NDP MARINE RTDI STRATEGIC PROGRAMME 2005

BIDI

Biological **D**ata **I**ntegration

Geographic Area: Irish coastline
Project Coordinator: Valerie Cummins
CMRC Contact: Yassine LASSOUED

Project Final Report

Due date of deliverable: 31 January 2008
Actual submission date: 18 April 2008

Start Year of project: 2005 **Duration:** 2 years

Lead contractor for this deliverable: CMRC
Contributing contractors: MI

Version 0

Dissemination Level		
PU	Public	
CO	Confidential, only for project participants	**X**

Executive Summary

The overall objective of the BIDI (**Bi**ological **D**ata **I**ntegration) project is to review existing biological datasets within the Marine Institute, analyse them with respect to integration with both Arc Marine and the Marine Institute Data Model (MIDM), and assess the scientific value of undertaking this process.

This document constitutes the BIDI project final report. It provides a summary of the work that has been done as well as the results that have been achieved during the lifespan of the project. The report also shows how the project objectives have been achieved.

Grosso modo, the project's objectives have been successfully achieved despite the human resources problems faced in the middle of the project. The project succeeded in delivering a high quality integrated biological data model based on the Arc Marine data model and compliant with the existing Marine Institute Data Repository. The selection of the data design options that led to the final data model was based on standard criteria including both physical and logical levels. These criteria were prioritized by MI data custodians, users and managers, as well as the project partners, and were therefore the basis for the design options selection. Feedback from all stakeholders was extremely important.

The final data model was extensively validated by all project partners, both from the logical and the domain point of view. It was implemented and a sample geodatabase is available now (see http://workshop1.science.oregonstate.edu/fri07). Data were successfully loaded into a sample geodatabase. In addition, a report was generated specifying the necessary reformatting and transformation processes needed for loading the current MI data into the geodatabase.

Finally, the integrated biological data model has been well promoted within the MI, but also among international institutions such as the Danish Hydrological Institute (DHI) – Water and Environment, and the British Oceanographic Data Centre (BODC).

Contents

1 Introduction

This document constitutes the BIDI project final report. It provides a summary of the work that has been done as well as the results that have been achieved during the lifespan of the project. The report also shows how the project objectives have been achieved.

The document is organized as follows. Chapter 2 summarizes the work that has been done as part of the BIDI project as well as the achieved results. The chapter consists of four sections, each summarizing the work and results within a work package (WP1 to WP4).

In Chapter 3, we provide a brief summary of the difficulties that have been encountered during the project. Chapter 4 is dedicated to lessons that have been learned from the BIDI project. Finally, we conclude in chapter 5.

2 Work and Achievements

In this chapter, we run through the project work packages and summarize the work that has been done, and the results that have been achieved, within each work package. We also report on the methodology that we have adopted as well as its efficiency in terms of satisfying the project's initial objectives.

The chapter is organized in five sections, each focusing on one of the project's five work packages: WP1: Review, WP2: Refine, WP3: Extend, WP4: Synthesize, WP5: Project Management. Then, a final section is aimed to conclude.

2.1 Work Package 1: Review

As per the project description, work within WP1 has been structured and carried on according to the following four tasks:

1. **T1.1:** Training needs assessment,
2. **T1.2:** Review of the ESRI Arc Marine data model (AMDM),
3. **T1.3:** Knowledge of 'real world' MI work flows and datasets,
4. **T1.4:** Identification of user needs.

As part of T1.1, we identified the need of trainings in relation to the AMDM by different means, including:

- AMDM Documentation available through the AMDM Web site,
- Online tutorials available through the ESRI Web site,
- A technical workshop on the Marine Data Repository (MDR) and its AMDM-based model (Marine Institute Data Model, or MDM),
- Documentation on previous and ongoing projects and use cases that implemented the AMDM, a number of them being available through the AMDM Web site.

Training by all these means allowed us to undertake an overview of both the AMDM and the MDR, which is the objective of T1.2. The study highlighted the AMDM advantages (such as intuitiveness of the model, interoperability, autonomy, ease of database management, etc.) vs. its disadvantages (such as model complexity, need of robust infrastructure, high budget, etc.). Conclusions and recommendations have been made in relation to using the AMDM. The main outcome of them is that the AMDM should be used for building the MI biological geodatabase in order to facilitate integration with the existing MDR; meanwhile the output model should be kept intuitive and should be designed in a way that does not compromise the system's scalability and performance.

In addition to the technical MDR workshop, a data familiarization workshop was organized. While the former was focused on the data storage work flows (from the data owners to the data managers), the latter was more focused on the data collection work flows (from the data collectors to the data owners). Both workshops with available documentation on the MI Web site allowed a very good understanding of the real world MI work flows and datasets. The data familiarization workshop, in addition, was an opportunity to meet with data owners and helped in identifying user needs which is the object of T1.4.

2.2 Work Package 2: Refine

Work package 2 has been structured and carried out according to the following four tasks:

1. **T2.1:** Review Marine Institute biological datasets,
2. **T2.2:** Review the AMDM schema/class diagrams,
3. **T2.3:** Extend base classes in the model that best match the MI datasets,
4. **T2.4:** Align physical model with extended base classes.

The data familiarization workshop has been a starting point for the identification of prioritized application and datasets. Further communication with the MI data management group as well as the data custodians allowed us to prioritize a number of fisheries surveys and Harmful Algal Blooms (HAB) datasets. These datasets are:

- Acoustic surveys,
- Nephrops underwater TV surveys,
- Larval surveys,
- Deep water surveys
- Groundfish survey
- Harmful algal blooms

Intensive effort has been spent on T2.2 in order to understand the AMDM. A deep understanding of the AMDM has been reached by using all the methods identified in T1.1. This allowed us to identify the AMDM elements useful for the representation of the MI biological datasets. Work within task T2.2 also included familiarization with the MI data model, which extends a former version of the AMDM, with the aim to identify commonalities with the biological data. Elements such as the MDM business and Measurement tables have been identified as common entities with the targeted biological data model.

As part of tasks T2.3 and T2.4, a good understanding of the existing biological datasets' schemas and terminologies was necessary before starting the modeling phase. In the next step, we integrated the existing biological data schemas into one integrated schema independently from the AMDM. Work here built on the ongoing FSS Surveys data integration project. The so-obtained schema in called integrated biological data schema (Figure 2-1).

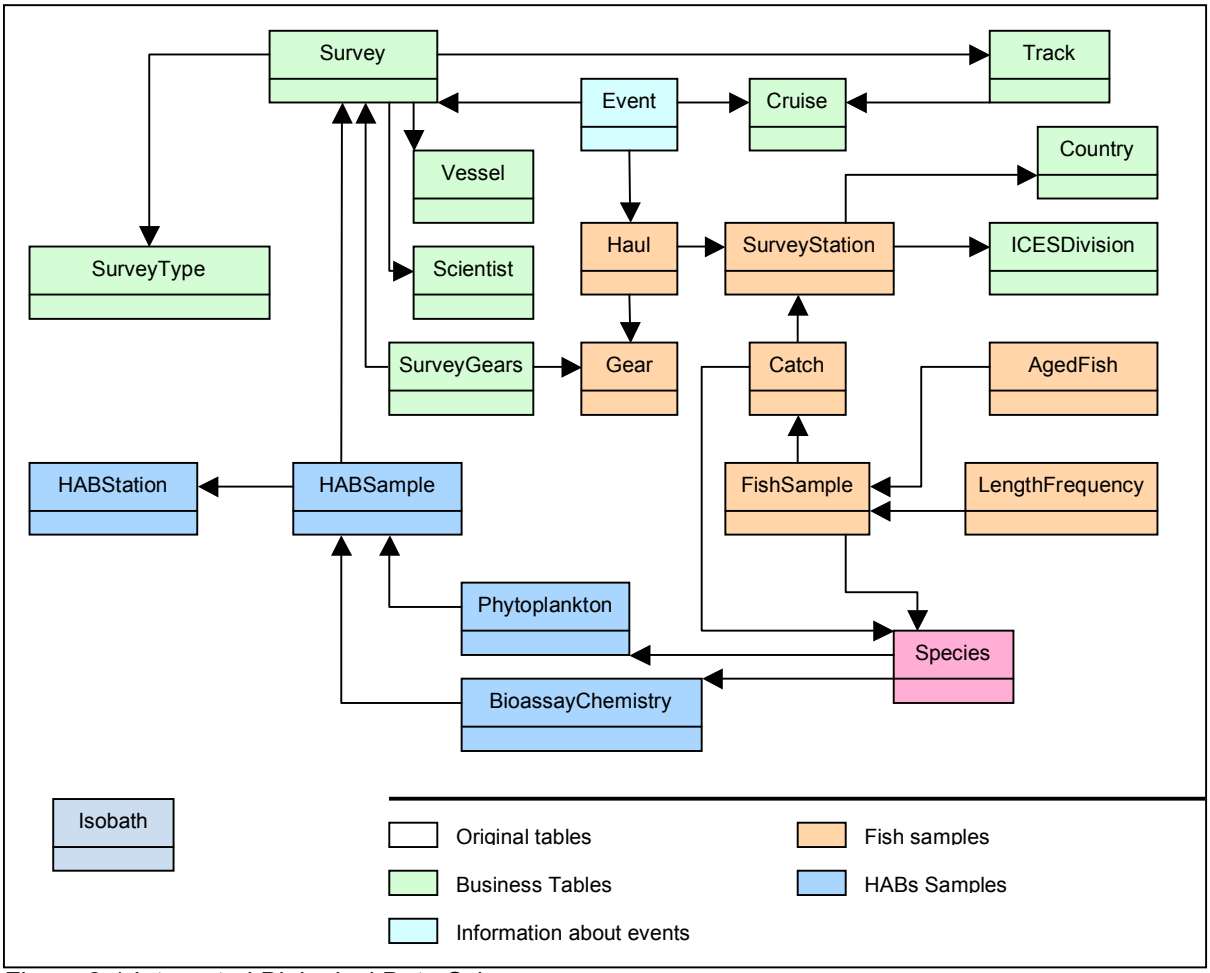

Figure 2-1 Integrated Biological Data Schema

Then, we proposed an alignment of the integrated schema with the existing MDM. Aligned entities concerned only business and measurement tables. This resulted in a new data model (schema) called aligned integrated schema (c.f. Figure 2-2).

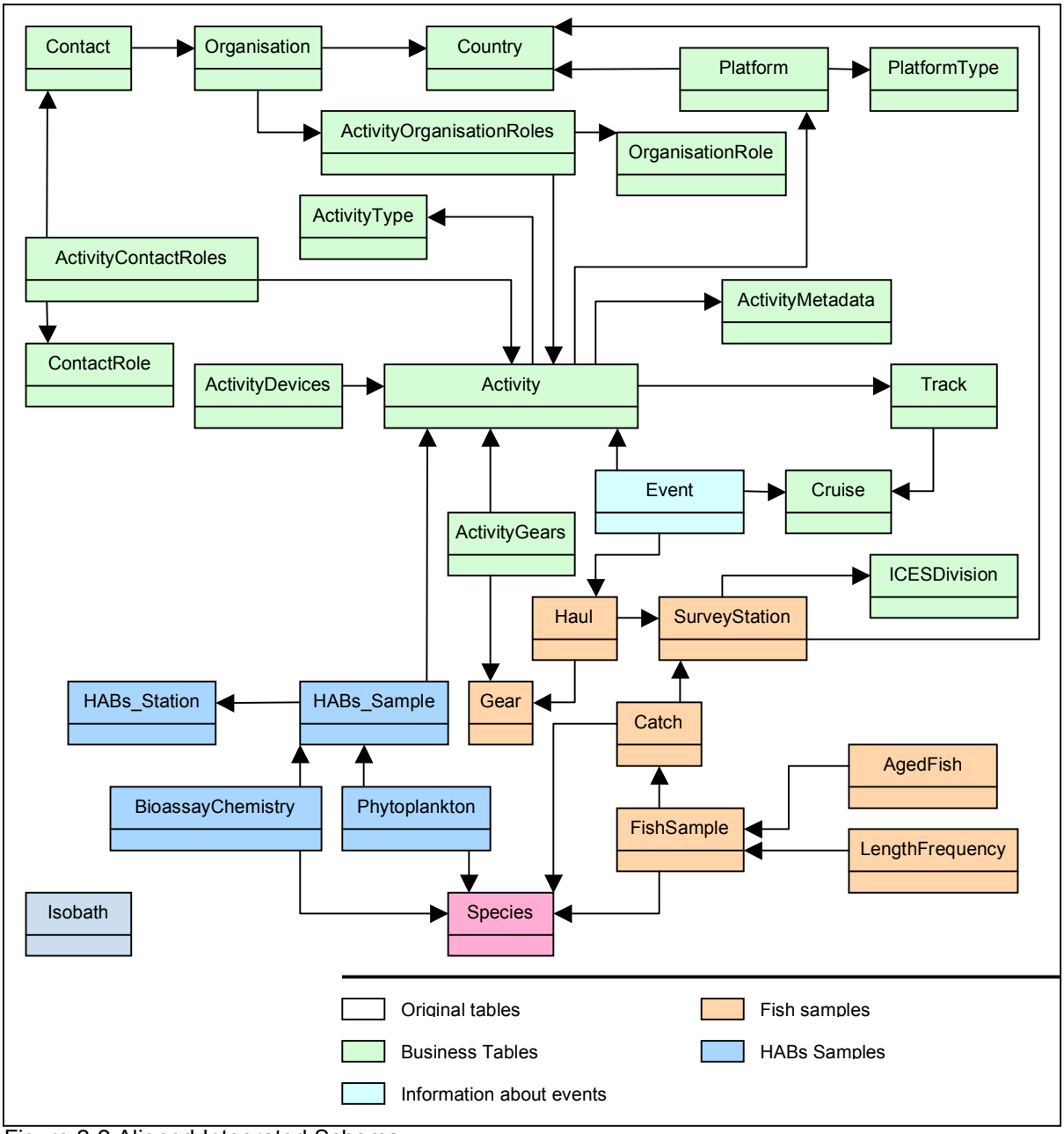

Figure 2-2 Aligned Integrated Schema

Once the integrated aligned schema was defined, we focused on aligning it with the AMDM by extending the latter. We proposed several options for extending the AMDM in order to support the integrated aligned biological data model. This work was iterative as it needed input from all the project partners as well as the data custodians and the MI data management group. Also, we defined a set of criteria, both from the physical and logical levels, for evaluating the quality of the different proposed design options. Logical level criteria included ease of understanding, ease of correct query formulation, harmony with the MDM, and harmony with the AMDM. Physical level criteria include normalization, storage efficiency, access efficiency, scalability, genericity, and ease of data transformation and loading. Because no data design option could be the best (or better than another) in an absolute way, it was crucial to understand the data custodians', users' and managers' priorities. Results from this work are available as part of the Data Design Options report which has been circulated among the project partners but also among the data custodians.

Based on the different evaluation criteria as well as iterative feedback from the different partners, and data custodians, users and managers, one data design option was selected at the end.

2.3 Work Package 3: Extend

The objective of work package 3 is:
1. **T3.1:** Extend the prototype database, the Marine Data Repository, to include selected biological datasets.

The data design option selected as result of tasks T2.3 and T2.4, was implemented in MS Visio 2000 as an extension to the physical MDR model (schema) and the AMDM. Further syntactic, logical and domain validation processes were reviewed by the project partners in order to ensure that the final physical data model be fully valid. The result is illustrated in the class diagrams of annex1.

OSU implemented a sample geodatabase using the final validated model. Considerable effort was spent in extracting, reformatting, transforming and loaded data into the deodatabase. Results of this work are available as a report which is included in annex 2 of this document. The report also highlights the major difficulties with existing data, their formats, their typing, structure, etc. It also specifies the necessary reformatting and transformation work needed for loading the current MI data into the geodatabase.

2.4 Work Package 4: Synthesize

As per the project description, the objective of work package 4 is:
1. **T4.1:** Review and promote the potential of the integrated data model,
2. **T4.**2: Develop tools to support enhanced data analysis based on the organization of data in the AMDM.

The potential of the integrated data model has been promoted at several meetings and workshops. For example, a meeting with the biological data custodians was organized in May 2007 at the MI. This meeting was the opportunity to present initial versions of the biological data model, then at its development phase, to the data custodians, to get their feedback on the data design, and also to promote the integrated model, which was very much appreciated by them.

Also, in July 2007, a BIDI session was organized in conjunction with the Coastal Atlas Interoperability Workshop (http://workshop1.science.oregonstate.edu/fri07). The session gathered all available? project partners (CMRC, MI, OSU), but also scientists and researchers from various organizations such as the Danish Hydrological Institute (DHI) – Water and Environment, the British Oceanographic Data Centre (BODC), the Oregon Coastal Management Program (OCMP) and the Department of Geography at the University of Washington. The objective of the session was to gather input from the various scientists and to promote the potential of the data model.

Also, as part of dissemination and of task T4.1, the BIDI data models were published on the main Arc Marine Web site (http://dusk.geo.orst.edu/djl/arcgis).

For human resources issues that will be explained in chapter 3, the progress on task T4.2 has not been as effective as progress in other tasks.

3 Difficulties

The difficulties experienced in the BIDI project are mainly human-resource-related. In 2006 Dr. Eamonn O'Tuama (project coordinator) left CMRC. A few months later (April 2007), Josu Ramirez, the main technical member of BIDI, left CMRC as well. A handover needed was necessary to insure the project continuity. But, for human resources issues (only two IT researchers left in CMRC), delays were not avoidable.

4 Lessons Learned

There are many lessons that we learned during the BIDI project both from good and bad practices.

Among the lessons learned from our successful practices we insist on the importance of familiarization workshops which allow rapid and good understand of the data that we are working with as well as their work flows.

Also, it is important to consider several modeling options at early stages of the data design and to have a set of standard evaluation criteria for these options. Communication and negotiation between all project partners and with data users, custodians and managers is crucial for understanding everybody's priorities (e.g. performance vs. normalization, etc.) before selecting data design options.

We learned that it is very important to have some implementation and prototyping tasks at early stages of a data modeling project as opposed to the idealistic theoretical approach. The idea is to detect as many problems as possible beforehand.

5 Conclusion