Data Management Architecture to enable Multinational Co-operation

Andrew Conway, Marine Institute (Ireland), andrew.conway@marine.ie
Adam Leadbetter, Marine Institute (Ireland), adam.leadbetter@marine.ie

Wastage, expending resources in a needless way, is a prevalent risk across almost every project undertaken. It is often seen as an unavoidable byproduct of undertaking any project. In particular, it has a tendency to negatively impact projects which involve more than one organisation. Communications are duplicated and parallel systems are often spun up, leading to redundancy. Data management efforts are not exempt from this, but tactics can be deployed to mitigate against some of the larger resource wasters. The INTERREG VA COMPASS project, a cross-boundary project involving organisations from Northern Ireland, the Republic of Ireland and Scotland, aimed to establish a regional network for ocean monitoring of marine protected areas, incorporating the application of tactics to minimise wastage within the project.

Instead of each partner organisation developing their own individual data management processes in isolation, with limited consultation or collaboration, the COMPASS project took the approach of agreeing on architecture at the project's outset. This architecture informed the data management choices that would be made by each partner. Key to the design of the core data architecture was a tacit agreement that all project data, systems and resources would be shared without discourse.

In terms of the local technical infrastructure used by each partner, it was decided early on that no prescription would be made to each partner on how their data would be stored (for example, SQL or THREDDS). This eliminated effort that might be wasted by any organisation having to translate data from their preferred local data management solution in order to conform to a data structure set out exclusively for a single project with finite resources and time-frame.

Once that had been decided, methods for transformation and sharing of the data were agreed upon. A federated data management system, where nodes of data are held and controlled at a remote location, was determined to be the most efficient mechanism for data sharing. The principles of federation keep a single data source for each partner's data and remove the need to keep multiple replicated sources synchronised across the project as updates occur. This was particularly beneficial when data were being added in near-real-time and also when the partners were from different nations which may have specific local legislative or community requirements for environmental data publishing (e.g. EU's INSPIRE or MEDIN standards in the UK). The system used to provide federation within the project architecture was NOAA's ERDDAP data server.

The ERRDAP data server is open source software written in Java that provides a platform through which data can be shared between partners and also published more widely. Developed out of NOAA's Monterey Laboratory, the platform-agnostic ERDDAP, or the Environmental Research Division's Data Access Program, builds upon the open-source ideals of the OPeNDAP, WCS, SOS and OBIS standards. It offers a consistent, yet an easy-to-use way of downloading or viewing scientific data in a variety of formats. Through all of this, ERDDAP also provides the ability to generate RESTful web API links in a

relatively straightforward manner, which makes the process of integrating data into web-based applications simple.

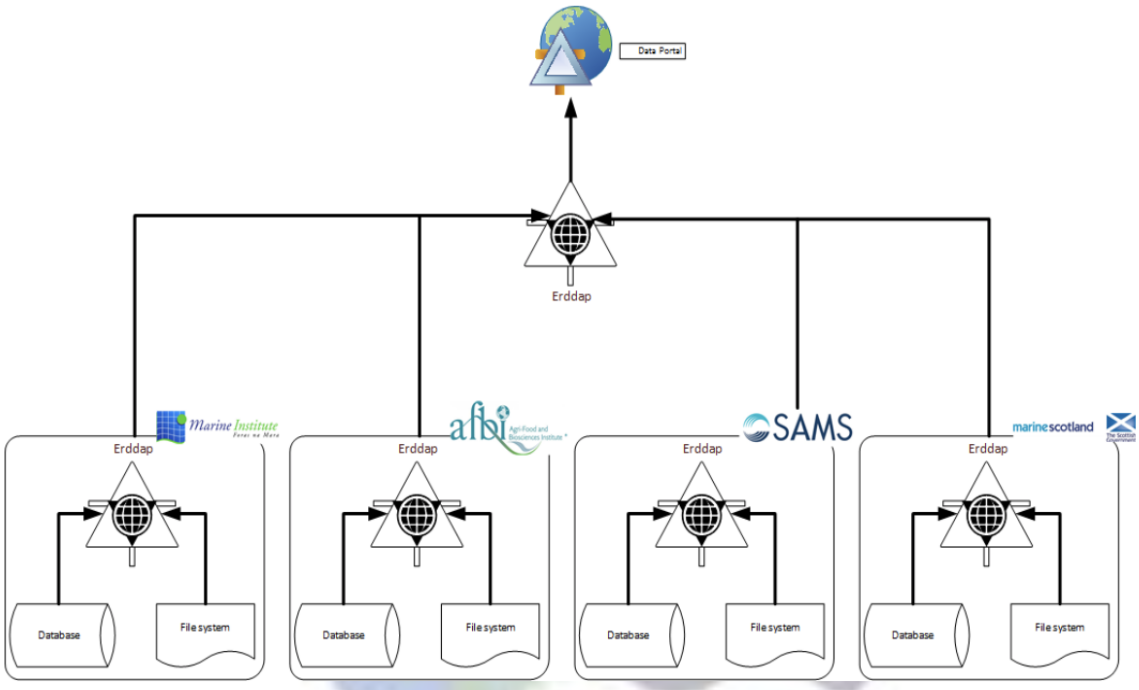

Figure 1. Anticipated high-level architecture for the COMPASS data management architecture

The use of a federation of ERDDAP servers (Figure 1), deployed at each partner location, guaranteed that each contributing organisation could maintain full control of their data while also contributing to a larger dataset, which was not dependent on any single partner to keep publicly available. Previously, this could only be attained through a single partner hosting the sum total of the project output data. This generally would involve significant duplication of data and a significant monetary cost. Further, each partner is equally involved in data publishing. Rather than one partner ring-fencing the data management/publication role, the skills obtained as well as the infrastructure deployed, are available for future use by all partners.

The establishment of these ERDDAP nodes now could have potential benefits in the future. Rather than the accepted norm of project-specific web portals for hosting data, which eventually fade into oblivion as projects age following completion, these ERDDAP servers can now act as a ready-made hub for current and future data. A beacon for publishing and a single point for data management and maintenance. Thus, the use of this data management architecture has alleviated those costs and enabled more time and resources to be spent on scientific analysis both now and hopefully into the future.

Overall, the data management architecture generated for the INTERREG VA COMPASS project, shows that through data management, multinational cooperation can be cultivated.