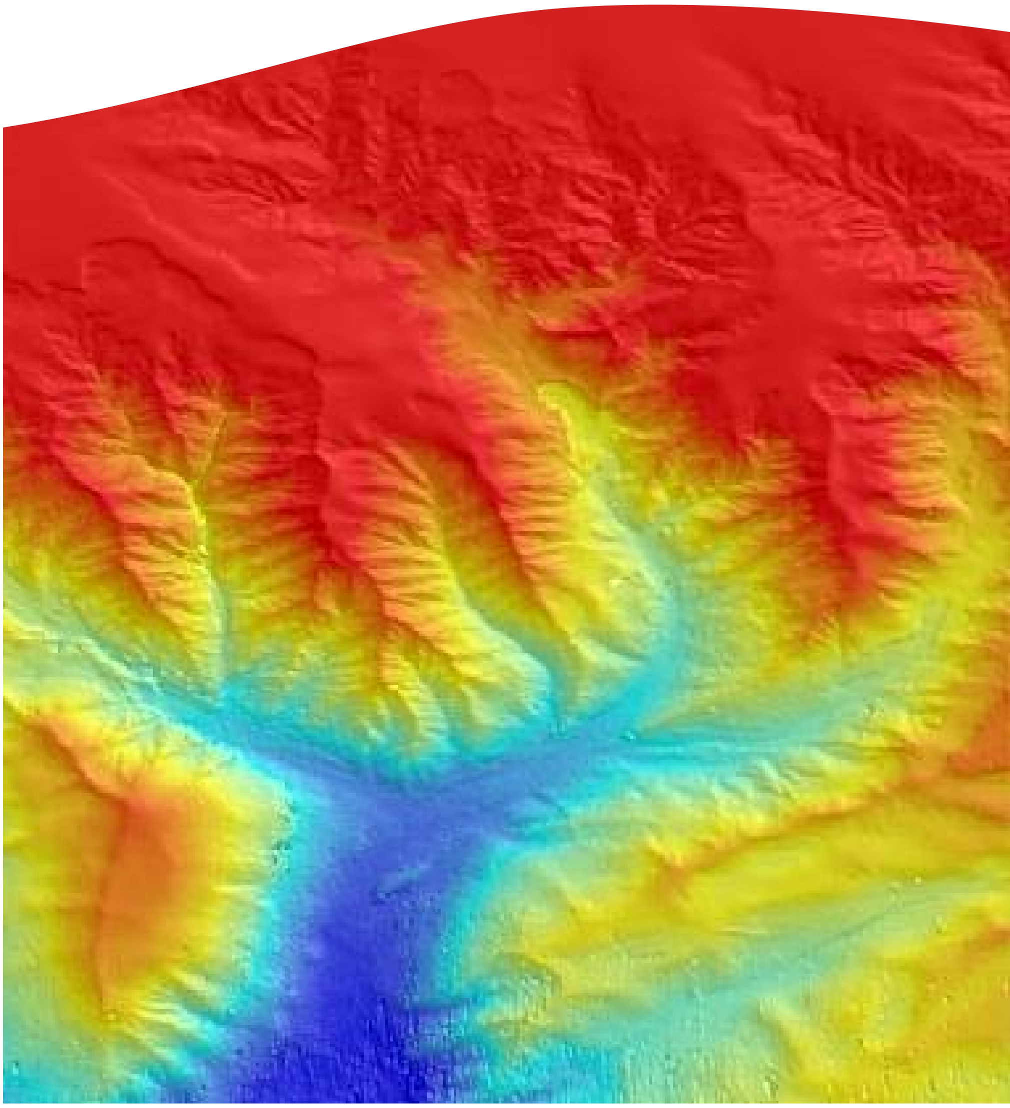
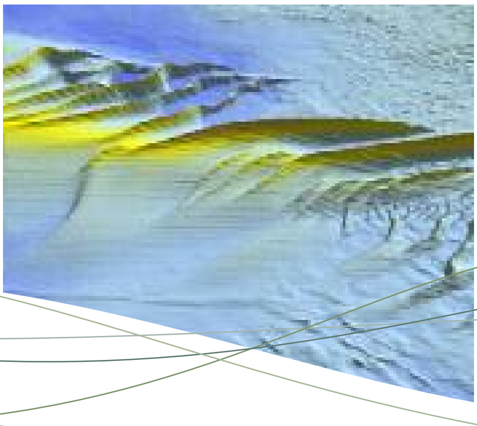
Marine Research Sub-Programme
(NDP 2007–’13) Series

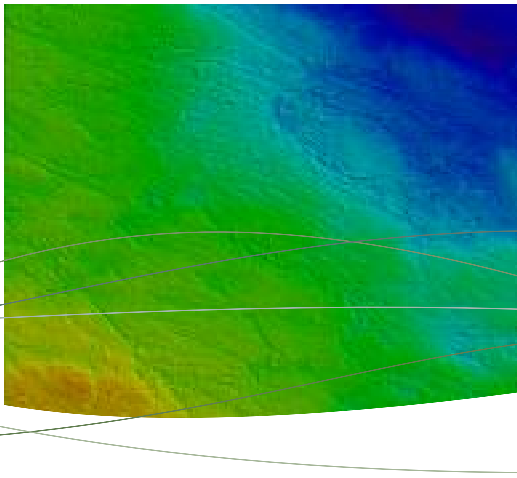
GeoDI - Geoscientific Data Integration

Project-based Award

SeaChange
Casadh na Taoide







Lead Partner: University College Cork

NDP
National Development Plan
Transforming Ireland

Marine Institute
Foras na Mara

The Marine Institute is the national agency which has the following functions:
"to undertake, to co-ordinate, to promote and to assist in marine research and development and to provide such services related to research and development that, in the opinion of the Institute, will promote economic development and create employment and protect the marine environment" Marine Institute Act 1991.

Sea Change: A Marine Knowledge, Research & Innovation Strategy for Ireland

Sea Change—A Marine Knowledge, Research & Innovation Strategy for Ireland 2007-2013—was launched in early 2007 and was the outcome of extensive analysis and consultation with government departments, state agencies, industry and the third-level sector. It outlines a vision for the development of Ireland's marine sector and sets clear objectives aimed at achieving this vision, namely to:

1. Assist existing, and largely indigenous, marine sub-sectors to improve their overall competitiveness and engage in activity that adds value to their outputs by utilising knowledge and technology arising from research.
2. Build new research capacity and capability and utilise fundamental knowledge and technology to create new marine-related commercial opportunities and companies.
3. Inform public policy, governance and regulation by applying the knowledge derived from marine research and monitoring.
4. Increase the marine sector's competitiveness and stimulate the commercialisation of the marine resource in a manner that ensures its sustainability and protects marine biodiversity and ecosystems.
5. Strengthen the economic, social and cultural base of marine dependant regional/rural communities.

The Sea Change strategy was developed as an integral part of the government's Strategy for Science, Technology and Innovation (SSTI) and the Marine Institute as the lead implementation agency is working within SSTI policy and with government departments and agencies to deliver on the Strategy.

The Marine Institute managed Marine Research Sub-Programme, one of eight sub-programmes within the Science, Technology and Innovation (STI) Programme of the National Development Plan 2007—2013, targets funding to meet the objectives of the Sea Change strategy.

Over the lifetime of Sea Change, funding will be provided for:

- Project-Based Awards
  - Strategic Research Projects
  - Applied Research Projects
  - Demonstration Projects
  - Desk/Feasibility Studies
- Researcher Awards
  - Strategic Research Appointments
  - Research Capacity/Competency Building
  - Post-Doctoral Fellowships
  - PhD Scholarships
- Industry-Led Research Awards
  - Company Awards
  - Collaborative Awards
- Infrastructure Awards
  - Infrastructure Acquisition
  - Access to Infrastructure

Further copies of this publication can be obtained from:
Marine Institute, Rinville, Oranmore, Co. Galway, Ireland or www.marine.ie

Cover Images © INFOMAR

ISSN: 2009-3195

NDP
National Development Plan
Transforming Ireland

Marine Institute
Foras na Mara

SeaChange
Casadh na Taoide

Marine Research Sub-Programme 2007-2013

Project-based Award

GeoDI - Geoscientific Data Integration

(Project Reference: PBA/KI/07/001)

Lead Partner:	University College Cork (UCC)
Project Partners:	Oregon State University (OSU) and subcontractor United States Geological Survey (USGS)
Author(s):	Yassine Lassoued
Project Duration:	01 February 2008 to 31 May 2011

Acknowledgments

This project (Grant-Aid Agreement No. PBA/KI/07/001) was carried out under the Sea Change strategy with the support of the Marine Institute and the Marine Research Sub-Programme of the National Development Plan 2007–2013.

The project was led by the Marine Geomatics Team of the Coastal and Marine Research Centre (CMRC), formerly the Coastal and Marine Resources Centre, University College Cork.

Disclaimer

Responsibility for the information and views presented in this report rest solely with the authors and do not necessarily represent those of the Marine Institute. Neither the authors nor the Marine Institute accept any responsibility whatsoever for loss or damage occasioned or claimed to have been occasioned, in part or in full, as a consequence of any person acting, or refraining from acting, as a result of a matter contained in this publication.

Project Partners

Dr Yassine Lassoued
Marine Geomatics
Coastal & Marine Research Centre
University College Cork
Haulbowline
Cobh
Co. Cork

Prof Dawn Wright
College of Science
Department of GeoSciences
Oregon State University
104 Wilkinson Hall
Corvallis
Oregon 97331-5508
USA

Table of Contents

Executive Summary

Large volumes of geoscientific (i.e., geological and geophysical) datasets have been gathered by the Marine Institute (MI) and its partners over the past number of years, notably through the current INFOMAR (Integrated Mapping for the Sustainable Development of Ireland's Marine Resource) and previous INSS (Irish National Seabed Survey) and MESH (Mapping European Seabed Habitats) programmes. A key challenge now exists to derive maximum value from these very costly and valuable products by integrating these geoscientific datasets together, and with other resources such as biological, chemical, and environmental data. This will allow for an ecosystem approach in the analysis of marine and geoscientific data, a holistic and more sophisticated view of change in the status of the marine environment, and thus improve the quality of scientific advice.

The Geoscientific Data Integration (GeoDI) project aimed to address this challenge by examining the critical issues involved in the integration of Irish marine geoscientific datasets, and by assessing tools and services for enhanced management, discovery, access, and analyses of geoscientific data.

The main outcomes and results of the GeoDI project are:

- A set of reports reviewing existing technologies, standards, models, and best practices related to the integration, management, and delivery of geoscientific datasets with associated recommendations and proposed process changes aimed at improving geoscientific data management.
- An integrated, flexible, and scalable geoscientific data model based on Arc Marine, compatible with existing MI databases (Marine Data Repository (MDR) and Biological Data Integration (BIDI) model).
- A set of geoscientific ontologies for use in data, metadata, and extract, transform, and load (ETL), built based on existing well-established vocabularies.
- A complete system specification for integrating, managing, and delivering geoscientific resources (data, metadata, and ontologies), including detailed specification of the subsystems and how they fit together and interact with each other. The specification demonstrates practical uses of ontologies and includes a semantic web service (SWS) which is being further advanced by the EU FP7 NETMAR project in order to be submitted to a standardised body.

- A semi-automatic, generic, and ontology-based ETL tool for loading datasets into the integrated geoscientific database, which may be further developed and customised to other systems or databases.
- The Integrated Geoscientific Information System (IGIS), a complete system for integrating, managing, and accessing geoscientific resources. In addition to standard services, the IGIS includes the following components which may be reused individually (e.g. by the Irish Spatial Data Exchange (ISDE)):
    - A SWS for accessing the geoscientific domain ontology,
    - A catalogue service mediator that allows access to distributed catalogue services and solves semantic conflicts between these,
    - The GeoDataOnline portal, available at http://gdo.ucc.ie.
- A report, with recommendations, assessing potential analyses and services and identifying tools and approaches to facilitate geospatial analysis of geoscientific data.
- Three publications in two conferences (Remote Sensing and Phtogrammetry Society Annual Conference, 2010 and E-Science Grid Facility for Europe and Latin America (EELA2), 2009) and a book (Coastal Informatics – Web Atlas Design and Implementation).

1. Project Description

1.1. Problem Addressed

Large volumes of geoscientific (i.e. geological and geophysical) datasets have been gathered by the Marine Institute (MI) and its partners over the past number of years, notably through the current INFOMAR (Integrated Mapping for the Sustainable Development of Ireland's Marine Resource) and previous INSS (Irish National Seabed Survey) and MESH (Mapping European Seabed Habitats) programmes. The collected datasets are generated by a variety of instruments during different surveys, and are stored in a variety of formats and representations.

A key challenge now exists to derive maximum value from these very costly and valuable products and from the national data acquisition effort, by integrating these geoscientific datasets together, and with other resources such as biological, chemical, and environmental data. This will allow for an ecosystem approach in the analysis of marine and geoscientific data, and a holistic and more sophisticated view of change in the status of the marine environment, thus improving the quality of scientific advice.

1.2. Aims and Objectives

The Geoscientific Data Integration (GeoDI) project aimed to address the challenge as outlined above by examining the critical issues involved in the integration of Irish marine geoscientific datasets, and by assessing tools and services for enhanced management, discovery, access, and analyses of geoscientific data.

The GeoDI project had the following specific objectives (SO):

SO-1. Review existing geoscientific datasets within the MI and its partners, including analysis with respect to integration with other datasets and assessing the value of so doing.

SO-2. Review international best practice for the management of marine geoscientific data.

SO-3. Specify a suitable data model to cater for the chosen geoscientific datasets, which would allow the data to be integrated with other MI data holdings such as the Marine Data Repository (MDR) and/or more recently the biological data repository designed as part of the Biological Data Integration (BIDI) project.

SO-4. Develop ontologies for geoscientific resources which would:

- Provide a geoscientific knowledge base with rich semantics that can be shared, reused, and queried over the Semantic Web (SW),
- Improve metadata interoperability by developing ontology terms for values of elements and attributes of metadata instances.

SO-5. Define data transformation and load routines based on the semantic mappings already identified. More specifically, investigate the use of Artificial Intelligence (AI) and Constraint Programming (CP) to develop automatic and semi-automatic Extract, Load and Transform (ETL) tools for marine and geoscientific data and ontologies.

SO-6. Specify automated processes for the generation of metadata sufficient to allow successful identification, location, and analyses of data by users.

SO-7. Specify suitable data output / delivery methods for dissemination and integrated analysis of the data.

SO-8. Identify process changes that would improve the management of the data.

SO-9. 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.

SO-10. Assess potential analyses and services that can be made available internally in the MI and externally by integrating geophysical/geological data with other MI datasets.

SO-11. Identify and evaluate tools and approaches / workflows to facilitate geospatial analysis and querying of the geophysical/geological data.

1.3. Methodology

The GeoDI project was implemented through a series of six R&D work packages to meet the specific objectives outlined above (c.f., section 1.2). Figure 1 illustrates the GeoDI work packages and their interconnections.

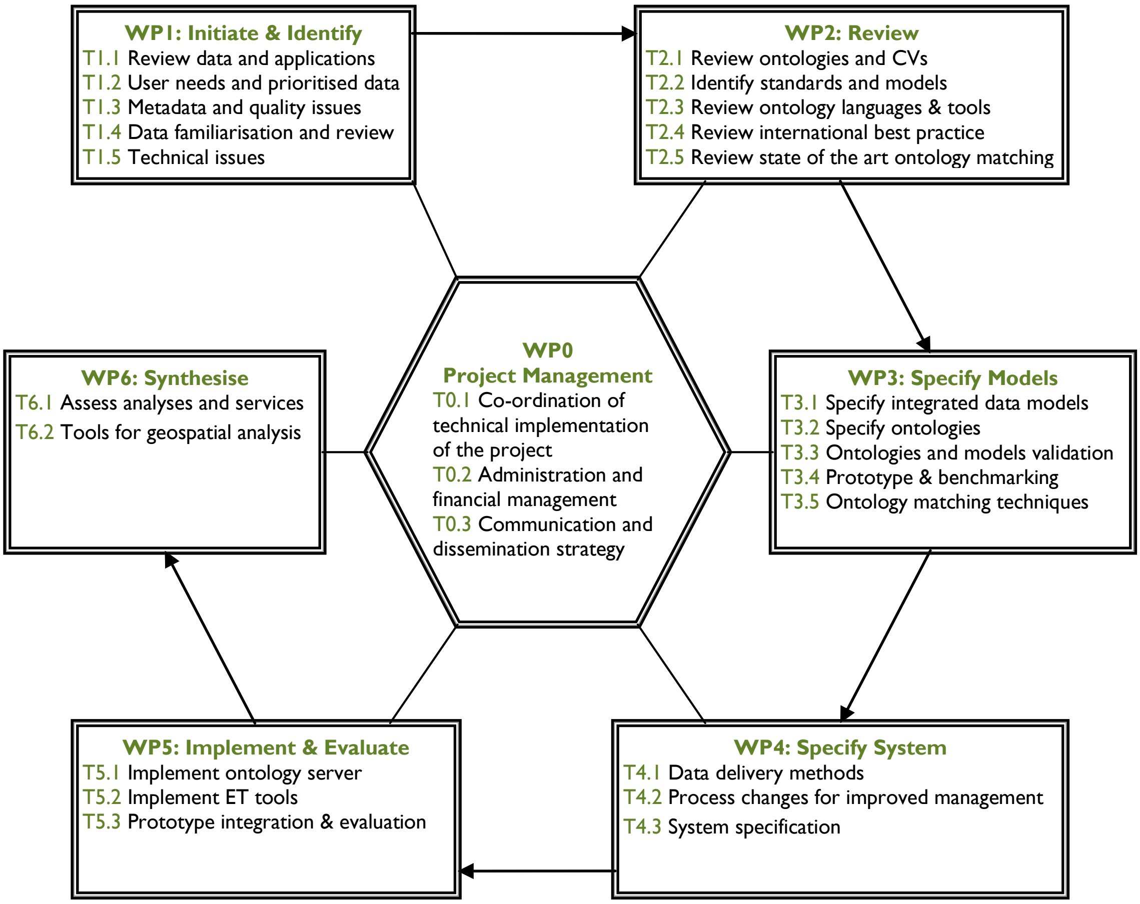

WP1: Initiate & Identify
T1.1 Review data and applications
T1.2 User needs and prioritised data
T1.3 Metadata and quality issues
T1.4 Data familiarisation and review
T1.5 Technical issues
WP2: Review
T2.1 Review ontologies and CVs
T2.2 Identify standards and models
T2.3 Review ontology languages & tools
T2.4 Review international best practice
T2.5 Review state of the art ontology matching
WP6: Synthesise
T6.1 Assess analyses and services
T6.2 Tools for geospatial analysis
WP0
Project Management
T0.1 Co-ordination of technical implementation of the project
T0.2 Administration and financial management
T0.3 Communication and dissemination strategy
WP3: Specify Models
T3.1 Specify integrated data models
T3.2 Specify ontologies
T3.3 Ontologies and models validation
T3.4 Prototype & benchmarking
T3.5 Ontology matching techniques
WP5: Implement & Evaluate
T5.1 Implement ontology server
T5.2 Implement ET tools
T5.3 Prototype integration & evaluation
WP4: Specify System
T4.1 Data delivery methods
T4.2 Process changes for improved management
T4.3 System specification

Figure : GeoDI Work Packages

1.4. Work Carried Out

The GeoDI work packages were successfully carried out and completed according to the description of work. The project partners worked closely with the Griffith Geomatics for Geosciences partners, the Joined-up Data for Joined-up Thinking partners, the EU FP7 NETMAR (Open Service Network for Marine Environmental Data) project partners, and the

MI. All project results and outputs were reviewed for quality control internally and by the MI, and then made publicly available through the GeoDI website[1].

A four-month extension allowed the GeoDI project partners to further collaborate with the NETMAR project to advance the specification and development of the semantic framework, which was then fed back into the GeoDI project.

1.4.1. Work Package 1: Initiate & Identify

Objectives:

- Review existing geoscientific datasets and current applications (use of data) within the MI and its partners
- Identify user needs and prioritised list of datasets, current and future applications and outputs
- Review and understand the process used to collect data and the datasets semantics
- Identify any problems regarding data quality, resolution or metadata
- Analyse datasets and identify technical issues regarding modelling, integration and application.

Work package 1 was structured and carried out according to the following tasks:

- T1.1. Review data and applications
- T1.2. Identify user needs and prioritised data
- T1.3. Identify metadata and quality issues
- T1.4. Data familiarisation and review
- T1.5. Identify technical issues.

CMRC engaged with the MI on reviewing the various types of geoscientific datasets available from the INFOMAR, INSS, and MESH programmes, and on identifying the technical problems related to these datasets with regard to their management and integration. CMRC also worked closely with the MI on identifying geoscientific data applications and user requirements. For these purposes, several methods were used:

- Face-to-face progress meetings and workshops, notably the INFOMAR Seminar (MI, 13-02-2008) and the GeoDI data familiarisation workshop (MI, 22-07-2008). The former was an opportunity to gain an overall understanding of the INFOMAR project scope and data collection techniques and applications, while the latter was more focused on GeoDI, and

[1] http://geodi.ucc.ie

covered data collection and products, data and metadata management, and user requirements. The GeoDI data familiarisation workshop and progress meetings were an opportunity to meet with geoscientific data users from the MI and to understand their applications and requirements.

- Questionnaires were sent to geoscientific data users identified by the MI. The questionnaires allowed CMRC to collect more information on user requirements.
- Investigation of representative sample geoscientific datasets made available by the MI, which allowed CMRC to have a deeper understanding of how data were represented and stored, and to identify the technical issues related to these data.
- Internal CMRC meetings with marine geology experts, which allowed CMRC and the Cork Constraint Computing Centre (4C) IT researchers to get a good understanding of marine geoscientific datasets, data collection, and analysis techniques.

WP1 resulted in:

- Deliverable D1.1*'Review of Geoscientific Datasets'* - a report containing an inventory of the available geoscientific data, including their structures and formats, identifies data and technical issues and user requirements.

1.4.2. Work Package 2: Review

Objectives:

- Review standard and existing use case ontologies and controlled vocabularies, and identify mappings with the prioritised datasets
- Identify and examine existing standards and data model use cases for marine and/or geophysical data
- Review ontology languages, tools, and servers
- Review international best practice for the management of large geoscientific data
- Review state of the art ontology matching and merging techniques.

Work package 2 was structured and carried out according to the following tasks:

- T2.1. Review ontologies and controlled vocabularies (CVs)
- T2.2. Identify and examine data standards and models
- T2.3. Review ontology languages, tools and servers
- T2.4. Review international best practice
- T2.5. Review state of the art ontology matching and merging.

As part of T2.1, CMRC worked with Oregon State University (OSU) on identifying and documenting existing ontologies and controlled vocabularies of relevance to geoscientific data and metadata, e.g. the Glossary of Technical Terms of the Geological Survey of Ireland (GSI), the British Geological Survey (BGS) Geoscience Vocabularies, the National Aeronautics and Space Administration (NASA) Global Change Master Directory (GCMD), and the SeaDataNet controlled vocabularies. CMRC paid particular attention to explaining the common uses of ontologies and controlled vocabularies, as these often remain unknown to scientists, and they provided example of projects using ontologies.

As part of T2.2, OSU worked with CMRC on identifying and subsequently assessing existing international and European standards and data models. The review included the most common data, metadata and semantic web standards from the World Wide Web Consortium (W3C), the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC), the European Committee for Standardisation (CEN), and the Infrastructure for Spatial Information in Europe (INSPIRE) directive. The reviewed data models were Arc Marine, GeoSciML, and Arc Geology. Recommendations were made regarding the standards and models to adopt, and how they fit together. This work allowed us to select the Arc Marine data model for integrating the marine geoscientific datasets.

As part of T2.3, CMRC reviewed existing ontology languages, tools, and servers. The work built on results from the previous EU FP6 InterRisk project, and included recommendations regarding the tools and languages to use in GeoDI, subject to evaluation.

As part of T2.4, OSU and CMRC worked together on reviewing international best practice for the management of large volumes of geoscientific data. The review covered important developments in the management of marine geoscientific datasets, and exemplar projects showcasing international best practice. The work also included recommendations aimed at improving several aspects of the integrated geoscientific database (IGDB) such as performance, design, scalability, memory, and backup and metadata management. CMRC liaised with the E-Science Grid Facility for Europe and Latin America (EELA2)[2] project and investigated the use of Grid technology for the management and analysis of geoscientific datasets. An approach for distributed management, discovery, and access of large geoscientific datasets on a Grid infrastructure was proposed.

[2] http://www.eu-eela.eu

As part of T2.5, 4C reviewed and documented existing ontology schema matching techniques, ETL tools, and application programming interfaces (API). 4C delivered an extensive review of the field, including a discussion of what tools might be used in the project, and how they work with the various forms of GeoDI data (including Excel files, MS Access files and Shape files). Discussion focused on whether or not building in-house tools would be a better option - which is what was eventually decided. Other important issues were also discussed prior to critical decisions being made. These included the realisation that multiple matchers were necessary to perform the matching task for the GeoDI data (for example, schema-level, element-level, data-type schema, and relation schema matchers). This work was reviewed and commented on extensively by Dr. Richard Wallace, 4C, particularly during the early stages.

WP2 resulted in:

- Deliverable D2.1 *'Review of Ontologies and Controlled Vocabularies'* - a report documenting existing semantic resources, and explaining how they may be used.
- Deliverable D2.2 *'Identification of Standards and Models'* - a report examining existing standards and data models, with recommendations on how to use them and how they interact.
- The selection of Arc Marine as the basis for developing the integrated geoscientific data model (IGDM).
- The selection of Web Ontology Language (OWL) and the Simple Knowledge Organization System (SKOS) for developing the geoscientific ontologies.
- Based on D2.2, a book chapter was authored entitled *'Coastal Atlas Interoperability'* in Cummins, V., Dwyer, N. and Wright, D.J. (eds.) (2011) *Coastal Informatics – Web Atlas Design and Implementation.* USA, IGI. ISBN 978-1-61520-815-9 (hardcover) – ISBN 978-1-61520-816-6 (ebook).
- A presentation was delivered entitled *'G^{2}Library, a Grid Geoscientific Library'* at the second EELA2 conference in Choroni, Venezuela, November 2009. The presentation proposed a Grid library for managing, discovering, and accessing large marine geoscientific datasets on the Grid.
- Deliverable D2.3 *'Review of Ontology Languages and Tools'* – a report including recommendations regarding the ontology tools and languages to use in the project.
- Deliverable D2.4 *'Review of International Best Practice for the Management of Large Volumes of Geophysical/Geological Data'.*
- Deliverable D2.5 *'Review of ETL and ontology/Schema Matching Techniques and Tools'.*

1.4.3. Work Package 3: Specify Models

Objectives:

- Specify and develop integrated models for geoscientific data
- Develop ontologies for geoscientific data and resources and for metadata keywords
- Validate the developed ontologies and models
- Implement and test prototype
- Develop and test automatic or semi-automatic ontology merging and matching techniques.

Work package 3 was structured and carried out according to the following tasks:

- T3.1. Integrated data models
- T3.2. Specify ontologies
- T3.3. Validate ontologies and models
- T3.4. Prototype implementation and benchmarking
- T3.5. Ontology merging and matching techniques.

As part of T3.1, CMRC proposed a few initial design options for the representation of the various marine geoscientific data types based on previous projects, notably the Irish Marine Aggregate Initiative (IMAGIN), the BIDI project and the MDR. The proposed design options were based on the Arc Marine data model. They were discussed with OSU, USGS, and the MI in order to identify the most suitable design features. The selection of the design options was based on user requirements identified in WP1, and on advice from Dawn Wright (one of the creators of Arc Marine) and Brian Andrews (USGS). The selection of the design options enabled the completion of an integrated geoscientific data model, the building blocks of which are illustrated in Figure 2.

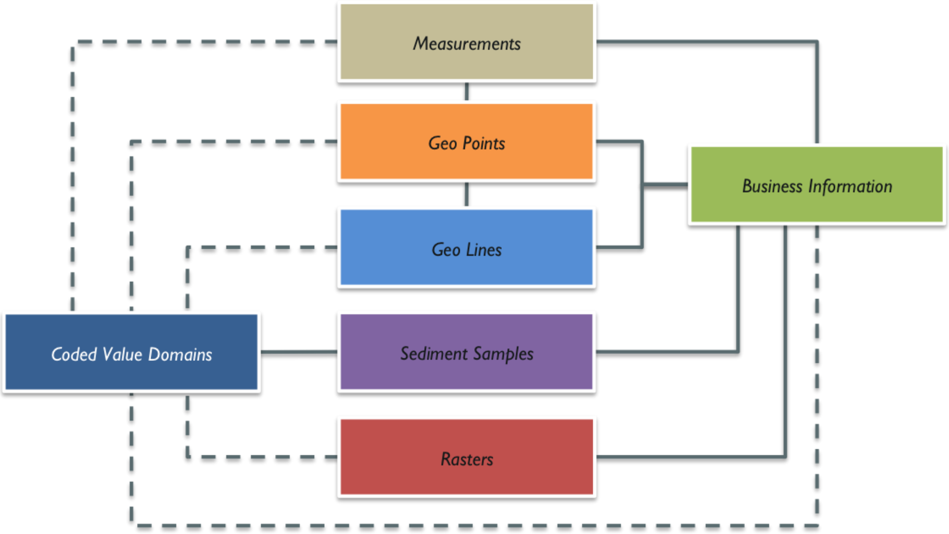


Figure : Overview of the Integrated Geoscientific Data Model Packages

As part of T3.2, CMRC and 4C specified three practical purposes for developing ontologies and controlled vocabularies in GeoDI (c.f., Figure 3). Based on these purposes, CMRC and 4C specified three ontologies and controlled vocabularies: domain ontology; data model ontology; and data controlled vocabularies. The specification was based on existing semantic resources recommended as part of WP1, namely the SeaDataNet and NERC vocabularies, the GCMD science keywords, the NASA GCMD instruments and sensors, and the USGS thesaurus. Others resources such as the INSPIRE themes, the General Multilingual Environmental Thesaurus (GEMET), and the BGS Discovery Metadata Keywords were identified during the later stages of the project.

Both the data model and ontology designs were the result of a collaboration effort between IT and marine geology specialists. As part of T3.3, OSU and USGS reviewed the data model and ontologies specified and developed by CMRC and 4C. Feedback from the review was subsequently used to improve the data model and ontology as part of an iterative process. It should be noted that the GeoDI domain ontology has been extended by the Griffith Geomatics for Geosciences project team to include terrestrial topics.

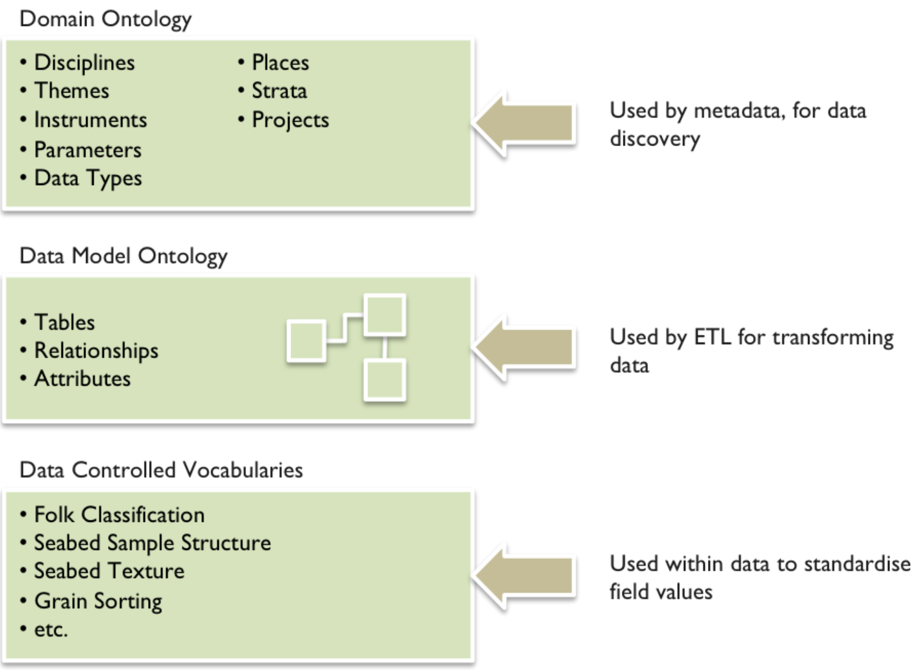


Figure : Geoscientific Ontologies and Controlled Vocabularies and their purposes

As part of T3.4, CMRC implemented and tested the integrated data model in Microsoft Sequel Server (MSQL) 2008, the database management system used by the MI. This would allow compatibility between the generated code (SQL scripts) and the MI software infrastructure. CMRC and 4C then tested loading various types of sample data provided by the MI, for example seabed samples, grain size analysis, single beam, and multibeam data. This allowed 4C to identify problems encountered during the data loading process, which were then taken into consideration in the design of the ETL tool.

As part of T3.5, and based on the recommendation of T2.5 and the results of T3.4, 4C designed and implemented an intelligent ontology matching tool based on a multi-strategic approach, in which a set of ontology matchers are used to discover mappings between two data structures (represented as ontologies), each using its own technique. Then, the results of all the matchers are combined. The proposed approach was based on existing ontology and schema matching techniques from the Database and AI communities. The approach was tested with a sample dataset provided by the MI, and the results were encouraging, with the tool having an accuracy rate of 75 percent.

WP3 resulted in:

- An IGDM available in both the relational and object-relational models, publicly available as MS Visio diagrams from the GeoDI web site (http://geodi.ucc.ie).
- An SQL script for installing the IGDM in a MSQL 2008 database.

- The following set of ontologies and controlled vocabularies:
  - Geoscientific domain ontology available as an OWL file. The latest version can be accessed at http://geodi.ucc.ie/ont/20110429/geoscience.owl.
  - A data model ontology, available as an OWL file, and delivered as part of the ETL tool.
  - A set of data controlled vocabularies, available as part of the data model and the SQL script.
- Deliverable D3.2 *'Selected Data Model and Ontologies for Geoscientific Data'* - a report that documents the data model and ontologies listed above.
- *'Ontology based Automatic ETL for Marine Geoscientific Data'* - an article published in the proceedings of the Remote Sensing and Phtogrammetry Society Annual Conference (RSPSoc'2010), 1st – 3rd September 2010, Cork, Ireland.
- A tested intelligent ontology matching tool for detecting mappings between heterogeneous geoscientific data structures.

1.4.4. Work Package 4: Specify System

Objectives:

- Specify suitable data delivery methods for dissemination and integrated analyses of the data
- Identify process changes which would improve the management of the data
- Specify the complete system.

Work package 4 was structured and carried out according to the following tasks:

- T4.1. Data delivery methods
- T4.2. Process changes for the improvement of data management
- T4.3. System specification.

As part of T4.1, and based on the recommendations made in T2.2, T2.3, and T2.4, CMRC specified the data, metadata, and ontology delivery methods for dissemination and integrated analyses. The proposed delivery protocols and formats were based on the INSPIRE directive and the W3C, ISO, and OGC standards.

As part of T4.2, and based on the data management problems identified in WP1, international best practice, and on experience gained as part of previous and ongoing parallel projects, CMRC specified process changes which would improve the management of geoscientific data, metadata, ontologies, and controlled vocabularies. The specification addressed both high-level problems, such as conformance with standards, and detailed and practical problems such as primary keys, specific data requirements, ontology population, etc.

As part of T4.3, CMRC specified the Integrated Geoscientific Information System (IGIS) as the solution to manage, integrate, and access the various geoscientific resources dealt with by the GeoDI project, i.e. data, metadata, and semantic knowledge (c.f., Figure 4). The proposed architecture is service oriented, which means that it packages functionality as a suite of interconnected interoperable web services that may be reused as building blocks for developing new systems. In addition, it is based on INSPIRE-compliant standard communication protocols, web services, and data format that facilitate interoperability with external systems.

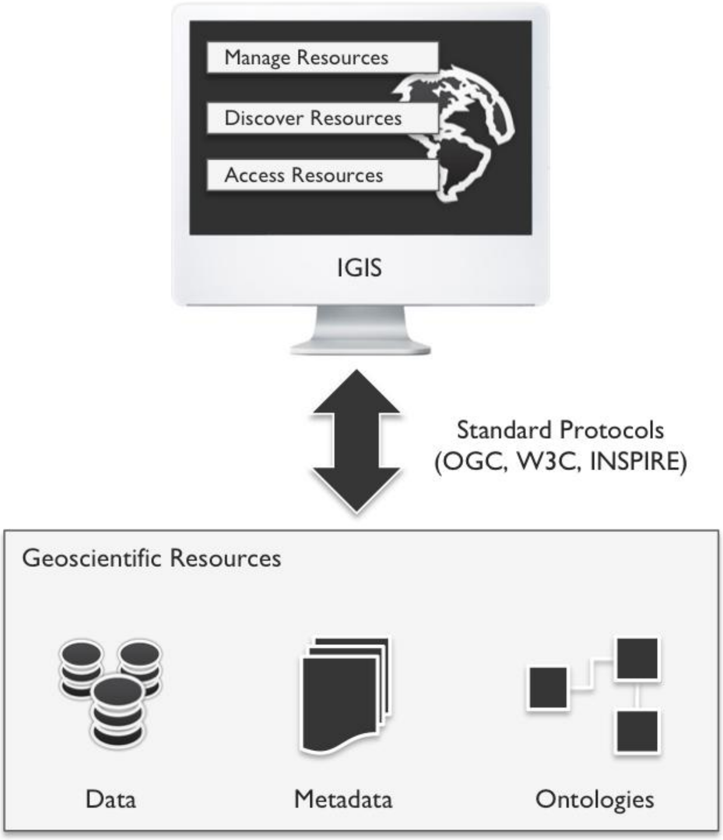


Figure : The objective of IGIS is to provide integrated management of, discovery of, and access to geoscientific resources, all based on standard protocols, services, and formats

In addition to standard web services, the IGIS specification supports a new service which is the Semantic Web Service (SWS). The SWS is a high-level SKOS web based service that handles most common semantic queries required by external applications and clients. The SWS specification is a joint effort between the GeoDI team and the ongoing NETMAR project. It will be taken over by NETMAR, advanced and submitted to a relevant standardisation body with the aim of making it a standard for querying SKOS thesauri. This specification was reviewed and approved by two semantic web experts: Prof. Peter Fox, from the Rensselaer Polytechnic Institute (RPI); and Pr. Oscar Corcho, from the Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid (UPM).

The IGIS specification also includes a web portal, GeoDataOnline, for accessing the various IGIS services and resources. The portal specification includes an ontology browser, a data discovery interface, a metadata viewer, and web mapping and data access interfaces.

The IGIS specification allows for integration with the services developed as part of the Griffith Geomatics for Geosciences project.

WP4 resulted in:

- Deliverable D4.1 *'Data Delivery Methods'* – a report which is the result of T4.1.
- Deliverable D4.2 *'Process Changes for the Improvement of Data Management'* – a report which is the result of T4.2.
- Deliverable D4.3 *'System Specification'* - a specification report that defines the IGIS architecture and components, and specifies the interfaces between the various components and how these fit together. The report reaches a level of technical detail that makes it easy for developers to implement the IGIS.

1.4.5. Work Package 5: Implement & Evaluate

Objectives:

- Implement an ontology server for delivering ontologies
- Implement intelligent ETL tools based on the automatic ontology merging and matching techniques
- Extend the prototype using the ETL tools and perform further testing and evaluation.

Work package 5 was structured and carried out according to the following tasks:

- T5.1. Implement ontology server
- T5.2. Implement intelligent ETL tools
- T5.3. Prototype extension and integration.

As part of T5.1, CMRC implemented a triple store using the PostgreSQL database management system and the Jena ontology framework. The triple store is hosted on the CMRC server and it contains the geoscientific domain ontology. Loading ontologies into the triple store was facilitated through the ontology-loading tool developed by CMRC. On top of the triple store, CMRC developed a SWS conforming to the specification identified under D4.3. The SWS is currently running on the CMRC server and available at the following URL: http://gdo.ucc.ie:8080/gdo/SWS. Examples of how to interact with this service are provided as part of output in D4.3.

As part of T5.2, 4C developed a semi-automatic ETL tool for the IGDB based on the GeoDI Extractor, the GeoDI Transformer, and the GeoDI Loader. The Transformer includes the matching tools developed as part T3.5. These tools were all written in Java. In addition a user-friendly interface was built, also in Java. The interface includes options for the user to provide input to the System at critical junctures (e.g. if the ETL cannot suggest a mapping from data to schemas), to verify results before the System continues and possibly to override results provided by the GeoDI System. These are important features for making the System more practical and easy to use. The GeoDI ETL tool supports ESRI Shapefile, MS Excel, MS Access, and CSV (comma-separated-values) files. It is also capable of automatically converting common data formats such as date formats and geographic coordinates.

As part of T5.3, CMRC implemented the proposed IGIS according to its specification as defined in D4.3. Due to time constraints, some components had to be prioritised on their importance to the overall architecture. Therefore some, such as the ontology server and the mapping interface, were not implemented. However, the majority of components identified in D4.3 were fully implemented.

The project partners and the MI tested the implemented system. As there was no plan in the description of work for an external user evaluation, CMRC evaluated the IGIS against the IGIS specification, reported any issues, and provided recommended solutions for them.

WP5 resulted in:

- An ontology loader for loading OWL or RDF ontology files into a triple store based on PostgreSQL or MSQL 2008. The tool is developed in Java using the Jena API. It is a standalone executable jar file to be run from command-line.
- A semi-automatic ETL tool for loading Shapefile, MS Excel, MS Access, and CSV data files into the IGDB, and generating basic metadata associated with them such as keywords, bounding box, and time stamp. In addition to this tool we provided a script for loading large data, such as multibeam, that are not supported by the ETL tool.

- The IGIS, a complete system for integrating, managing, and accessing geoscientific resources. In addition to standard services, the IGIS includes the following components which may be reused individually (e.g. by Irish Spatial Data Exchange (ISDE)):
    - A SWS for accessing the geoscientific domain ontology. This service was developed in Java using the Jena API and is publicly available at http://gdo.ucc.ie:8080/gdo/SWS.
    - A catalogue service mediator that allows access to distributed catalogue services and solves semantic conflicts between these. This service was developed in Java and is accessible at http://gdo.ucc.ie:8080/gdo/CS;
    - The GeoDataOnline portal, which was developed in Adobe Flex 3, available at http://gdo.ucc.ie.

1.4.6. Work Package 6: Synthesise

Objectives:

- Assess potential analyses and services that can be made available internally or externally
- Identify tools and approaches to facilitate geospatial analysis and querying of geoscientific data.

Work package 6 was structured and carried out according to the following tasks:

- T6.1. Assess potential analyses and services
- T6.2. Tools and approaches for geospatial analysis and querying.

As part of this work package, OSU and CMRC identified and documented potential analyses and services that can be made available to the geoscientific database, and also tools and approaches to facilitate geospatial analysis and querying of geoscientific data.

WP6 resulted in:

- Deliverable D6.1 *'Potential Analyses, Services, and Tools for Geospatial Analysis and Querying'*– a report which is the result of T6.1.

1.5. Difficulties

A few technical and non-technical difficulties were encountered during the GeoDI project, which are summarised below:

- On 11 March 2008 a fire in the CMRC building completely destroyed their offices. Fortunately, no data related to the project were lost due to the backup strategy of the CMRC GeoDI team. However researchers were displaced and equipment lost, and this resulted in a delay of approximately 3 months. In order to minimise the impact of this delay, some tasks from WP2 (Review) were brought forward, notably T2.1 (Review of ontologies and controlled vocabularies) and T2.2 (Identification of data standards and models).
- The IGDM was initially developed as an object-relational data model following the Arc Marine model, and implemented in the Open Source PostgreSQL database management system which supports the object-relational model, with which CMRC were familiar. However at a later stage, CMRC was required to implement the database on MSQL 2008, as this is the database management system used by the MI. Therefore, the data model needed to be re-implemented as a flat relational model as MSQL 2008 does not support the object-relational model. A few technical issues were encountered (mostly related to converting an object-relational model to a relational one), but were solved through more interactions with the MI and the Joined-up Data for Joined-up Thinking project. Ali Al Othman, CMRC researcher who is an expert in MS technologies, was seconded to the project in order to minimise delays caused by the reimplementation of the data model and the database.
- One of the main difficulties encountered by 4C in building the ETL tool involved the format of the information on which the system is based. This data was not collected with any specific automatic processing in mind, and was therefore not tailored to that end. Thus, there tended to be a large degree of variability in the input, which had to be addressed as well as the anticipated problems with errors. Another difficulty was in building the data model ontology itself. This had to be done with an eye to an existing data model, which tended to influence the person building the ontology in a certain direction both because of the way that terms are classified (i.e. how things are divided up) and what has been omitted. For the latter this involved both discovering the missing terms and deciding where they should go in the overall scheme.
- Another difficulty related to the ETL tool development was the time initially allocated to this task (one year including data familiarisation, research, and development), which was

not sufficient to develop a fully operational tool that supports all types of data available. Given the complexity of the problems related to loading geoscientific data, the ETL development could have been a project in its own right. To optimise the quality of the results, CMRC contributed around two additional person-months to the development of the ETL.

- Work package 6 was initially scoped in the description of work reliant on the outcomes of the data mining project (defined strategic PBA/KI/07/02), which was cancelled. Due to the lack of inputs, CMRC had to investigate additional literature in order to identify data mining applications for the integrated geoscientific database to satisfy components of this work package.

1.6. Intellectual Property

All the outputs of the GeoDI project are publicly available at no cost. The specification and best practice reports generated by the project partners are publicly available online through the GeoDI website. These are research outputs of interest to the marine community that may be exploited for both commercial and non-commercial purposes. The software developed by UCC such as the SWS, the catalogue service mediator, and the ETL tool will be made freely available for non-commercial use through a licensing mechanism. CMRC already uses a similar mechanism for licensing the Marine Irish Digital Atlas (MIDA) software which will be adopted for the GeoDI software.

2. Results and Outcomes

Work carried out in the GeoDI project resulted in the following (as detailed in Section 1.4):

- A set of reports reviewing existing technologies, standards, models, and best practices related to the integration, management, and delivery of geoscientific datasets, with recommendations and proposed process changes aimed at improving geoscientific data management.
- An IGDM based on Arc Marine and compatible with existing MI databases, namely the MDR and the BIDI model. The model is both flexible and scalable.
- A set of geoscientific ontologies for use in data metadata and ETL, built based on existing well-established vocabularies.
- A complete system specification for integrating, managing, and delivering geoscientific resources (data, metadata, and ontologies), including detailed specification of the subsystems and how they fit together as a whole and interact with each other. The system specification is based on international and European standards (W3C, ISO, OGC, and INSPIRE). The specification demonstrates practical uses of ontologies and how these can be linked to data and metadata. The specification also includes a SWS, which is being further advanced by the EU FP7 NETMAR project in order to be submitted to a standardised body.
- A semi-automatic, generic, and ontology-based ETL tool for loading datasets into the integrated geoscientific database, which may be further developed and customised to other systems or databases.
- The IGIS, a complete system for integrating, managing, and accessing geoscientific resources. In addition to standard services, the IGIS includes the following components which may be reused individually (e.g. by ISDE):
    - A SWS for accessing the geoscientific domain ontology. This service was developed in Java using the Jena API and is publicly available at http://gdo.ucc.ie:8080/gdo/SWS.
    - A catalogue service mediator that allows access to distributed catalogue services and solves semantic conflicts between these. This service was developed in Java and is accessible at http://gdo.ucc.ie:8080/gdo/CS.
    - The GeoDataOnline portal, which was developed in Adobe Flex 3, and is available at http://gdo.ucc.ie.
- A report outlining recommendations assessing potential analyses and services, and identifying tools and approaches to facilitate geospatial analysis of geoscientific data.

- Three publications:
  - An article *'Ontology based Automatic ETL for Marine Geoscientific Data'* published in the proceedings of the Remote Sensing and Phtogrammetry Society Annual Conference (RSPSoc'2010), 1st – 3rd September 2010, Cork, Ireland.
  - A presentation *'G^{2}Library, a Grid Geoscientific Library'* at the second EELA2 conference in Choroni, Venezuela, November 2009. The presentation proposed a Grid library for managing, discovering, and accessing large marine geoscientific datasets on the Grid.
  - A book chapter entitled *'Coastal Atlas Interoperability'* in Cummins, V., Dwyer, N. and Wright, D.J. (eds.) (2011) *Coastal Informatics – Web Atlas Design and Implementation.* USA, IGI. ISBN 978-1-61520-815-9 (hardcover) – ISBN 978-1-61520-816-6 (ebook).

3. Impacts and Benefits

The GeoDI project allowed the enhancement of expertise, and internal and external building of capacity in the fields of geoscientific data integration. It also provided opportunities for the innovative application of ontologies, AI, service-oriented architectures, and Grid computing to problem solving in geoscientific data integration, management, and access. The project enabled the transfer of expertise between Ireland and the USA, and between the individual partner institutes and individuals. GeoDI was a successful cross-disciplinary approach to geoscientific data management (geosciences, database management, and AI).

The GeoDI project was an opportunity for all the partners to collaborate with other related projects, notably the Griffith's Geomatics for Geosciences, the Joined up Data for Joined up Thinking, and the EU FP7 NETMAR project.

The GeoDI project allowed the specification and development of a semantic framework that has since been adopted by the NETMAR project, and will be further advanced and submitted to a standardisation body in order to provide a standard for accessing semantic resources.

This project gave 4C some exposure in new areas of endeavour, specifically in marine science and, more specifically, information technology for marine science. There is potential for increasing the scope of 4C's work, particularly in the domain of combining knowledge representation with combinatorial optimisation.

According to Brian Andrews (USGS), "*the results of the GeoDI project described in the deliverables are an extremely valuable resource for any international organisation with a large marine geophysical mapping program; and an equally large data management and dissemination program. Throughout the GeoDI project I have shared results with colleagues here in the USGS Coastal and Marine Geology Program (CMGP)*[3] *and also in our Knowledge Management Working Group*[4]*. The overall approach of the GeoDI project was thorough and included input from multidisciplinary project partners. I have encouraged the USGS CMGP to use the GeoDI project as a template and initiate a similar project. While I have no definite plans for follow-up research, I will continue to use the deliverables as a reference and also promote the GeoDI project as an excellent example of a 'holistic' project that includes all the important steps for the successful management and delivery of marine geophysical data to an international user community.*"

[3] http://marine.usgs.gov/index.php

[4] http://woodshole.er.usgs.gov/projects/project_get.php?proj=2921BO8&style=html

Brian Andrews specifically highlighted D2.1, '*Review of Ontologies and Controlled Vocabularies'* and D2.2, '*Identification of Standards and Models'* as being of particular use to the USGS. He concluded "*I have gained a new understanding of the organisation and methodological approach for a project of this scope and size. The USGS works with the same data types and volumes (size) but does not have a similar approach because our data are spread all over US and international waters. Nevertheless we could adopt a similar data model as a mechanism to centralise and manage our similar datasets. In addition I learned a tremendous amount about the importance of controlled vocabularies, granular- or tiered- metadata, and ontologies.*"

Acronyms

4C	Cork Constraint Computing Centre
AI	Artificial Intelligence
API	Application Programming Interface
BGS	British Geological Society
BIDI	Biological Data Integration
CMRC	Coastal and Marine Research Centre
CSV	Comma Separated Values
EELA2	E-Science Grid Facility for Europe and Latin America project
ETL	Extract, Transform and Load
FP6	European Union Framework Programme 6
FP7	European Union Framework Programme 7
GCMD	Global Change Master Directory
GeoDI	Geoscientific Data Integration Project
IGDB	Integrated Geoscientific Database
IGDM	Integrated Geoscientific Data Model
IGIS	Integrated Geoscientific Information System
INFOMAR	Integrated Mapping for the Sustainable Development of Ireland's Marine Resource
INSPIRE	Infrastructure for Spatial Information in Europe
INSS	Irish National Seabed Survey
ISDE	Irish Spatial Data Exchange
ISO	International Organization for Standardization
MDR	Marine Data Repository
MESH	Mapping European Seabed Habitats
MI	Marine Institute
MSQL	MicroSoft Sequel Server
NASA	National Aeronautics and Space Administration
NETMAR	Open Service Network for Marine Environmental Data Project
OGC	Open Geospatial Consortium
OSU	Oregon State University
OWL	Web Ontology Language
SKOS	Simple Knowledge Organization System

SWS	Semantic Web Service
USGS	United States Geological Survey
W3C	World Wide Web Consortium

www.marine.ie

Headquarters

Marine Institute
Rinville
Oranmore
Co. Galway
Tel: +353 91 730 400
Fax: +353 91 730 470
Email: institute.mail@marine.ie

Marine Institute Regional Offices & Labratories

Marine Institute
80 Harcourt Street
Dublin 2
Tel: +353 1 476 6500
Fax: +353 1 478 4988

Marine Institute
Furnace
Newport
Co. Mayo
Tel: +353 98 42300
Fax: +353 98 42340