OCEANUM
OCEAN NUMERICAL

Insights, lessons, and recommendations from the DIA Marine Data Innovation Project

30 June 2022

LINZ management of the project

Excellent oversight and management was provided by LINZ throughout, with a supportive structure that created a responsive and positive environment for the cross agency engagements. Both the project managers (Anna and Ashley) were key drivers of success and should take credit for keeping folk on task and ensuring the overall vision was aligned within each party's use case. The project had initial challenges beyond the usual shake down period for new teams in collaboration. We started in December and so the summer season did not allow real focus to develop until February. More time than anticipated was spent in developing the use cases, and the initial two sprints were disparate and unfocussed as a result. However, once all had a clear understanding of the project and the plan to work towards, the subsequent sprints were a productive pathway to execute upon. Use of the monthly demos to the stakeholder group was a useful and highly visible focal point to show progress. It would have been nice to have a few physical meetings during the project, but COVID didn’t allow that. As the supplier to the POC, we were well supported throughout and felt the project management by LINZ was setting all the parties up for the best chance of success. It might seem obvious, but ensuring good management and governance structures are in place is a recommendation for DIA in future innovation projects they may fund.

DoC use case

Early identification of datasets and third parties would have allowed more progress in the beginning, but the excellent resourcing and solid commitment shown during the second half of the project led to a successful POC outcome, in our opinion. We gained an understanding of the complex and challenging task that DoC has at hand, and the clear benefit to be gained from having an efficient data discovery process readily available to key staff and contractors.

The collaboration between NIWA, DoC and Oceanum in developing the analysis and end-user application was effective and straightforward. The DataMesh provided a common point of reference for the data used in the analysis, and allowed the rapid development of a custom App to expose the analysis to stakeholders.

MPI use case

The MPI use case that was tested during the POC was relatively simple, and made use of a set of individual but related datasets, the national Cetacean distributions. This use case was a good opportunity to demonstrate the ability of the DataMesh to connect to multiple individually published raster layers and expose these as a single virtual dataset that could be analysed as a single entity. Connecting to the MPI geospatial services was straightforward and collaboration around getting datasets published to support the use case worked well.

Our understanding of this use case was for an analytical tool that was:

1. Connected to the all of the underlying data sources
2. Executable within the overall data sharing platform
3. Distributed and shareable through the data sharing platform

The initial intention was to integrate the DataMesh with a tool that had already been developed by MPI in ArcGIS Pro. However, both the proprietary nature of the ESRI product and licensing restrictions made this approach challenging.

As the core analysis itself was satisfied by a spatial aggregation operator in the DataMesh itself, the easiest approach to expose the necessary analysis was to create a custom DataMesh export. This allows the complete analysis workflow to be carried out within the DataMesh itself. The downside to this approach is the current inability of MPI to change the internal functionality of the export function. While it satisfies the second requirement above, it doesn't really offer a way for an external party to download, potentially modify and share the analysis functionality.

As an intermediate work-around Oceanum decided to simply deliver a custom app that exposes and describes the analysis for the use case. There are more elegant ways of achieving this outcome in a more generic and extensible way, however this was not achieved in the POC timeframe.

The primary insights gained from this use case were:

1. It is challenging or impossible to integrate directly with proprietary applications, as they lack stable and/or published extension points to hook the DataMesh to.
2. While custom analysis functionality can be added to DataMesh as export functionality, the current architecture means that it cannot be modified by users. This creates some degree of vendor lock-in and may cause maintenance issues.
3. Deep integration with most existing platforms would involve developing toolboxes or extensions within the applications themselves (for example ArcGIS Pro extensions). However these approaches would have their own development and maintenance overheads separate and parallel to a core data sharing infrastructure.
4. There is a need for a general platform to share analysis workflows and apps with other organisations or the public. This goes beyond the scope of a data sharing platform but would provide very useful capability.

Te Arawhiti use case

We were surprised by the impact that standard business IT practices had by inadvertently introducing barriers to the type of enabling technology that the DataMesh seeks to deliver. This is an important lesson from the perspective of a supplier, and one that can be expected to be repeated elsewhere. The mitigation steps are relatively simple, but rely on having a willing and engaged IT department to implement. The lesson here is to have the support of the IT department at the outset of the project. It is also worth noting that these issues were experienced by a cloud-based platform, requiring only internet connectivity. A system requiring on-premises installation might face far worse barriers to implementation.

The datasets provided by Te Arawhiti were available from the start of the project and overall were easy to integrate. However, the size of two of the layers presented some initial challenges as they exceeded the maximum number of features accessible in a single request to the ESRI feature service. This limitation was overcome by some internal modification of the DataMesh drivers, however illustrated the potential need to work around external limitations in contributing data services.

LINZ use case

This use case highlighted the current gap in available COP systems to rapidly ingest and display spatial and point data that has a temporal component. As a supplier, the exercise revealed functionality we need to include in a complementary service currently on our development roadmap, particularly for visual display of environmental variables and derivatives of those data.

We gained insight to the prevailing landscape of competitive tension when both NIWA and MetService declined to be involved in the data sharing exercise with government agencies. This was a surprise, because such an exercise was a demonstration of a pathway to impact for the products and services that these organisations provide. In times of national emergency, such tensions and the consideration of environmental data as proprietary IP is clearly not in the country's best interest. This is a systemic and enduring issue, and one that we hope will be considered and addressed in the Te Ara Paerangi – Future Pathways Programme.

In contrast to NIWA and MetService, the inclusion of data from Predictwind, a New Zealand based service providing ultra high-resolution wind forecasts, was very straightforward. These data were utilised during the LINZ exercise and were the best wind information available. The DataMesh integration showed the potential to add commercial service providers into a national capability, and have these resources to hand during a national emergency and potentially as an ongoing resource for other uses. We remain hopeful that the national agencies can find a way to incorporate data sharing under emergency situations into their business model.

Oceanum team insights

The regular TOG meetings had a positive and cohesive effect on the cross agency engagements. In hindsight we should have brought more technical updates to these meetings so that the partners were closer to the developments as they occurred.

The collaboration with all parties in the POC was effective, enjoyable and efficient. Each agency had their own use case and idea of what they wanted from the POC. It was challenging to balance different expectations against what the Oceanum DataMesh provides and the time available to configure and customise the core platform to optimise each set of needs.

Clearly defining the functional scope for any future cross-agency capability would be important to avoid any disappointment or misunderstanding of what it might be capable of. In particular in defining the specific scope related to:

1. Data discovery, searchability and categorization
2. Connectivity to external data services and associated administration
3. Storage capability
4. Data import functionality for a wide range of data types
5. Subsetting, interpolation, aggregation and transformation
6. Format conversions for downstream data uses
7. Support for different data models (e.g. vector, raster, multi-dimensional)
8. Support for live and/or continuously updating data
9. Authorization and permissions around datasets and their administration
10. Data delivery APIs, compatibility with community standards (e.g. OGC)
11. Administration APIs for managing registration, metadata, authorization and quotas
12. Visualisation of datasource metadata
13. Visualisation of the actual data
14. Advanced dataset analysis

15. Ability to create collaborative packages or workspaces
16. Toolsets to enable development of external facing apps
17. Hosting apps and analysis answering specific environmental questions
18. Deep integration with existing external applications and toolsets
19. Public access to data and functionality
20. User Interface for all the above

It is worth noting that the set features and requirements for a platform suitable for marine data would in fact satisfy numerous other data domains within government and industry. On this topic, the DataMesh is being deployed into a range of industrial and science research applications (port, navigation and offshore facilities operations, object detection from space, national aquaculture investigations, wind energy prospecting, defence science activities, the NZ Antarctic Science Platform, and more).

While the Oceanum DataMesh is an existing platform that was used to test the POC, any other platform that satisfies a defined set of requirements would of course be appropriate. However it is our view that internal development of such a platform within New Zealand government agencies would be extremely challenging.

Overall, the Oceanum involvement in the POC confirmed the company's vision and purpose, which is to provide enabling technologies to deliver real-world outcomes from environmental data. Without in any way implying criticism, it has reinforced our view that government and national science agencies are currently poorly-equipped internally to provide the necessary data infrastructure, data engineering and end-user tools to support their core purpose and maximise the return on public investment. We speculate that some of the reasons include the persistence of relict systems that gather increasingly high levels of technical debt and the prevalence of proprietary and vendor-locked solutions throughout the agencies. Our view is formed from the rapid advances that have been made in the open source data science community in recent years, building on cloud-native architectures and deploying in collaborative frameworks that adhere to transparent standards.