave Radiation (units: W/m^2)
Current Calibration Coefficient: 3.71 μV/W/m^2
Dome Correction Factor: 3.187
Sigma = 5.6704e-8
Emissivity = 1
A = 0.0010295
B = 0.0002391
C = 0.0000001568

Processing Conversions Continued:

Longwave Radiation Continued

LW_case (units: mV) = 1 / (0.0010295 + 0.0002391 * ln(UWLW_PIRcase * 1000) + 0.0000001568 * ln(UWLW_PIRcase * 1000)^3)

LW_dome (units: mV) = 1 / (0.0010295 + 0.0002391 * ln(UWLW_PIRdome * 1000) + 0.0000001568 * ln(UWLW_PIRdome * 1000)^3)

Convert Calibration Coefficient (units: mV/W/m^2): (1 / 3.71) * 1000 = 269.542

LW = 269.542 * UWLW_PIR + 5.6704e-8(1(**LW_case**^{4}) + 3.187 ((**LW_case**^{4}) - (**LW_dome**^{4}))

J Data Archives

Home: https://www.esrl.noaa.gov/gmd/grad/
Data: https://gml.noaa.gov/dv/data/?site=brw
Webpage: https://gml.noaa.gov/obop/brw/
BSRN Archive: https://bsrn.awi.de/

Fig. 1. Example datagram for albedo rack data collected at Barrow Station in Utqiaġvik, Alaska. Notice that datagrams do not need to follow the exact schematic presented, and can vary to include more or less information.

On-site visualization. Ideally, any instrument operating in the field has software tools for visualizing the data on-site as it is collected. Information is provided for locating and accessing these visualizations, and if appropriate, instructions for interpreting them and identifying problems such as out-of-range values or engineering issues such as overheating. Screenshots of visualizations (i.e., plots, figures) support associating the visualization with the instrument being monitored if technicians do not have immediate access to a manual or direct communication with the researcher.

Raw output file structure and naming convention. This is a snapshot of a raw data file format detailing variable name and units. If the variable name has been abbreviated or coded in the raw data file, it may be necessary to have an ancillary key (see Fig. 1). It is important to also include the directory path where the data are stored locally as well as the file naming convention. If data are hosted in several locations, those directory paths should also be listed. Several lines of the outcoming data can also be shown as an example below the header information.

Ingest data file structure and naming format. Ingest data are distinguished from raw data by processing to convert (if necessary) from analog signals (e.g., voltages like mV) to a physical variable (e.g., irradiance). Included here are details about conversion calculations, internal constants, and any quality-control techniques used to preprocess the data. It is important to list any publications that demonstrate the type of quality control or calculation techniques used. Frequently in the process of converting from raw data to ingest data, various housekeeping occurs to take care of time stamp irregularities so that data can be divided into regular increments (usually hourly, daily and/or monthly). It should be noted that ingested data are not yet quality controlled for outliers or are manually curated.

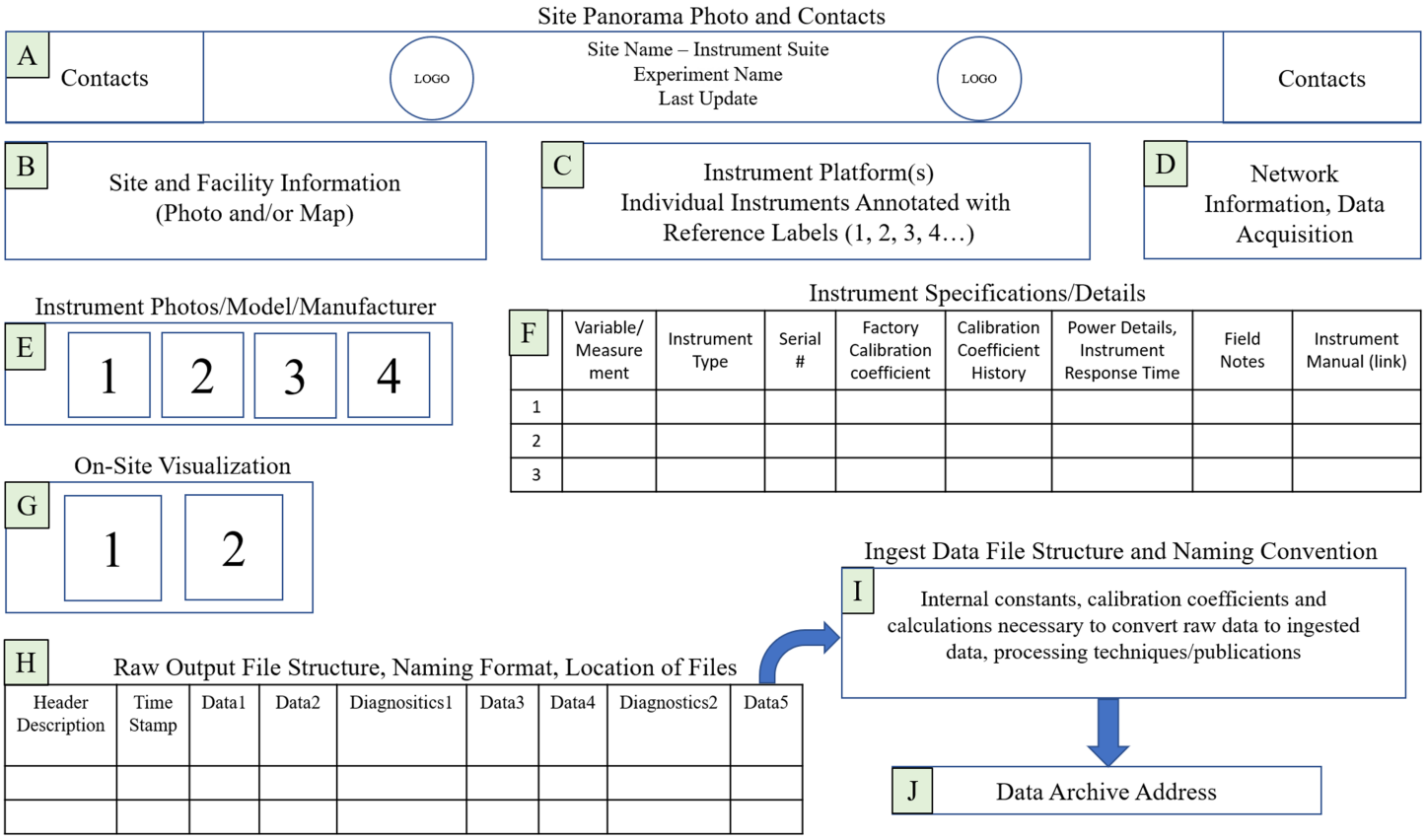


Fig. 2. Essential components of a datagram (A–J details are described above in text).

Data archive address. Data archive web addresses where the data have been submitted are listed, including archives, portals, and host FTP pages. It is also important to list where the data are being hosted by the principle investigator.

Figure 1 and Fig. ES1 are examples of datagrams generated for data collected by atmospheric instruments, but it is important to note that datagrams can also be utilized to organize data outputs from ecological sampling, ocean profiles, snow samples, subsurface data, and beyond. A generalized example to the essential components schematized in Fig. 1 can be found in Fig. 2, as a guideline to develop datagrams for any type of scientific data collected in the field.

Summary

In recent years, data science publications in journals such as *Earth System Science Data* (www.earth-system-science-data.net/) have become increasingly popular ways to comprehensively document details of the measurement site, the instruments, the measurement and analysis practices, and the resulting variables and, most importantly, to provide access and a reference to an observation-based dataset (Boike et al. 2019; Maturilli et al. 2013). While it might appear that the information provided by a datagram is redundant of a data science paper, a datagram is distinctly different in several aspects, including that it exists during deployment, it likely evolves as data are gathered, it can be viewed in poster format with no text explanations, it can be printed and thus accessed when off grid, it can be used as an engineering design tool, and it is a living document that can be updated as frequently as is necessary. Datagrams fill the gap between machine-oriented metadata and data science articles by providing metadata useful to a wider variety of specialties, like technicians, post-docs, engineers, data managers, and researchers alike—anyone who physically or virtually comes in contact with the instrument or the data that are collected.

Acknowledgments. We thank Christopher J. Cox, Jim Wendell, Emiel Hall, Chris Cornwall, Diane Stanitski, Brian Vasel, Allison McComiskey, Bryan Thomas, Ross Burgener, Christine Smith, and Leslie Hartten. We also appreciate the productive comments from the reviewers. This research was partially supported by the NOAA Physical Sciences Laboratory and the Global Monitoring Laboratory, and was also supported by NOAA's Global Ocean Monitoring and Observing Program/Arctic Research Program.

References

Boike, J., and Coauthors, 2019: A 16-year record (2002–2017) of permafrost, active-layer, and meteorological conditions at the Samoylov Island Arctic permafrost research site, Lena River delta, northern Siberia: An opportunity to validate remote-sensing data and land surface, snow, and permafrost models. *Earth Syst. Sci. Data*, **11**, 261–299, https://doi.org/10.5194/essd-11-261-2019.

Guptill, S. C., 1999: Principles and technical issues. *Geographical Information Systems*, 2nd ed. John Wiley & Sons, 677–692.

Maturilli, M., A. Herber, and G. Konig-Langlo, 2013: Climatology and time series of surface meteorology in Ny-Ålesund, Svalbard. *Earth Syst. Sci. Data*, **5**, 155–163, https://doi.org/10.5194/essd-5-155-2013.

Nogueras-Iso, J., F. J. Zarazaga-Soria, J. Lacasta, R. Bejar, and P. R. Muro-Medrano, 2004: Metadata standard interoperability: Application in the geographic information domain. *Comput. Environ. Urban Syst.*, **28**, 611–634, https://doi.org/10.1016/j.compenvurbsys.2003.12.004.

Pallickara, S. L., S. Pallickara, M. Zupanski, and S. Sullivan, 2010: Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections. *IEEE Second Int. Conf. on Cloud Computing Technology and Science*, Indianapolis, IN, IEEE, 573–580, https://doi.org/10.1109/CloudCom.2010.99.