Keeping Track of Samples in Multidisciplinary Fieldwork

RESEARCH PAPER

PÅL GUNNAR ELLINGSEN
LARA FERRIGHI
ØYSTEIN GODØY
TOVE MARGRETHE GABRIELSEN
*Author affiliations can be found in the back matter of this article

ABSTRACT

We here present methods, tools and results for efficiently collecting metadata and tracking samples collected in the field. The samples and metadata were collected during scientific cruises conducted by amongst others marine biologists, oceanographers, geochemists and marine geologists in the Nansen Legacy project. It is here reported on the successful development and implementation of a system for labeling, tracking and openly publishing metadata from the cruises. The results and tools have been made openly available, as they are suitable for a range of situations, from the individual scientist working in the field to large research missions.

CORRESPONDING AUTHOR:
Pål Gunnar Ellingsen
Department of Electrical Engineering, UiT The Arctic University of Norway, 8505 Narvik, Norway; Department of Arctic Biology, UNIS – University Centre in Svalbard, 9170 Longyearbyen, Norway
pal.g.ellingsen@uit.no

KEYWORDS:
Data managment; Sample tracking; Metadata; Darwin Core; CF Standard Names

TO CITE THIS ARTICLE:
Ellingsen, PG, Ferrighi, L, Godøy, Ø and Gabrielsen, TM. 2021. Keeping Track of Samples in Multidisciplinary Fieldwork. *Data Science Journal*, 20: 34, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-034*

INTRODUCTION

The EU directive on Public Sector Information was replaced by the Open Data Directive in 2019 (European Commission 2019). This is in line with recent developments at the national level (e.g. in Norway) where there is an increasing focus on how to ensure that data collected by scientists is made available to other research groups and the public (for instance GO FAIR 2020). The directive does not specifically identify how to share data, but there has been an increasing focus in recent years on standardised documentation of, and interfaces to data. These efforts ensure that data are easily accessed and reused across communities without detailed personal knowledge on the data production/collection. The core principles needed to achieve the access and reuse are outlined in the FAIR guiding principles (Wilkinson et al. 2016). An exemplification of these principles is the global data exchange managed by the World Meteorological Organisation (WMO) which achieved operational exchange of meteorological, hydrological and oceanographic data even during the cold war. The foundation for the success of WMO was harmonised encoding of data regardless of the data provider. Every consumer of the data knew what was measured, when it was measured and how it was measured. This principle has since evolved (on its own) in many different communities within a broad range of fields. Noteworthy examples here are the Global Biodiversity Information Facility (GBIF 2020) and Ocean Biodiversity Information System (OBIS) (UNESCO 2020) providing large datasets and networks on biodiversity.

The work presented here is part of an ongoing project named The Nansen Legacy, which is a Norwegian project with more than 100 participants from ten institutions. Participants in the project come from a broad range of fields, including marine biology, oceanography, meteorology, geology, geochemistry, and they are accustomed to handling data in different ways. The Nansen Legacy constitutes an integrated Arctic perspective on climate and ecosystem change, from physical processes to living resources, and from understanding the past to predicting the future. The data produced and published by the project follows the FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al. 2016) guiding principles. Additionally the project aims at publishing the metadata of data collected during a cruise online shortly after the cruise finishes and before processing the data. Within these metadata, information on what was sampled/recorded, where and when it was recorded and by whom should be captured. The primary purpose of these metadata is to improve internal communication within the project in preparation of publishing the full datasets, but the information is also made publicly available allowing external scientists to follow the activities of the project. To allow the event logging to be a practical tool for the project, it became necessary to develop a standardised way of recording this information. Since no system supporting the required workflow was identified existing standards and approaches were adapted to serve the needs of the project. Recently (Damerow et al. 2021) proposed a similar approach for tracking samples in multidisciplinary ecosystem sciences, presenting an exhaustive review of the different methods used to handle sample identifiers and related metadata practices. Our work, which has been running in parallel, has focused on marine sciences, while theirs have covered terrestrial ecosystem fieldwork.

BACKGROUND

Metadata, data about data, is well known to be of significant importance (Damerow et al. 2021) and can be mined for information (Palacios-Abrantes et al. 2019). It is needed in the process of publishing the recorded data to for instance (GBIF 2020), OBIS (UNESCO 2020) and SeaDataNet (Pecci et al. 2020, Schaap and Lowry 2010). Commonly individual scientists have recorded the metadata in their field journal and then recorded them on a computer together with the data when back from the field. This works for one scientist, but with increasing international collaboration (Wagner et al. 2017) and the corresponding exchange of information between scientific communities, there is a growing need to further develop existing standards of metadata. This collaboration is known to lead to a greater need for FAIR datasets (McNutt et al. 2016, Stall et al. 2019).

There are many types of metadata and in the context of the Nansen Legacy, event, discovery and use metadata are of particular interest. This implies that the project has a need to record information about events producing data, document these data in a manner suitable for making data Findable once published and finally ensure that users accessing the data are able

to understand and interpret them. These categories are not independent and in the following we provide some background information used when developing the chosen approach. The approach applied aims to ensure that proper documentation is captured when documenting the events that produce data, and to simplify the process of preparing discovery and use metadata once the datasets are prepared for publication.

Initial work concentrated on evaluation of standards for discovery metadata and in particular standards that are embedded with the actual data (are linked to a use metadata standard). Two approaches covering this aspect are commonly used for the disciplines that the Nansen Legacy covers, that is the Attribute Convention for Dataset Discovery (ACDD 2014) and Darwin Core. ACDD is a discovery metadata standard that is closely linked to NetCDF and the Climate and Forecast Conventions (CF) (Lawrence et al. 2006, Eaton et al. 2020), while Darwin Core is integrated in Darwin Core Archive. Both capture discovery level information in the data files and also addresses the need for use information.

ACDD (Attribute Conventions Data Discovery) defines keywords and how to use these to capture discovery information for the data. It covers the basic requirements for discovery metadata, but lacks the specificity required to provide details on the data collection process and the content of the data. For the latter issue this is covered by the CF-conventions. The CF-conventions describe how to organise and annotate the data in order to ensure that humans and computers are able to understand them. It addresses the need to properly identify variables, units of variables, how missing values are encoded, whether data values are spatially or temporally averaged etc. A key component of the CF conventions is the CF Standard Names (CF Standard Names 2020, Lawrence et al. 2006) which enables a semantic framework for describing the variables contained in a dataset|. CF is currently usually used in the context of the NetCDF (Unidata 2020) file format, although the principles, as outlined in the CF data model, could be implemented in other formats (e.g,. GeoJSON (Butler et al. 2016)).

While the CF conventions focus on geoscientific (meteorology, oceanography, glaciology etc) information, a similar approach serving a different community (biodiversity) is provided by the Darwin Core Archive (DwC-A) (Wieczorek et al. 2012). This approach also focus on a semantic framework for description of the content, but relies on a spreadsheets as carrier (container) for the data. A DwC-A dataset is a ZIP-file containing a number of spreadsheets (comma separated files) and a metadata file (Ecological Metadata Language – EML). One of most impressive employment of Darwin Core is within the GBIF (2020), where billions of recorded biological occurrences are catalogued.

While both these metadata standards are governed by active groups (community driven), with good management structures in place and well rooted in the respective scientific communities the primary challenge is to ensure that scientists are adopting and utilizing them. The threshold to navigate and learn semantically structured standards like these is quite high for most scientists, thus there is a need for tools simplifying the workload. Furthermore, both ACDD/CF and DwC-A are approaches for publishing data, but they are not directly suitable to capture events (that initiates the creation of a dataset) during field campaigns/cruises. This is a very critical step in the scientific workflow/data life cycle since observations are collected and annotated with metadata. If done properly, subsequent publishing of data (a requirement by most funding agencies) is simplified. When working in the field, scientist have traditionally recorded sampling information in some form of logbook. Often the samples themself have been labeled with handwritten labels following a system which is often personal (unique) and rarely reproducible. This method, which to our experience is employed in more or less an organised fashion, works satisfactorily only in smaller groups.

In larger projects where several participants are recording metadata, some form of standardised approach is required, as also noted by Damerow et al. 2021. If great care is not taken, interpretation of standardised approaches will often diverge through individual adaptations. In this context, individual adaptations are sources for errors and misunderstandings. How severe these mistakes can be, depends on what was recorded, the quality control procedures and the longevity of the metadata. For instance, recording time in local time instead of UTC can be considered a marginal difference by individuals, but can cause huge problems in subsequent analysis of the data by other scientists. Similarly, non-standardised recording of dates can result in a record being misinterpreted by months. In order to ensure that sampling performed in the

context of the Nansen Legacy is consistent across cruises/field activities, the Nansen Legacy has developed sampling protocols (The Nansen Legacy 2020, Version 1, Version 2, Version 3, Version 4.2, Version 5, Version 6), i.e. descriptions on how to perform various sampling activities. This ensures that observations are comparable within activities of different partners, but also in time during the project. These sampling protocols combined with standardised event logging are crucial to ensure reproducible research.

This paper presents the approach and toolkit developed and implemented by the Nansen Legacy in an effort to solve many of these challenges. Furthermore it explores methods for tracking large amounts of samples and their metadata in a field setting, with a strong focus on standardisation of information within the biodiversity domain. Herein there will be a focus on the traceability between physical and digital representation of samples.

APPROACH

This section will present the approach by addressing the following issues:

1. how to track samples
2. how to ensure that the necessary metadata are recorded during a cruise
3. how to minimise the additional workload of logging on scientists

These challenges are highly interlinked, see ***Figure 1*** and need to be addressed in a complete framework. Software and data sources are covered in sections after this section for clarity.

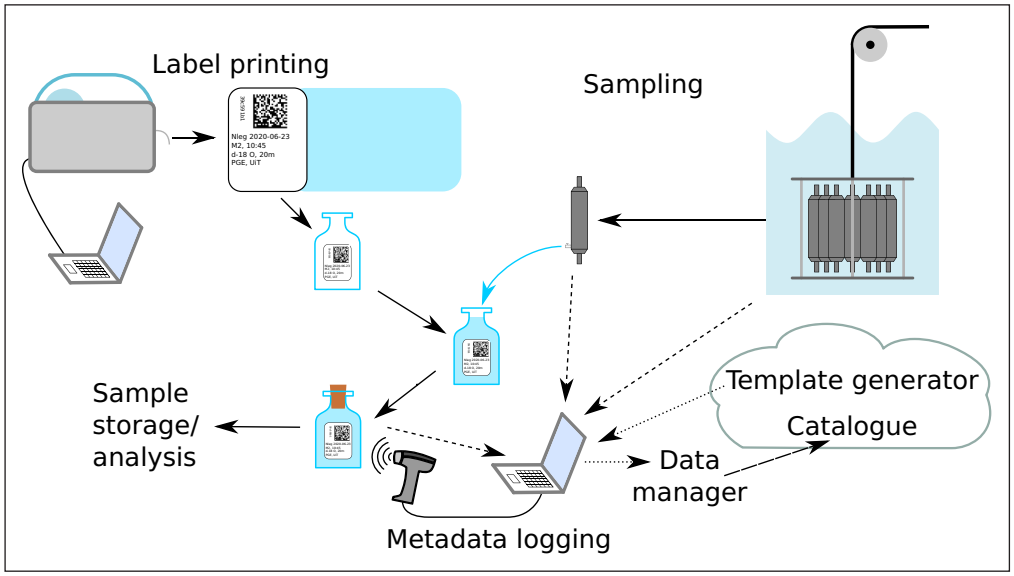

Figure 1 An example workflow on the cruise for logging the metadata, from label printing, to sampling and metadata recording. The steps are not numbered as the timing of the different steps vary between different users. The catalogue webpage allows the scientists to search the metadata. Details on the labels can be found in Figure 2 and the methodology for IDs can be found in Figure 4.

TRACKING SAMPLES

In the planning phase of the Nansen Legacy project, it was apparent that in order for several groups to be able to compare results and share samples, there needed to be a way of tracking the samples, labeling the samples consistently and logging the metadata from them. For tracking samples it was decided to follow the same approach as the parcel delivery industry, by using standardised printed labels. Printing is done using network connected label printers (Zebra GX430t), where the print job is generated by custom written webpage (hosted on a virtual machine on the ship) or a stand alone program. There are several advantages in using a webpage. One is that several different sized labels can be used, each with a dedicated printer, tracked by the web server. In this project, two different label sizes (25 × 50 mm and 19 × 25 mm) were used, each with its own printer. Another advantage is that the printing is made independent of computer drivers and operating system. This results in a lower threshold for using the system and significantly reduces the technical problems. Another advantage is that any needed updates are easily deployed.

The content on the labels were organised with mandatory data matrix (DM) 2D barcode (Plain-Jones 1995) and some text lines for the scientist to write information, see ***Figure 2***. The DM encodes a uniquely generated UUID (Universally Unique IDentifier) number in the form of a 36 character hex string (including the four -), see decoded label in ***Figure 2***. For control the first 8 hex characters is printed next to the DM code as shown in the figure. In the case of tiny samples/containers, a label with only the UUID DM code are used, see the small labels in the figure. DM was chosen over QR codes due to the slightly higher error correction and the support for physically smaller codes, important for the small labels used in the project. The error correction was important for the project as work was carried out all year round at sea and on sea ice in the Arctic, sometimes in bad weather, resulting in damaged labels.

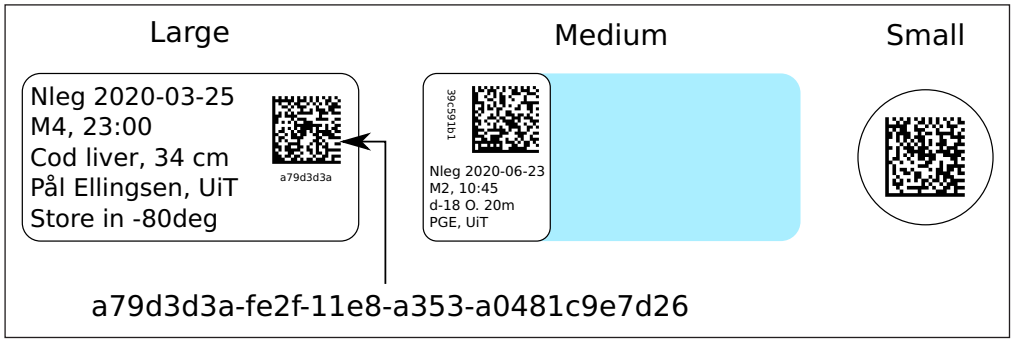

Figure 2 Examples of the labels in use, classified by their size: Large 25 × 50 mm, medium 19 × 25 mm (excluding the wrap around plastic) and small 10 mm diameter. The DM code in the Large label is decoded into the UUID below it.

The UUID code ensures that every sample has a unique identifier, that statistically is expected to be unique throughout the world. As it is not practical to write by hand, the encoding of it into a DM code, enables it to be read into a computer using a dedicated 2D barcode reader (in the project Opticon OPI-3601 USB readers were used) or a smartphone. Dedicated barcode readers are convenient, as they can be used to directly input data in a spreadsheet field or search bar, while a smartphone is slower, but more readily available. Since the DM codes carry error correcting parts, it is a robust way of uniquely identifying the sample. It should be noted that the DM code does not contain any information about the sample (metadata). The advantages of this will be discussed below.

The choice of UUIDs over alternatives (Guralnick et al. 2014, Damerow et al. 2021), was due to several considerations. It can easily be generated on any computer and it works as an id in databases, enabling the id to follow the sample from the sample database and into the published data. Moreover, appropriate versions of UUID (for example version-1) contains the time at which the identifier was generated, thus allowing for additional checking of the metadata records (an UUID produced a year later then the sample was recorded could raise a warning flag).

LOGGING OF METADATA

To facilitate the logging of sampling metadata during a cruise, it was, after considerable discussion within the project data management group, decided that the best would be to use spreadsheets. Spreadsheets were chosen as they are widely used by scientists, can be edited without specialised software or network connection, and are machine readable, given standardised fields and layout. Standardisation was achieved by generating custom tailored templates, based on standard names and cell validation. With over a hundred scientists, and a goal of making this useful for other types of field recordings, it was decided to write a template generator usable by the scientists to generate their own template. The generator generates the spreadsheet with metadata cells (fields) locked down to certain data formats.

As the majority of samples collected would be from marine biology, it was decided to base the template generator on Darwin Core standard names. The generator was based on the DwC Excel Template Generator (GBIF Norway 2020) developed by GBIF Norway (UiO Naturhistorisk museum 2020) for Darwin Core. The use of Darwin Core for the standard names, solved some of the challenges, though for a lot of samples and data, the parameters and characteristics that need to be recorded in the metadata were not part of Darwin Core. For these it was decided to use the CF standard names (CF Standard Names 2020), but written on the form of Darwin

Core. The original CF format was kept as a keyword in the Python dictionary used to define the names, ensuring easy conversion back to CF in the event of data being produced as NetCDF files. For names that did not fit any of these standard names (examples are bottle number for the rosette and number of endoparasites), the Darwin Core naming format was used to create new names. These were created though discussions with the relevant scientists within the project.

One of the challenges with Darwin Core, is that it sets weak limits on the parameter inputs. This makes it prone to invalid inputs, like a time of 25:00 and a latitude of 92 degrees. To limit the possibility of such errors, input values were restricted by defining input types (integer, float, date, time, list or string) and limits (time between 00:00.0 and <24:00 etc). Spreadsheets support these through cell data validation properties. After filling out a spreadsheet, the values were rechecked in the spreadsheet checker, catching mistakes made by overwriting the validation (one weakness of spreadsheets). On the cruises the checker was available from the onboard web server. Further checks were made when the metadata was imported into the central database (covered later) used to track all the samples. With respect to the Darwin Core standard, the *eventDate* parameter which supports a full ISO 8601 date, was split into a *eventTime* parameter (only time in UTC) and *eventDate* parameter (only date in format *YYYY-MM-DD*), allowing us to lock down the time and date values in the spreadsheet. When using the metadata in a Darwin Core data product these can easily be joined into a valid *eventDate*.

The spreadsheet generator was written in Python and runs on a standard web server as a two part page, see ***Figure 3*** where the left side displays available choices, grouped by similarity and level of requirement. The right side displays the documentation for the parameter under the cursor. This includes its definition, unit and validity. For some of the parameters additional information was added on top of the Darwin Core definition. This could be for instance its limits. The page loads with the required parameters selected, determined by a schema selectable via drop down menu. Configuration of which parameters are available, required, groupings etc. is done in a configuration *yaml* file, making it easy to tailor this to a give project.

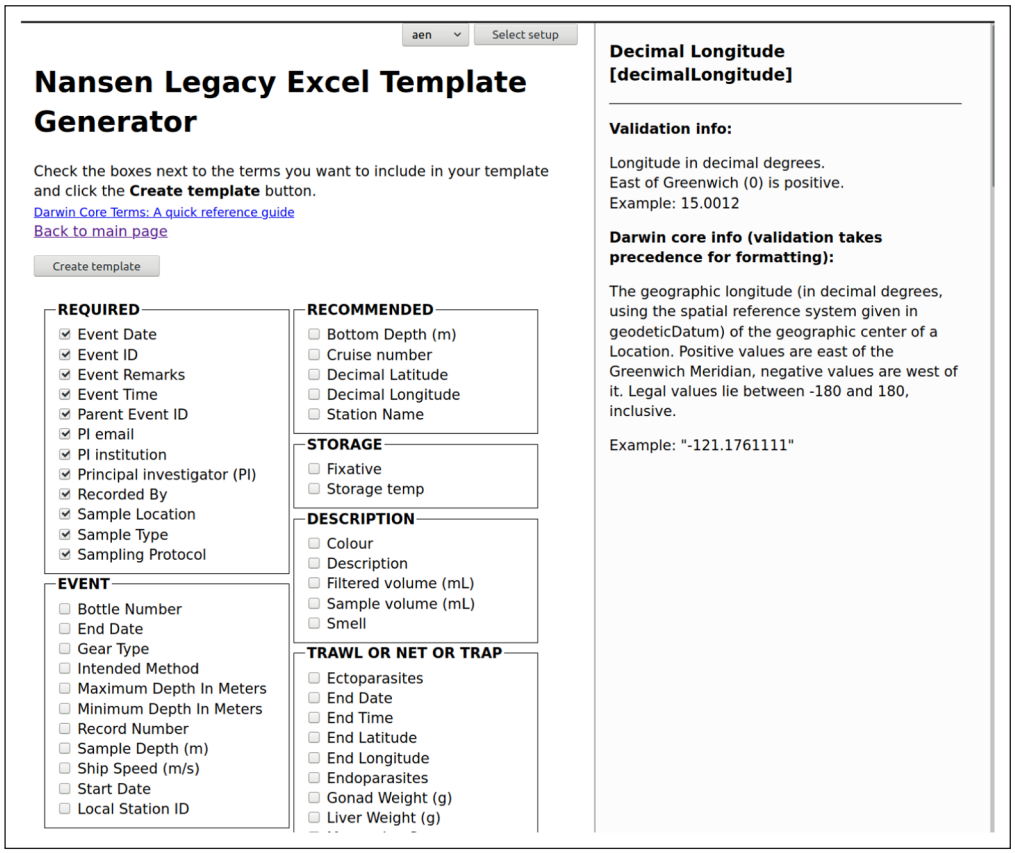

Figure 3 The front end of the template generator as set up for the Nansen Legacy project (selected by the dropdown at the to), running on the SIOS page *https://www.sios-svalbard.org/cgi-bin/darwinsheet/?setup=aen*. Only some of the possible parameters are show, with the additional information for the *Decimal Longitude* parameter shown on the right.

Once the desired parameters are selected, the spreadsheet can be generated, resulting in a spreadsheet (.xlsx file) which can be opened by most modern spreadsheet programs (Microsoft Excel, LibreOffice, etc.). It contains one sheet containing high level metadata (title, project,

responsible person, e-mail) which is relevant to all the samples logged, and one sheet for the sample log. The scientist is now ready to log the samples taken (either physical samples or samples in the form of instrument measurements). These are all identified with a UUID, either read from the DM with a 2D barcode reader or generated in the form of instrument measurement using a UUID generator (onboard web page).

Once the scientist finishes recording all of the metadata (and data), the sheet can be checked for mistakes using the spreadsheet checker (hosted as a webpage by a VM), which only checks columns generated by the template generator (order does not matter). This means that if the users decided they wanted to add some of their own notes, these are not parsed. Valid columns are identified by standard names contained in a hidden row in the spreadsheet. At the end of the cruise, the error free spreadsheets are submitted to the data manager, which uses custom tools to read the spreadsheets into the sample database, which is then verified and published online.

SAMPLE RELATIONS

When sampling, it is rare that only one sample is collected from a piece of gear or area. Additionally, samples are often sub-sampled and maybe these again are sub-sampled, and so on. These relations are important to track for several reasons. Firstly it is common to investigate several properties of the same volume, like for example the pH, TOC, oxygen content and salinity of a volume of water. To collect data from such sub-sampling, the relation between different samples needs to be recorded. Secondly by tracing these samples, it is possible to maintain knowledge of which samples went where and what was sampled. Lastly the knowledge of these relations can be exploited to reduce the workload of recording sample metadata.

To keep track of the tree of relations, a parent-child system was employed. In this system only the child needs to know who its parent (singular) is. This is achieved by letting the child (sub-sample) record the UUID of the parent as a *parentEventID* parameter. An example would be that the liver (child) from a fish would set its *parentEventID* to the UUID (*eventID*) of the fish (parent), while setting its own *eventID* to a new UUID, see ***Figure 4***. In the case of the Nansen Legacy cruises, the top of the tree of relations was chosen as the deployment of a piece of sampling gear. In other applications, it could for example be a measurement station, a certain field site or a sample. ***Figure 4*** shows how this builds a tree of relations. The logging of a deployment is done by one responsible participant, or a computer (cruise logger), which assigns the gear deployment UUID (*eventID*) and records its metadata. On the Nansen Legacy cruises this metadata contains position, time, date, gear type, depth, station name and similar information. The participants then do not need to log these parameters for samples they collect from the deployed piece of gear, as long as they relate the sample to its parent (either a sample or the gear). To easier identify their sample in the spreadsheet, most find it convenient to log some of the parent metadata. For cases where multiple scientists need access to the same parent or in the case where it is not a labelable sample, for instance reusable Niskin bottles, A4 sheets with 2D barcodes with the *parentEventID*s were distributed. An example here would be a rosette with 12 Niskin bottles. The rosette deployment is logged as a given gear deployment and given an *eventID* (UUID). The 12 Niskin bottles are then assigned 12 unique UUIDs, see first part of ***Figure 4***, which are printed on an A4 page as 2D barcodes. They are then logged in the spreadsheet using the deployment *eventID* as their *parentEventID*. When scientists want to log samples taken from the Niskin bottle, they can then then scan the relevant barcode (containing the UUID) from the A4 sheet as their samples' *parentEventID*.

When the spreadsheets are collected at the end of the cruise, all the entries are put into a PostgreSQL database. When recoding in the field, a parent and its children could be located in different spreadsheets. This is solved in the database, where the parents for every child is found, all the way to the top. Checks are then run to ensure that there are no orphans (children without parents). Once this is done, fields that have been flagged as inheritable are written onto all the children, grandchildren,.... If a child has one of the parameters already recorded, a flag for the parameter is checked. If the flag shows that the parameter follows strong inheritance, the data is overwritten, if it is weak, the data is kept. What is strong and weak, depends on if the parameter could be different between parent and child. Most parameters are strong (e.g. location, station), but for instance the depth range is weak, as for instance different nets in one multinett could be deployed at different depth ranges.

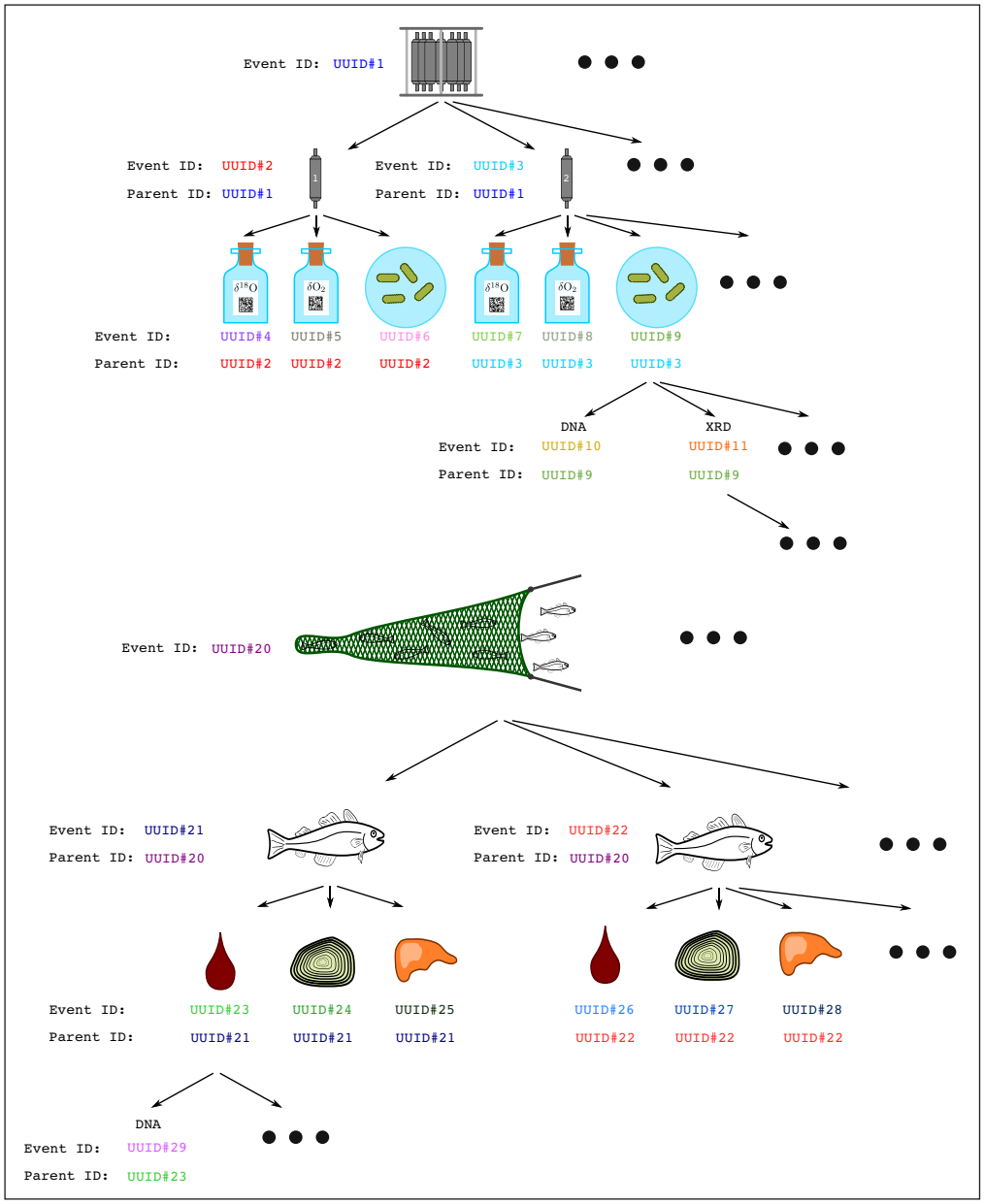

Figure 4 The figure shows two examples of parent-child relation trees. Both trees display the inheritance of UUIDs from parent to child.

DATA AVAILABILITY

The metadata that is presented here is available freely from the Svalbard integrated arctic observing system (SIOS) web portal at *https://sios-svalbard.org/reports/aen_multi*. On the same page links to the sampling protocols can be found.

CODE AVAILABILITY

The code developed is freely available from GitHub. There are a number of different repositories published there, see ***Table 1*** for the details. The code is split in several repositories to ensure that the different components can be deployed independently. All of the code is GPL-3 licensed.

DESCRIPTION	WHERE
Spreadsheet generator	*https://github.com/SIOS-Svalbard/darwinsheet*
Drupal modules	*https://github.com/SIOS-Svalbard/Aen_sample*
Spreadsheet to database conversion	*https://github.com/SIOS-Svalbard/AeN_data*
Label printing	*https://github.com/SIOS-Svalbard/AeN_print*
Additional documentation	*https://github.com/SIOS-Svalbard/AeN_doc*

Table 1 The table lists the available software and the link to the GitHub repository with it.

DISCUSSION

The tools and approach described were developed as part of the Nansen Legacy project with the goal to ensure a standardised event logging and sample tracing within the project. A particular focus was on the preparation for data for publication and sharing of information within the project during this phase. Additionally focus was placed on simplifying the data publication process for the scientist through collection of information in a standardised form already during data collection. Standardisation for metadata collection was achieved through the use of the spreadsheet template generator. It was also important to standardise the collection of data. This was covered by requiring the research groups within the project to agree on measurement protocols, which were collected into a common document and published (The Nansen Legacy Version 1 2020). There will always be updates, new procedures and improvements to these protocols, which, when deemed necessary, were incorporated by releasing new versions of the document (The Nansen Legacy 2020, Version 1, Version 2, Version 3, Version 4.2, Version 5, Version 6). Changes to protocols need to be managed to ensure that the changes do not affect the comparability of the resulting data unnecessarily. Application of the protocols and templates ensure that sampling at different times and by different scientist are conducted in a repeatable way, producing data which can be easily reused both inside and outside of the project.

Another aim was to ensure that these tools could fit into other projects and workflows conducted by the participating scientists. To allow for this, all the tools were developed using open standards and are freely available on the SIOS (Svalbard integrated arctic earth observing system) webpage. This webpage is not accessible on a cruise, but access to a similar setup was delivered by a virtual machine deployed on the ship.

On a ship, with access to servers and working areas, the approach worked well after an initial startup period, which included guidance and demonstrations given to the participants. Bringing one data manager on the initial cruise further helped in the implementation. Through experienced based transfer of knowledge, the concept was further spread into new user groups. A solid commitment from the group leaders ensured a successful adoption.

The separation of label printing and logging, results in some extra work by having to enter the UUID into the logging sheet (by scanning the DM), but it significantly increases the flexibility, as a label is not in use before it is logged. Thus, the small labels with only the DM code can be printed in bulk and only used when needed. If some marked sample holders were not used, the labels were still valid for a future use. Printing of labels and labelling should be done in advance of sampling, improving planning of the sampling strategy, and reducing the workload during sampling. Some of the metadata could be registered ahead of time in the spreadsheet, linking the UUIDs and the metadata, further reducing the workload during sampling.

Within the first year of the project, there were several occasions where scanning a DM on a sample was the only way of identifying it, as the other markings were either rubbed off or wrong. This alone showed the possible improvements gained with the presented strategy. In the process of reading the spreadsheets into the database, inconsistencies and mistakes were found, including orphaned samples (missing parent) and could, in most cases, be fixed with the help of the person recording these, as these were discovered shortly after the cruise. The standardised measurement protocols (available from the SIOS page) and referencing each sample to the relevant protocol, provided another check for correctness of the metadata.

In the process of implementing the tools, it was found to be of utter importance to have enough training and guidance available to busy participants. The commitment from the project leaders help pull the methods past their initial issues and bugs, into a system capable of increasing the production of FAIR data. The tools developed here should be easily extendable to other fields, especially the template generator has been shown to be powerful in reducing errors and making different groups agree upon common parameter names.

Most fieldwork is not conducted on large cruises with a ship full of facilities for hosting virtual machines and multiple label printers. To address the needs of these scenarios, the methodology has been tested in different smaller settings. One example is going out in a smaller boat to take a local station. In such a case the sample containers were labeled before going into the field, removing the need to bring a printer into the field. Labeling beforehand requires a bit more planning on which samples to take, but it saves time in the field. For logging the metadata,

the spreadsheet is generated before going into the field, and then filled in during the fieldwork. Information can also be pre-filled before going out, as the sample containers with the IDs are already printed on the labels.

Based on the tools and approaches developed, metadata can be published shortly after data collection with the relations between samples established. Printing of labels and use of the template generator has encouragingly also been adopted outside of the project, showing the utility of these tools for the wider scientific community. Thus, these methods can provide a foundation for improved production of FAIR data.

LIMITATIONS

The presented method was designed specifically for the Nansen Legacy project, but with a generic perspective to make it useful for scientists in the domains represented within Nansen Legacy. This implies that there are some limitations of the approach.

One is that the identifiers, the UUIDs, do not resolve to any of the metadata, nor links to a webpage or similar (like DOIs, URIs, ...). An example of this limitation would be someone scanning a sample DM, getting an UUID and not knowing to search the SIOS webpage for the metadata. This is only a limitation for select samples (small labels) analysed by someone without knowledge of the project. It allowed for generation of UUIDs at any time, and without the need to contact any online resources. Within the project the need to trace UUID was solved by an online public searchable database. This visibility of the database to search engines is limited and could be improved by adding for instance *schema.org* representation.

The solution for pre-publishing event metadata between the cruise and the full data publication does not comply with all the FAIR principles. It is not interoperable as custom field names are used without an API resolving these, nor does it fulfill requirements to be reusable as it lacks a clear specification of a license. It is however somewhat findable both through the search interface and general search engines, accessible as the data can be downloaded as *.tsv* (tab separated) files with the standard names as well as investigated on the webpage. Lastly it is also somewhat reusable, as the field definitions and export formats are mostly standard. As the primary purpose of the approach presented here is to improve data sharing prior to data publication and simplify the data publication of FAIR data, this was considered a substantial improvement over existing approaches like field diaries.

Another issue is the potential inconsistency between what is printed on the physical sample labels and what is recorded in the metadata sheet. The choice to keep label printing and metadata logging separate, was done intentionally. Enforcing a link by directly printing for instance some preselected fields in the spreadsheet might save some time. Such a link would need careful software and web server designs and would put limitations on what information could be on the labels. As the labels have limited room for information it was found important to allow scientists to decide the information themself, as these labels need to be understandable after the cruise for scientists and laboratory personnel. Here most of the groups already had established labeling requirements, that could be incorporated in the separated procedure. An opening for custom labeling text comes with a reduction in standardisation, though the assumption was that mandatory DM with the UUID was enough. In practise most scientists used the date and name template format inserted by the printing page, supplied as a gradual standardisation effort.

The resulting database of samples will over time contain samples that have been lost (in processing) and incorrect sample location. These deficiencies could be corrected by building a system for updating the samples continuously, but it was deemed too time consuming and workflow invasive and it was decided that the participating scientists and institutions were responsible for tracking the samples in their systems once transferred from the ship/field to their locations. A work in progress in the project is to update the event metadata with a DOI referring to the published dataset containing the data from the sample.

CONCLUSION

Sample tracking in multidisciplinary fieldwork is a complex topic, which is also developing together with an increasing need of documenting (meta)data in a standard and FAIR way,

sharing data and resources among scientists and, not least, using data management tools and workflows to facilitate the work of researchers to produce sharable science. In this work, we have presented a method to approach sample tracking, which has been implemented and used in a large project, involving many scientists from different institutions and with different backgrounds, but that can be considered general enough to be reused in different contexts. The methodology presented, can be further improved and adapted to better respond to the needs of a specific campaign, but the underlying approach would still be valid, as it is based on well known standards, as Darwin Core, the Attribute Convention for Data Discovery, as well as the Climate and Forecast convention, which have been extended and adapted to fit the needs of such large research campaign.

By adopting and extensively using this approach over several research expeditions, strengths and weaknesses have both appeared. For example, the method we proposed requires quite some effort and time to be in place, probably more than previously adopted approaches, but the overall advantages outweigh greatly this issue. Some of the advantages includes an improved overview over the sampling effort, a common understanding of important parameters and sampling procedures, and a simplified data publication process through capture of high quality and standardised metadata early in the data collection process. The latter in compliance with the ESIP data management training material prepared by Ruth Duerr, "The 5P's matter" – Prior Planning Prevents Poor Performance (Russell 2013).

One of the greatest strengths of the approach is its ability to to work in a range of fieldwork and scientific cruise situations, from individual scientists to large projects. The approach is not unique, similar approaches have been implemented by other communities, but valuable lessons learned are gained through the implementation in one of the largest data ocean cruise data collection efforts in Norway to date.

ACKNOWLEDGEMENTS

This research was funded by the Research Council of Norway through the Nansen Legacy project (NFR-276730). We acknowledge Christian Svindseth, UiO Natural History Museum for the first version of the template generator (*https://github.com/umeldt/darwinsheet*). The participants of the Nansen Legacy are thanked for their willingness to participate and give feedback in the development, testing and implementation of tools and approaches presented here.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Pål Gunnar Ellingsen has written most of the article, with important input from the co-authors. The software was developed by Pål Gunnar Ellingsen with contributions from Lara Ferrighi. Øystein Godøy and Tove Margrethe Gabrielsen have contributed to the development of the methodology used for printing and metadata collection.

AUTHOR AFFILIATIONS

Pål Gunnar Ellingsen *orcid.org/0000-0002-3331-5581*
Department of Electrical Engineering, UiT The Arctic University of Norway, 8505 Narvik, Norway;
Department of Arctic Biology, UNIS – University Centre in Svalbard, 9170 Longyearbyen, Norway

Lara Ferrighi *orcid.org/0000-0001-5221-1377*
Norwegian Meteorological Institute, P.O. Box 43 Blindern, 0313 Oslo, Norway

Øystein Godøy *orcid.org/0000-0001-6410-3488*
Norwegian Meteorological Institute, P.O. Box 43 Blindern, 0313 Oslo, Norway; Svalbard Integrated Arctic Earth Observing System – Knowledge Centre, Norway

Tove Margrethe Gabrielsen *orcid.org/0000-0001-5801-4569*
Department of Arctic Biology, UNIS – University Centre in Svalbard, 9170 Longyearbyen, Norway;
Department of Natural Sciences, CCR – Centre for Coastal Research, University of Agder, 4306 Kristiansand, Norway

REFERENCES

ACDD 1.3 Change Summary and Record; Telecon Minutes. ESIP Commons, December 2014. 2014. URL: http://commons.esipfed.org/acdd%7B%5C_%7D1-3%7B%5C_%7Dreferences (visited on 2020-03-02).

Butler, H, et al. 2016. *The GeoJSON Format GeoJSON*. URL: https://tools.ietf.org/html/rfc7946. DOI: https://doi.org/10.17487/RFC7946

CF Standard Names. 2020. URL: http://cfconventions.org/standard-names.html (visited on 2020-02-28).

Damerow, JE, et al. 2021. Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. In: *Data Science Journal*, 20(1). ISSN: 16831470. DOI: https://doi.org/10.5334/dsj-2021-011

Eaton, B, et al. 2020. *NetCDF Climate and Forecast (CF) Metadata Conventions, v 1.8*. URL: http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.html (visited on 2020-06-03).

European Commission. 2019. *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information*.

GBIF Norway. 2020. *DwC Excel Template Generator (2020) source code*. URL: https://gbif-norway.github.io/dwc-excel-template-generatorjs/%20https://github.com/umeldt/darwinsheet (visited on 2020-06-24).

GBIF.org. 2020. *GBIF Home Page*. URL: https://www.gbif.org/what-is-gbif (visited on 2020-06-23).

GO FAIR. *GO FAIR*. URL: https://www.go-fair.org/ (visited on 2020-06-03).

Guralnick, R, et al. 2014. The trouble with triplets in biodiversity informatics: A data-driven case against current identifier practices. In: *PLoS ONE*, 9(12): 1–13. ISSN: 19326203. DOI: https://doi.org/10.1371/journal.pone.0114069

Intergovernmental Oceanographic Commission of UNESCO. 2020. *OBIS (2020) Ocean Biodiversity Information System*. URL: www.iobis.org (visited on 2020-06-23).

Juliano Palacios-Abrantes, et al. 2019. A metadata approach to evaluate the state of ocean knowledge: Strengths, limitations, and application to Mexico. In: *PLoS ONE*, 14(6): 1–18. ISSN: 19326203. DOI: https://doi.org/10.1371/journal.pone.0216723

Lawrence, BN, et al. 2006. Maintaining and Advancing the CF Standard for Earth System Science Community Data. In: *Development*, 1–12. URL: http://cf-pcmdi.llnl.gov/documents/white-papers/cf2%7B%5C_%7Dwhitepaper%7B%5C_%7Dfinal.pdf.

McNutt, M, et al. 2016. Liberating field science samples and data. In: *Science*, 351(6277): 1024–1026. ISSN: 0036-8075. DOI: https://doi.org/10.1126/science.aad7048

Pecci, L, Fichaut, M and **Schaap, D.** 2020. SeaDataNet, an enhanced ocean data infrastructure giving services to scientists and society. In: *IOP Conference Series: Earth and Environmental Science*, 509(1): 10–12. ISSN: 17551315. DOI: https://doi.org/10.1088/1755-1315/509/1/012042

Plain-Jones, C. 1995. Data Matrix identification. In: *Sensor Review*, 15(1): 12–15. ISSN: 02602288. DOI: https://doi.org/10.1108/EUM0000000004265

Russell, T. 2013. Prior planning prevents poor performance. In: *Australian Pharmacist*, 32(2): 28.

Schaap, DMA and **Lowry, RK.** 2010. SeaDataNet – Pan-European infrastructure for marine and ocean data management: Unified access to distributed data sets. In: *International Journal of Digital Earth*, 3(SUPPL 1): 50–69. ISSN: 17538947. DOI: https://doi.org/10.1080/17538941003660974

Stall, S, et al. June 2019. Make scientific data FAIR. In: *Nature*, 570(7759): 27–29. ISSN: 0028-0836. URL: http://www.nature.com/articles/d41586-019-01720-7. DOI: https://doi.org/10.1038/d41586-019-01720-7

The Nansen Legacy. 2020. Sampling Protocols: Version 1. In: *The Nansen Legacy Report Series*, 9. DOI: https://doi.org/10.7557/nlrs.5715

The Nansen Legacy. 2020. Sampling Protocols: Version 2. In: *The Nansen Legacy Report Series*, 10. DOI: https://doi.org/10.7557/nlrs.5716

The Nansen Legacy. 2020. Sampling Protocols: Version 3. In: *The Nansen Legacy Report Series*, 11. DOI: https://doi.org/10.7557/nlrs.5717

The Nansen Legacy. 2020. Sampling Protocols: Version 4.2. In: *The Nansen Legacy Report Series*, 12. DOI: https://doi.org/10.7557/nlrs.5718

The Nansen Legacy. 2020. Sampling Protocols: Version 5. In: *The Nansen Legacy Report Series*, 13. DOI: https://doi.org/10.7557/nlrs.5719

The Nansen Legacy. 2020. Sampling Protocols: Version 6. In: *The Nansen Legacy Report Series*, 14. DOI: https://doi.org/10.7557/nlrs.5720

UiO Naturhistorisk museum. 2020. *GBIF Norway (2020)*. URL: https://www.gbif.no/ (visited on 2020-06-24).

Unidata. 2020. *NetCDF*. Boulder, CO. URL: https://www.unidata.ucar.edu/software/netcdf/. DOI: https://doi.org/10.5065/D6H70CW6

Wagner, CS, Whetsell, TA and **Leydesdorff, L.** 2017. Growth of international collaboration in science: revisiting six specialties. In: *Scientometrics*, 110(3): 1633–1652. ISSN: 15882861. DOI: https://doi.org/10.1007/s11192-016-2230-9

Wieczorek, J, et al. 2012. Darwin core: An evolving community-developed biodiversity data standard. In: *PLoS ONE*, 7(1). ISSN: 19326203. DOI: https://doi.org/10.1371/journal.pone.0029715

Wilkinson, MD, et al. 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data*, 3: 1–9. ISSN: 20524463. DOI: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Ellingsen, PG, Ferrighi, L, Godøy, Ø and Gabrielsen, TM. 2021. Keeping Track of Samples in Multidisciplinary Fieldwork. *Data Science Journal*, 20: 34, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-034*

Submitted: 18 December 2020
Accepted: 20 October 2021
Published: 10 November 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.