RC1

This reviewer thoroughly enjoyed reading this manuscript, and found the presented data management approach and strategies to be commendable and significantly valuable; outlining the technical and cultural aspects considered for the creation, management and publication of FAIRer data resulting from large, coordinated, oceanographic research efforts. Telling this story, which includes successes and lessons learned, can serve as a model for future efforts, further facilitating culture change toward a more FAIR and open oceanographic research community.

The manuscript is also well organized and easy to read. This reviewer has general and minor specific comments that may help improve the readability of the text as it relates to the underlying FAIR/Best Practices theme being conveyed. These are outlined below:

General:

> *The manuscript frames the strategies and implementations in the context of addressing the FAIR principles. The authors mention how FAIR is often referenced in abstract or incomplete ways in section 5.1 (lines 214-216), however the text itself does not relate each strategy or implementation to the specific principles themselves in a holistic or comprehensive manner.*

While this reviewer agrees with all statements that predicate the principles on machine readability or improving reuse as an ultimate goal, only superficial references were made as to how the project's implementations satisfied or related to each principle or sub-principle. It would be helpful in justifying or substantiating these statements by the authors to either organize the text along the principles, or link the statements as they occur in the text to the explicit principle or sub-priniciples they address and how. This is just a suggestion as it is realized this could require substantial revision.

This author considered organising the manuscript along the principles (FAIR) when drafting it, but found this to be challenging. This is in part because individual aspects of data management influence the adherence of data to more than one of the principles. For example, choice of data format determines both how interoperable and reusable data are, but also can affect accessibility (e.g. netCDF files can be served over OPeNDAP whereas other formats can not).

It is, however, a good idea to refer to which principles are considered at different stages in this section. We have included this sentence towards the end of the introduction:

"To clarify how the FAIR principles were considered throughout the project, each relevant aspect is annotated inline using the corresponding initial — F (Findable), A (Accessible), I (Interoperable), and R (Reusable) — shown in parentheses."

We have adjusted the text in accordance to this statement. We have decided not to refer to specific subgoals within the principles as they have not been introduced properly within this manuscript - and one reviewer has called for the manuscript to be more concise.

> *This reviewer cautions against the use of 'fair-compliance'. The FAIR principles are just that, principles. As such, there are a variety of strategies to employ that improve the alignment or achievement of data toward their end goal. But implying data can meet some metric to be FAIR or unFAIR is incorrect. Rather, data can be _more_ FAIR along a continuum, or relative to other data. The term 'fair-compliance' implies a pass/fail measurement against some metric.*

This is an important point that has been made by both reviewers. We have removed any phrasing that suggests that the FAIR principles are some pass/fail metric that data can comply with.

This relates especially to section 5.3.1, where lines 325-331, implying some data centers do not comply with mentioned metadata standards such as ISO 19115, or GCMD DIF, thus reducing their catalog's findability. While it is agreed that the selection process the Nansen Legacy project employed was robust and laudable, the standards referenced are not the only ones available, and if a center employs, for example, schema.org or EML, they may still achieve findability. The challenge with FAIR is that multiple implementations can be community or center -specific, thus, commonly used formats may still need crosswalks between them to improve 'center interoperability'.

We agree with this and have listed these standards only as examples. We have slightly adjusted the text to make this clearer, also including schema.org and EML.

> *In reference to the conclusions drawn by the Nansen Legacy project, that a two pronged blend of technology and supportive cultural environment is needed for success, in combination with the description of a committed leadership team and a dedicated data management committee, it is curious that no findings or recommendations of the cultural side of changing a research group to a more data management savvy one was provided.*

We have expanded the Discussion and Summary section to address this important point.

> *Likewise, it would be helpful to understand (with even a brief description) the process of the governance development process for the data management effort. It is truly impressive and worth understanding its early evolution to serve as an example for future projects and programs.*

We have expanded the introduction to include a description of this.

“Effective data management was prioritised from the very start of the project, beginning with the preparation of the proposal. A dedicated team consisting data managers and scientists from all partner institutions, outlined the principles that would govern the project and incorporated these into the first draft of the project’s data policy (The Nansen Legacy, 2021) and data management plan (The Nansen Legacy, 2024) at the outset. The leadership team’s involvement was crucial in ensuring these foundational documents were both well-conceived and effectively implemented. These documents served as the cornerstone for all subsequent data management activities discussed in this paper.

The project allocated resources and competence through a dedicated work package on data management led with complementary expertise. This included experience from international data management structures (from e.g. World Meteorological Organization), genetic database systems and physical and biological field work. A dedicated full time data manager was appointed to plan, develop and support the data handling. In addition, a data management resource group with data managers from each partner institution was established to strengthen collaboration, facilitate harmonised handling across disciplines and institutions, and support a legacy beyond the project period. This may facilitate further development and cultural change. Project management provided funding for training and data publishing workshops to ensure broad involvement.”

Specific:

> *Section 4.3 beginning on line 179 describes lessons learned from the project’s data storage and sharing strategies, citing the challenges of governing data sharing across a large collaborative multi-institutional project, but offers no recommendations for future improvements. The international GEOTRACES program also has a data management committee that coordinates data sharing and the branding of GEOTRACES data is based upon project requirements being met. Researchers who were invested in their work being labeled as GEOTRACES strived to meet the requirements, which include the sharing of datasets for inclusion into the branded GEOTRACES Intermediate Data Product (IDP). Perhaps the visibility and gravitas of one’s research output holding the project brand could assist in the data sharing culture shift. I wonder if the Nansen Legacy project considered such branding strategies.*

This looks like an interesting initiative. However, in section 4.3 we are referring to data that are not ready to be made publicly available, and how to facilitate early sharing of unpublished data between project members. It seems that data shared via GEOTRACES would be available beyond project members. The approach chosen by Nansen Legacy was building on the principle adopted by MOSAIC for full transparency internally in the project prior to publication of the datasets. This addressed the need for internal collaborations and quality control, and embargo for educational purposes. The National e-Infrastructure for Research Data (NIRD) is a platform suitable for internal research collaboration in Norway.

The “branding strategy” we aimed for was to instill a culture of trust and respect, where sharing data is more common between normally competing scientists. Hopefully project alumni and those interested in the project will continue these practices. By involving institutional data managers within the project, we strengthened communication between data managers and scientists and between institutions, which we hope will lead to a collaborative cultural change on both a scientific and institutional level.

The reviewer raises a good point that no recommendations or future improvements are offered in the manuscript. We have added:

“Although data were not always shared through NIRD, overall sharing increased during the project. Several practices that supported this are worth carrying forward. From the outset, we aimed to build trust and respect among scientists and institutions who might normally be in competition. Involving institutional data managers within the project proved particularly valuable, strengthening communication and collaboration both between data managers and scientists, and across institutions.”

> *Section 5.2, line 221, in the statement “Nansen Legacy data should be published in FAIR-compliant data formats such as NetCDF files compliant with the Climate and Forecast conventions (Eaton et al., 2024) or Darwin Core Archives (Darwin Core Community, 2010) whenever possible (The Nansen Legacy, 2024).” it should be noted that NetCDF format is not necessarily FAIR in and of itself, but can be when semantically enabled terms such as CF are used in combination. This sentence is a bit leading, and would suggest removing the second instance of ‘compliant’ after NetCDF files: “... such as NetCDF files when accompanied by the Climate and…”*

Agree that the wording here could be better. We have changed this to “Nansen Legacy data should be published in FAIR data formats whenever possible (The Nansen Legacy, 2024). Recommended formats are NetCDF files that adhere to the Climate and Forecast (CF) conventions (Eaton et al., 2024) and Darwin Core Archives (Darwin Core Community, 2010).”

> *Line 224 “... which aims to make all data relevant to Svalbard available…” should be edited to “which aims to make all data relevant to Svalbard discoverable…” since the sentence is referring to data ‘findability’, unless the authors are also claiming accessibility via SIOS as well, in which case it can be edited to do so.*

Changed to “discoverable”

> *In section 5.3.3 Granularity (lines 391-392), the statement regarding avoidance of mixing feature types and vertical dimensions could use some clarity. For example, can surface observations at depth=0 not be combined with vertical profile observations? This seems perfectly acceptable. Likewise, there exist*

many time series datasets that consist of vertical profiles aggregated over time that are still easily usable in tabular or other formats. Although the thought is agreed with, the examples do not seem to convey or substantiate the thought.

We agree with this and have changed this bullet point to:

"Granularity in mixed-dimension datasets: Publish each feature type (vertical profile, trajectory, etc.) separately by default. Only combine when they share exactly the same coordinate axes and measurement context (e.g. a time-series of vertical profiles at a fixed location). Use explicit feature-type and dimension metadata — CF conventions' featureType is one example, but equivalent tags in other formats work just as well (Eaton et al., 2024)."

RC2:

This manuscript is well written and provides a clear and generally well described overview of the methodologies and lessons learned from the data acquisition and management processes implemented as part of the Nansen Legacy Project. It also includes some well-considered and important recommendations for future research programmes.

The information provided is potentially very useful to those seeking to initiate a similar activity, which is demonstrated by the fact that the data management framework adopted by the Nansen Legacy Project has already been shown to be transferable to other similar activities within other organisations.

The approach to data management adopted by the Nansen Legacy project highlights the potential impact of fostering and implementing good data stewardship practices, which not only encourages cultural change among researchers but also ensure that data is open and accessible for wider reuse.

The manuscript could benefit from being made more concise in some places. There is also some minor duplication of information between sections that might be removed to improve the readability of this paper.

We have made an effort to make the paper more concise by removing duplication of information. This has been done throughout but most prominently in the Data Publishing section.

Specific comments and suggested edits for improving the manuscript are given below:

Suggested language (phrasing / grammar) edits:

Line 7: Suggested edit for clarity: " The project also implemented a policy that mandates immediate data sharing…."

This has been changed as suggested.

Line 18: Suggested edit to improve phrasing / readability "……...to be compared,

synthesised and reused…..”

This has been changed as suggested.

Line 44: Is there a reason that the first letter of “Publishing” is capitalized. This is not consistent with the form used elsewhere in the manuscript

Thanks for identifying this typo. This has been changed.

Line 62: The sentence “Detailed protocols were developed in collaboration and published, coordinated by a senior engineer who worked closely with each researcher and group” is not well constructed and therefore difficult to read. Consider rephrasing to improve the readability and clarity.

Agree, this has been changed to “Detailed protocols were developed collaboratively and published, with coordination led by a senior engineer who worked closely with each researcher and research group.”

Line 119: Construction of the sentence is rather awkward. Suggested edit for clarity: “Scientists included contact details for the principal investigator, estimated timeline for publication and details of any relevant embargo period for each dataset …….”

Thanks for this improvement. We have made the change, but added “requested” before “embargo period” since the embargo period had to be approved.

Line 121: Suggested edit to correct grammar: “These tables were included in the project’s data management plan……”

Agree, this has been changed.

Line 167: Suggested edit to improve sentence construction and clarity: “However, permission from the relevant principal investigator was required before any use of the data.”

Agree and changed as suggested.

Line 173: Suggested edit to improve sentence construction: “Scientists were also encouraged to share their own Nansen Legacy data via the project area.”

Agree and changed as suggested.

Line 180: Suggested edit to correct grammar: "….data were instead often shared between project members via other methods."

Agree and changed as suggested.

Line 183: Suggested edit to correct grammar: "…..unfamiliarity with using the NIRD platform for many project members……"

Agree and changed as suggested.

Line 186: Suggested edit to improve sentence construction: "However, governing this sensitive topic at scale within the project was deemed to be challenging and impractical and was instead managed on a case-by-case basis when data access was requested by a project member."

Agree and changed as suggested.

Line 325: Suggested edit to improve phrasing / readability:" It is not practical for each data access portal to provide custom workflows to harvest metadata from each individual data centres."

Agree and changed as suggested, except for removing the 's' at the end of the sentence.

Line 351: It Is not clear what is meant by the word "opening" in this statement. Do you mean supporting open access and visualisation?

This was meant to refer to the physical opening of a file by some tool or software, but we agree this was not clear. Changed to 'access'.

Technical remarks

Line 99: Although the Darwin Core and CF Standard Names are implemented "where possible" it is unclear if a standard vocabulary has been implemented alongside these standardised terms to ensure consistency of terminology across all metadata records. It would also be useful if the metadata standard that has been implemented is clearly identified.
We have added "Other terms, not available in controlled vocabularies, were defined by the project to meet its specific needs."

Line 204 – 219: The purpose of including these selected project references is unclear. It does not enhance the paper and is superfluous in this context.

The original purpose of referring to these projects was to provide a motivation for people who have data to publish by informing them of examples of how their data could be used. However, we acknowledge that the manuscript is quite long so have removed this part to make it more concise.

A number of references are made to FAIR compliance. It should be noted that the FAIR Principles are intended to be a set of guidelines that provide the framework for making data findable, accessible, interoperable and reuseable . Line 239 states that data should "meet the FAIR Principles", suggesting that data is either compliant or it is not, which is not the case – it is not a binary assessment. Data that is FAIR is aligned with these guiding principles to some degree, and this alignment can be both assessed and improved. There are several FAIR assessment tools available that allow the "FAIRness" of a dataset to be determined and also indicate how this might be improved.

This is an important point that has been made by both reviewers. We have removed any phrasing that suggests that the FAIR principles are some pass/fail metric that data can comply with.

Line 234 suggests that there is a lack of tools to support delivery of FAIR data. However, there are several initiatives currently providing tools that support the adoption of the FAIR principles and delivery of data that is findable, accessible, interoperable and reusable. For example, the GO FAIR initiative has developed the FAIR Implementation Profile (FIP), a methodology that allows research communities to express their specific choices and practices for making data and metadata FAIR, which might be useful in this context.

It is important to include that there are existing tools, softwares and initiatives to help with this. The FAIR Implementation Profile is a nice initiative to refer to in this context. However, we still believe far more needs to be done to streamline the process. We have changed the text to this:

"Publishing FAIR data is new for many scientists and there is a learning curve associated with this. There is therefore a clear need for better support and guidance. While some tools and frameworks have been developed—such as the FAIR Implementation Profile by the GO FAIR initiative (Magagna et al., 2020), which enables communities to articulate their approaches to FAIR data—there remains a significant need for additional tools and software to support and streamline all aspects of the FAIR data publishing workflow."