



National Development Plan 2007—2013  
Science, Technology & Innovation Programme

## **Marine Research Sub-Programme**

**Call for Research Proposals  
July 2007**

### **KNOWLEDGE AND INFORMATION MANAGEMENT RESEARCH PROGRAMME**

## TABLE OF CONTENTS

<b>Project 1: 'Defined' Strategic Project</b>	Data Management for Marine Geological and Geophysical Datasets (including Integration with the Marine Data Repository)
<b>Project 2: 'Defined' Strategic Project</b>	A Practical Implementation of Data Mining Techniques for the Collection and Analysis of Very Large Marine Datasets

<b>Office use only:</b> <b>Project Reference:</b>	<b>PBA/KI/07/001</b>
--	----------------------

<b>NDP 2007-2013:</b> Science, Technology and Innovation (STI) Programme:  <b>Marine Research Sub-Programme</b>	
<b>Research Measure:</b>	<b>Policy Support</b>
<b>Research Programme:</b>	<b>Knowledge and Information Management</b>
<b>Project Type:</b>	<b>'Defined' Strategic</b>
<b>Project Title:</b>	<b>Data Management for Marine Geological and Geophysical Datasets (including Integration with the Marine Data Repository)</b>

**BACKGROUND:**

The Marine Institute has evolved over the last 16 years from an initial staff of 2 to approximately 200 today. The Institute's work programmes gather and produce a wide variety of marine datasets (oceanographic, environmental, biological and chemical data) in a diverse range of formats and standards (details can be found on [www.marinedataonline.ie](http://www.marinedataonline.ie)). Management of marine resources is moving from assessment of a single discipline to a more holistic or "ecosystem" approach, and is driving the demands for greater integration of data in order to derive information about the overall status of the environment. One of the most significant challenges arising from this is the collation and integration of these datasets in order to present them in an accessible and intuitive way.

A number of programmes have been initiated within the organisation to address these issues. The Marine Data Repository (MDR) spatial data warehouse project is one of these. The MDR integrates key physical and chemical datasets in a spatially enabled database and presents the data for query, visualisation and extraction through a web-based GIS interface. The MDR is based on the ArcGIS Marine Data Model (MDM), which has been developed over recent years by the marine community in collaboration with ESRI (<http://support.esri.com/datamodels>). The MDM has been developed using a community-based approach thus reflecting a collective experience in the storage and use of marine datasets in research. A considerable amount of information has been collated and made available publicly at <http://dusk.geo.orst.edu/djl/arcgis/index.html>.

Sea Change - A Marine Knowledge Research & Innovation Strategy for Ireland (2007-2013)<sup>1</sup>, includes a proposal for a National Coastal and Marine Information Service, whose objectives would include developing "the capacity and expertise required to integrate data and information in order to improve the quality of scientific advice", ensuring "that relevant data generated by research projects is passed, where appropriate, into use in monitoring and management programmes". The related RTDI Requirement is to provide "Integrated Data Management Programmes that focus on the development of frameworks for the integration of data from multiple sources—areas to be addressed include diverse data types, varying spatial

<sup>1</sup> [www.marine.ie/seachange](http://www.marine.ie/seachange)

resolutions, varying temporal resolutions, currency, archival, large volumes, semantics & ontologies”.

This proposal also links to objectives of other Sea Change programmes including the Policy Support Measure: Marine Environment Research Programme, the Industry Research Measure: Fisheries Resources Research Programme, looking to “Map the spatial and temporal distribution of spawning and nursery areas for fish stocks in the waters around Ireland and integrate with seabed survey data and oceanographic data” and the Discovery Research Measure programmes: e.g. the Rapid Climate Change Programme.

Over the past number of years the Irish National Seabed Survey, MESH and currently INFOMAR programmes have all gathered large volumes of geophysical/geological data in various file formats. As a partner in these programmes the MI has an interest in the management and interrogation of these datasets. Research is required to identify how best to store and manage these large volumes of geophysical/geological data, to ensure the security and optimal loading and retrieval of the data, and to also ensure that the data can be shared and integrated with other datasets. It is anticipated that the integration of geophysical/geological data with core physical, chemical, and biological data will support the development of multivariate spatial modelling techniques, increasing our understanding of environmental and biological processes in operation.

This research project will identify best practices for the data management of these geophysical/geological datasets, including data modelling which is based on or allows for integration with the MDR/MDM, and which can deal with storing and making accessible the very large volumes of data in question. The research will include a review of existing best practice in data management processes which have been implemented for similar data by other organisations internationally, including the use of standards e.g. for data interfaces. The investigation should also include how marine geophysical / geological data has been made available by these organisations to the wider public and scientific community e.g. OGC web services and GIS clients, and what products and services have been made possible using this data.

### **PROJECT RATIONALE:**

The aim of this project is to advance data management methods as they are applied to the geological and geophysical data in the Marine Institute. Geological / geophysical datasets have their own particular characteristics, including very large volumes of data, which need to be addressed. A specific example is the data gathered by the INFOMAR, and previously INSS, seabed survey programmes which have generated significant volumes of data (5 TBs+). Improvements in the processing and storage of the data will allow enhanced usage and analysis of the data, thus providing a greater return on investment for the often very large investment in capturing the data.

### **PROJECT AIMS:**

This research project aims to produce a reference for the storage and management of geological / geophysical datasets in the Marine Institute, with the goal of making the data easily available for discovery, access and analysis by Marine Institute and external end users.

The data generated by the INFOMAR programme will be a key dataset for the study. A major requirement will be the ability to integrate with (and evaluate using) the ArcGIS Marine Data Model as implemented in the Marine Data Repository by the Marine Institute.

## RESEARCH OBJECTIVES:

The Marine Institute wishes to invite suitably qualified researchers/research teams to:

- Review existing geophysical/geological datasets within the Marine Institute and its partners. This would include analysing them with respect to integration with other datasets and assessing the scientific value of so doing, including:
  - Identifying a prioritized list of datasets for assessment.
  - Identifying any data issues such as quality, resolution or metadata.
  - Reviewing the processes used to collect the data.
  - Reviewing dataset semantics, including mapping to standard ontologies and controlled vocabularies to both understand the data and facilitate the data discovery and integration processes.
  - Identifying any technical issues regarding modeling as a result of the analyse of the datasets.
- Review international best practice for the management of marine geophysical/geological data.
- Specify a suitable data model to cater for the chosen geophysical/geological datasets, which will allow the data to be integrated with other Marine Institute data holdings. This may include using the ArcGIS Marine Data Model, including features classes such as the Regularly Interpolated Surfaces feature class for bathymetric data and the Feature Lines feature class for sea wall or a shoreline data.
- Define data transformation and load routines as required, based on the semantic mappings already identified.
- Specify a process for the generation of granular metadata (preferably automated) sufficient to allow successful identification, location and analyse of data by users.
- Specify suitable data output / delivery methods, for dissemination and integrated analysis of the data.
- Identify process changes which would improve the management of the data.
- Implement a prototype data storage and retrieval system to test the model implementation, data loading and retrieval, including demonstrating support for open standard approaches to data discovery, data integration and analysis.
- Assess potential analyses and services that can be made available internally in the Marine Institute and externally to the wider scientific community and general public by integrating geophysical/geological data with other Marine Institute datasets.
- Identify and evaluate tools and approaches / workflows to facilitate geospatial analysis and querying of the geophysical/geological data.

## PROJECT DELIVERABLES:

1. An initial report specifying:
  - A prioritised list of datasets
  - A review of dataset semantics, including initial mapping to a standard ontology
  - Identification of any technical issues with the model structure as a result of the analyse of the datasets
  - Identification of any data issues such as quality, resolution or metadata
  - A high-level outline of the processes used to manage the data
2. A report on international best practice for the management of large volumes of geophysical/geological data
3. A document outlining possible modeling and controlled vocabularies / ontology options
4. A document outlining possible options for process change to improve the management of the data
5. A specification detailing the complete proposed solution, including details of the data management process changes, data mappings to the standard vocabulary / ontology, details of the data model to be used to store the data, details of how the data transform

- and load routine will operate, as well as details of data integration opportunities, data output services and data analysis opportunities, including possibilities for use of GIS tools
6. A prototype based on the above specification, to act as proof-of-concept for each area specified, i.e. data model, data transform & load, data integration, data output and data analyses including usage of GIS tools
  7. A report with a full evaluation of the specification and prototype
  8. A final project report, including implementation recommendations

### **ADDITIONAL SPECIFIC REQUIREMENTS FOR THIS PROJECT:**

The researcher(s) involved in this project will be required to liaise with Marine Institute personnel to identify specific datasets, processes used, and other project inputs.

This research should also take note of and compliment the ongoing work being done by Marine Institute partners in this area, notably by the GSI under the NDP Geoscience Sub-Programme (Science, Technology and Innovation Programme). To this end it is envisaged that there will be coordination and information sharing between this project and one or more projects under the Griffith Awards programme.

### **PROJECT STRUCTURE AND FUNDING:**

The project should be managed by a project co-ordinator, who will be responsible for ensuring that project management formalities, as well as research outputs, are delivered in a timely and presentable manner. The proposal should also clearly outline time commitment of existing and additional researchers. Liaison with Marine Institute personnel will be required on an ongoing basis. Funding will be provided for a 3 year project.

Further information is available in the *Guidelines for Applicants and Guidelines for Grantees for Project-Based Awards*.

### **INTENDED IMPACT:**

This research will provide best practice data management guidance for the complete management and analysis of geophysical / geological data within the Marine Institute. This will reduce the effort required to manage the data from capture to delivery and analysis, and will enhance the ease of access to geological / geophysical data within the Marine Institute. This project will also allow integrated analysis with other Marine Institute datasets, including those already held with the Marine Institute's Marine Data Repository data warehouse.

### **ADDITIONAL INFORMATION:**

#### ***ArcMarine Data Model***

"The ArcGIS Marine Data Model represents a new approach to spatial modelling via improved integration of many important features of the ocean realm, both natural and manmade. The goal is to provide more accurate representations of location and spatial extent, along with a means for conducting more complex spatial analyses of marine and coastal data by capturing the behaviour of real-world objects in a geodatabase. The model also considers how marine and coastal data might be more effectively integrated in 3-D space and time. Although currently limited to 2.5-D, the model includes "placeholders" meant to represent the fluidity of ocean data and processes. A key advantage of the data model is that it should help users to take fuller advantage of the most advanced manipulation and analysis capabilities of ArcGIS, particularly its support of more complex rules that can be built into its geodatabases, and of objects with not only attributes, but database rules (rudimentary behaviour)." -

<http://dusk.geo.orst.edu/djl/arcgis/index.html>.

The Marine Institute has adapted this model to best fit the requirements of the organisation..

**Reference Standards & Applications**

- The ArcGIS Marine Data Model (MDM) (<http://support.esri.com/datamodels>)
- Open Geospatial Consortium web services (<http://www.opengeospatial.org/>)
- International Hydrographic Organisation S-100 Framework
- Bathymetric Attributed Grid (BAG) format (<http://www.opennavsurf.org/>)
- GIS clients: Google Earth, NASA World Wind, web browser based applications

<b>Office use only:</b> Project Reference:	PBA/KI/07/02
---	--------------

<b>NDP 2007-2013:</b> Science, Technology and Innovation (STI) Programme:  <b>Marine Research Sub-Programme</b>	
<b>Research Measure:</b>	<b>Policy Support</b>
<b>Research Programme:</b>	<b>Knowledge and Information Management</b>
<b>Project Type:</b>	<b>'Defined' Strategic</b>
<b>Project Title:</b>	<b>A Practical Implementation of Data Mining Techniques for the Collection and Analysis of Very Large Marine Datasets</b>

## BACKGROUND:

Given the very large volumes of data arising from marine research programmes and sensor observations, the development of mechanisms to process and turn this data into usable knowledge in real or near-real time will be of great importance to research programmes in the marine sector.

The exponential growth in data volumes is leading to research organisations being unable to fully exploit the data generated i.e. becoming data rich but information poor. Automated analysis mechanisms such as data mining will become increasingly necessary for the management and exploitation of large datasets and real-time data streams. Proposed research programmes in the following areas will soon generate very large volumes of data:

- Coastal Sensor Networks
- Cabled Underwater Observatories
- Ocean Circulation and Climate Change Modelling
- Seabed & Resource Mapping
- Biotechnology / BioDiscovery

This project will investigate the practical implementation of data mining techniques<sup>1</sup>. The project will aim to select and implement working prototypes of at least two data mining techniques that would demonstrate genuine benefit to the analysis or management of marine datasets. The project will assess current datasets and investigate and assess the practical implementation of data mining scenarios that can be applied to those datasets. Examples would be rule based quality control of oceanographic data, event detection from real-time sensor data streams, ecosystem habitat classification, pattern analysis for fisheries and resource management etc.

<sup>1</sup> The Marine Institute has previously commissioned a relevant desk study "The use of data mining techniques for the quality control of scientific data" which is due to be finalised shortly.

## PROJECT RATIONALE:

The growth in scientific data volumes is presenting challenges for the management and analysis of the data. Increasingly automated techniques are required to convert huge amounts of data into useable information. This project proposes to apply Data Mining techniques to marine datasets, the principles of which can be applied to other datasets, to provide tangible scientific benefit.

## PROJECT AIMS:

To identify, implement and document data mining techniques for the management and analysis of Marine Environment Research Programme datasets. These techniques should improve the data management of the datasets and enhance knowledge discovery.

## RESEARCH OBJECTIVES:

The Marine Institute wishes to invite applications from suitably qualified researchers to:

- Identify two static and one streaming data sources from key Marine Environment Research programme datasets, e.g. Fisheries and Resource Management, INFOMAR Seabed Mapping, or OSS modelling or sensor data, to which useful and re-useable data mining analysis techniques can be applied e.g. to achieve change detection or data reduction. In the case of streaming data an external test data source with similar characteristics may be proposed;
- Identify the relevant science / data management objectives e.g. enhanced analyses, quality control, or data reduction, in conjunction with the Marine Institute programme teams to develop data mining criteria;
- Implement the data-mining criteria using the techniques identified to achieve the targeted objective; and
- Report on the implementations including a full analysis of the objectives, the processes used, the results, and a pros and cons analysis of the techniques used, with recommendations for future implementations of the techniques.

## PROJECT DELIVERABLES:

1. A report analysing the use of data mining techniques that can add value to the analysis of scientific datasets, including real-time streaming data from sensors.
2. A report analysing static and streaming data from Marine Environment Research programmes to which data mining techniques can be usefully applied (note: for streaming data, a similar external source may be proposed). This report will include possible data mining outputs for the specific programme, including their potential end-use application.
3. A documented specification detailing the programme datasets to be mined, the data mining techniques and technology to be used, the data mining criteria which will be implemented, and the expected output.
4. One or more prototype systems (as required) implementing the above specification.
5. A concluding report which analyses the methods used and the values achieved. The report should also make recommendations for the implementation of future systems, and an analysis of the reusability of the techniques (include a pro/con analysis).

### **ADDITIONAL SPECIFIC REQUIREMENTS FOR THIS PROJECT:**

The Marine Institute data infrastructure uses Microsoft technologies e.g. SQL Server. This project should use technologies which can coexist with this infrastructure and which are available without incurring significant additional licensing costs.

Liaison with Marine Institute personnel will be required to identify the datasets and processes that will form the basis for the project.

### **PROJECT STRUCTURE AND FUNDING:**

The project should be managed by a project co-ordinator, who will be responsible for ensuring that project management formalities, as well as research outputs, are delivered in a timely and presentable manner. The proposal should also clearly outline time commitment of existing and additional researchers. The project should be completed within 18 months.

Further information is available in the *Guidelines for Applicants and Guidelines for Grantees for Project-Based Awards*

### **INTENDED IMPACT:**

An increased capacity for the successful implementation of data mining techniques to add value to Marine Environment dataset analysis and management, most notably in the case of very large datasets and streaming data sources.

### **ADDITIONAL INFORMATION:**

Further information on the 2005 funded Desk Study: *The Use of Data Mining Techniques for the Quality Control of Scientific Data* is available on:

<http://www.marine.ie/home/funding/ndpfunding/FundedProjects/DeskStudies.htm>