

ROCKALL STUDIES GROUP

TECHNICAL REVIEW OF RSG DATABASE REQUIREMENTS

Jonathan Guard
(CSA Computing Services Ltd.)

Charlotte O'Kelly
(Informatic Management International)

John Wallace
(Informatic Management International)



Signed on behalf of CSA

*Viv Byrne
Managing Director
CSA Group*



Signed on behalf of Informatic Management

*John Wallace
Managing Director
Informatic Management*

EXECUTIVE SUMMARY

There are two key parties involved in the RSG project; RSG End User's (sponsor oil companies, government departments) and the RSG Project Teams. This report examines options and provides recommendations for how the data deliverables from the Project Teams can be managed and delivered to the End User's in a manner where the data can be fully utilised. The concept is simple, the Project Teams are collecting, assimilating and interpreting various different scientific data types which have been categorised under the broad categories of; Met-Ocean, Sub Surface, Environmental and Sea Bed. There is a high degree of interdependency between the various projects. The projects can also be classified on the basis of whether they are new data acquisition projects, data interpretation projects or compilation / methodology / research projects. The projects will result in a diverse range of products and data types from PhD thesis's to algorithms to mosaiced imagery.

The RSG End Users have their own operational data requirements. What formats can they accept, metadata requirements, industry standard requirements etc. To make matters even more complicated the technology of data transfer and data publishing is going through a period of rapid change. Formats, media standards and communication methods are changing at an ever increasing rate. Choice of data management solution will never satisfy all of the members and any solution will be a compromise, enabling majority satisfaction with the given choice. In order to provide options, the following key issues are addressed;

- 1) What are the Project Teams intending to deliver and in what formats (media, software format) and to what industry standards.
- 2) What are the requirements of the End Users with regard to acceptable formats (media, software format) and to what industry standards.

Case studies and other industry examples of databasing diverse geoscience data sets are important reference material and a review of some examples are given. The petroleum exploration industry is in the process of defining standards for; data models, data collection, sharing, archiving and transfer between different software packages. No clear global E&P standards have yet to emerge that cross the spectrum of scientific disciplines covered by the four project categories.

In the long run, data and the deliverables from the projects are the only tangible result of the project work. Some of the projects are of more relevance to the oil company members than others. However the RSG project work also has national importance and how the data can be disseminated to other non-RSG members (commercial, and non-commercial organisations) is an important area that has to be addressed by the management committee. Some of the projects for example contain data that has been licensed to some of the RSG members but not to all. How should this data be used and disseminated ?

Many of the Project Teams may not require or indeed wish for publication guidance. However, it is strongly recommended that a Data Handbook is produced to cover key areas such as; Metadata required, quality standards, transfer media, data formats acceptable etc.

Options for dissemination of the data can be broadly categorised under two headings: Direct Dissemination and Third Party Dissemination. The first category is in essence the simplest and involves the Project Teams delivering their project deliverables directly to the RSG Secretariat for onward distribution to the RSG members. Third Party Dissemination involves putting in place a new project team to develop and implement a data management solution. Information is provided on what areas need to be addressed before any solution is decided upon.

There are three pieces of software recommended to provide a data management solution for the RSG. The total system solution is provisionally referred to as IRSGS (Integrated RSG System). The three components include an RSG Data Inventory (RDI), the Generic RSG Information System (GRSGIS) and the RSG Web page. Each of the three software components can be implemented independently but, resources permitting, together will provide a strong technical foundation on which to build a complete integrated system for all RSG type projects.

1	INTRODUCTION.....	5
1.1	THE ROCKALL STUDIES GROUP	5
1.2	PROJECT METHODOLOGY	5
1.3	SCOPE AND LIMITATIONS OF THE STUDY	9
1.4	CURRENT INDUSTRY CONTEXT	9
1.5	REPORT DESCRIPTION	9
2	THE RSG PROJECTS	11
2.1	INTRODUCTION	11
2.2	OVERVIEW OF RSG PROJECTS	12
2.3	INFORMATION GATHERING METHODOLOGY	14
2.4	RESPONSE TO THE QUESTIONNAIRE SURVEY	15
2.5	SCIENTIFIC DIVERSITY	16
2.6	COLLECTION STANDARDS AND QUALITY ASSURANCE	17
2.7	DATA FORMATS.....	17
2.8	DELIVERABLE FORMATS	17
3	THE DATA DICTIONARY	20
3.1	INTRODUCTION	20
3.2	DEVELOPING A DICTIONARY	20
3.3	AN EXISTING DICTIONARY MODEL	20
3.4	THE RSG DICTIONARY	21
4	DATA USER REQUIREMENTS.....	24
4.1	INTRODUCTION	24
4.2	DATA PUBLICATION	24
4.3	AUDIENCE	24
4.4	CLASSES OF DATA	24
4.5	DATA DISTRIBUTION CONSTRAINTS.....	25
4.6	LEVEL OF FUNCTIONALITY	25
4.7	PERFORMANCE ISSUES	26
4.8	UPDATE-ABILITY	26
4.9	RSG MEMBERS FEEDBACK	26
4.9.1	<i>Data Types</i>	27
4.9.2	<i>Data Formats and Software used</i>	27
4.9.3	<i>Metadata</i>	27
5	DATA MANAGEMENT CASE STUDIES	28
5.1	INTRODUCTION	28
5.2	DATA MODELS	28
5.3	OPERATIONAL STANDARDS	31
5.4	DATA FORMAT STANDARDS	31
5.5	ARCHIVING MEDIA STANDARDS	31
5.6	COMMUNICATIONS AND INTERNET ACCESSIBLE DATABASES	32
5.7	RELEVANCE OF THE CASE STUDIES TO RSG DATA MANAGEMENT REQUIREMENTS	34
6	OPTIONS.....	35
6.1	INTRODUCTION	35
6.2	DIRECT DISSEMINATION – “THE CARDBOARD BOX OPTION”	35
6.3	METADATA INDEX – “ DATA INVENTORY”	35
6.4	DATABASE – “DATA DISTRIBUTION”	35
7	PROPOSED SOLUTION	37
7.1	INTRODUCTION	37
7.2	THE RSG DATA HANDBOOK	37
7.3	THE RSG DATA INVENTORY (RDI).....	37

7.3.1	Outline specification of RDI.....	38
7.4	THE GENERIC RSG INFORMATION SYSTEM (GRSGIS)	39
7.4.1	Core content for the GRSGIS.....	40
7.4.2	Outline functional specification for GRSGIS	40
7.4.3	The RSG Web Page.....	41
7.5	IRSGS IMPLEMENTATION	41
7.5.1	Introduction.....	41
7.5.2	Developing and Hosting System.....	41
7.5.3	Schedule.....	41
7.5.4	Funding.....	42
7.5.5	The IRSGS Team and their functions.....	43
8	SUMMARY	43
9	APPENDIX 1 – QUESTIONNAIRES	45
10	APPENDIX 2 – BRITISH OCEANOGRAPHIC DATA CENTER (BODC) DATA DICTIONARY EXAMPLE.....	46
11	APPENDIX 3 – METADATA FORM EXAMPLES	47
	Figure 1 . Flow chart summarising the overall PIP-RSG project flow.....	6
	Figure 2 Interdependency Diagram for the RSG Projects	14
	Figure 3 IRSGS Operational Flowchart.....	39

"This Project, including data and survey results acquired for the purpose, has been undertaken on behalf of the Rockall Studies Group (RSG) of the Irish Petroleum Infrastructure Programme Group 2 which was established by the Petroleum Affairs Division of the Department of the Marine and Natural Resources on 4 June, 1997 in conjunction with the award of exploration licences under the Rockall Trough Frontier Licensing Round. The RSG comprises: Agip (UK) Ltd, Anadarko Ireland Company, ARCO Ireland Offshore Inc, BG Exploration & Production Ltd, BP Exploration Operating Company Ltd, British-Borneo International Ltd, Elf Petroleum Ireland BV, Enterprise Oil plc, Mobil Oil North Sea Ltd, Murphy Ireland Offshore Ltd, Phillips Petroleum Exploration Ireland, Saga Petroleum Ireland Ltd, Shell EP Ireland B.V., Statoil Exploration (Ireland) Ltd, Total Oil Marine plc, Union Texas Petroleum Ltd and the Petroleum Affairs Division of the Department of the Marine and Natural Resources."

1 Introduction

1.1 The Rockall Studies Group

The Rockall Studies Group (RSG) has been created under the umbrella of the Petroleum Infrastructure Programme (PIP), whose overall aim is to promote hydrocarbon exploration activities in Ireland.

The objective of the RSG is to address common industry problems in the Rockall Trough by

- regional data gathering – Geology & Geophysics / geotechnical / environmental / metocean
- research projects, both applied and academic scholarships associated with the research
- research cruise sponsorship
- provision of a forum to facilitate co-operation
- affording the opportunity for Irish involvement.

The RSG will be in operation from 1997 – 2001 and has a total funding of £4.8 million Irish pounds.

1.2 Project Methodology

Project 98/18 was initiated in June 1998 with two project partners; CSA Computing Services Ltd and Informatic Management International. The objectives of the project are 3 fold;

- Produce a metadata dictionary for all data sets acquired by the RSG Projects
- An analysis of user requirements for the management of the data
- Options, recommendations and estimation of costs for the type of database and user interface that is required

The project was estimated to take 4 man months and was scheduled to be completed by the end of September 1998. The contract for the project was jointly signed on the 17th of August 1998.

The project involved 3 stages:

1	PROJECTS	Examines the RSG projects under a number of different areas with the primary objective of looking at each project's deliverables and their suitability for dissemination
2	USER DATA REQUIREMENTS	Examines the requirements of the RSG members who will be the primary audience for the project deliverables.
3	DATA MANAGEMENT OPTIONS	Examination of various data management options for project data and deliverables produced by the RSG projects.

Table 1 Project 98/18 Project Stages

In order to satisfy objectives 1 & 2, a process of using questionnaires, interviews and examination of contracts and project outlines was used to solicit information.

Sources for information for stage 3 were from projects completed, in-house discussions, journals, the WWW and previous similar case studies.

Figure 1 on the following page is a project flow chart showing the order of the issues addressed.

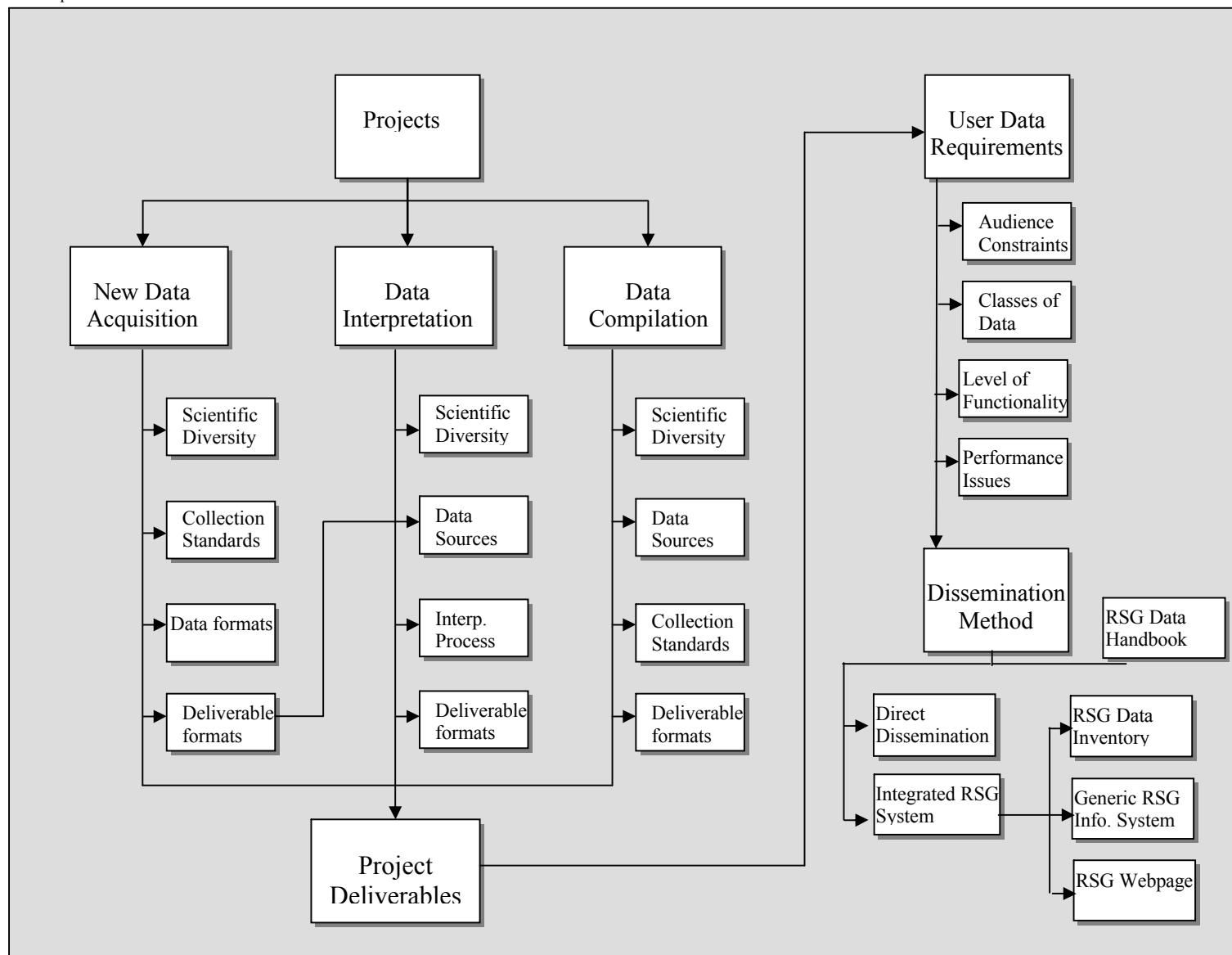


Figure 1 . Flow chart summarising the overall PIP-RSG project flow.

1.3 Scope and limitations of the Study

The technical database review has been restricted by some external factors during the study. In any study, and especially a study with such a variety of projects, both in terms of scientific diversity and project size, information gathering is challenged.

During the first stage of the project, the response to questionnaires by the various projects was initially slow. This can be attributed to the fact, that during the course of this study, RSG projects have still been undergoing contractual negotiation, and where therefore not in a position to comment on resulting data. In the case of the PhD funded projects, the studentships had barely started, and deliverables were still not known in any detail.

In the case of the larger projects, the questionnaires were passed from one person to the other with a general shift in responsibility. Large organisations were not in a position to fully complete the questionnaires in terms of data quality etc, as they were following RSG guidelines, rather than their own procedures.

Another limitation to the study was the lack of common structures between the different projects. The RSG needs to consider setting in place procedures for the future so that when proposals are being accepted they outline in a precise fashion the deliverables. This will enable a structured data management system from the beginning of projects.

It is essential that data management procedures are put in place to cater for the variety of data emanating from the RSG projects, and to ensure the quality and availability of the data to future users.

1.4 Current Industry Context

Data management is a key issue for any scientific project, in particular for a project that has both diverse scientific elements and also a diverse collection of teams and experts from many different disciplines and backgrounds. Currently there is much work going on in the industry to create synergies between different scientific disciplines both in terms of sharing and using data. However, it is not a simple matter, with different disciplines have their own software, hardware, quality standards and data formats.

Case studies such as the Irish Marine Data Centre's work on the EDAP project provide valuable insights into publication of data from a multidisciplinary research project. Likewise non profit organisations such as; POSC, PPDM and the OPEN SPIRIT Initiative are developing new data models and standards for the oil industry, allowing organisations, scientific disciplines and software to share data easier and more transparently.

1.5 Report Description

The report is divided into 7 chapters as outlined below.

1	INTRODUCTION	Introduction to the project, RSG
2	RSG PROJECTS	The RSG projects; their scientific diversity, standards to be used, deliverables and their formats and deliverable media to be used.
3	DATA DICTIONARY	Framework structure for a data dictionary for the RSG projects.
4	DATA USER REQUIREMENTS	Issues for the publication of the deliverables; who are the audience, are their data constraints, is there a need to distribute every project deliverable and some possible pitfalls.

5	DATA MANAGEMENT CASE STUDIES	A review of some relevant areas that the industry is currently looking at, and their relevance to the RSG project.
6	OPTIONS	Options for a data management solution.
7	PROPOSED SOLUTION	An Integrated RSG System solution for the RSG data management.

Table 2 Report Sections

2 The RSG Projects

2.1 Introduction

In June, at the start of this project, there were 25 projects (including this one) as part of the RSG. Not all of the projects had signed contracts and in particular projects related to Project 97/50 (Figure 1) were unsure of their status as this project was experiencing technical difficulties and delays.

The projects cover a multitude of scientific disciplines which have been categorised by the RSG management under 4 technical committees (Table 3). The Project Teams comprise experts and personnel from University Departments, Government Agencies, Semi-state Agencies, Consultancies, Private Consultants and Commercial Organisations.

Sub-Surface (SSTC)	Met-Ocean (MTC)	Environmental (ETC)	Sea Bed (STC)
97/2 Statistical characteristics of reflectivity patterns in deep seismic profiles	97/29 A Met-ocean strategy for the Rockall Area	97/14 TOBI acquisition	97/8 Fluid Inclusion studies of deep borehole cores
97/3 Structural elements nomenclature		97/14a TOBI – processing & interpretation	97/28 Sedimentological analysis of deep borehole cores
97/11 RAPIDS 3		97/14b TOBI – Quentin Huggett contract	97/34 High resolution biostratigraphy
97/16 Crystalline basement study		97/52 Environmental data gathering	97/50 Atlantic margin drilling
97/21 Seismic imaging study		98/6 Cetacean and sea bird monitoring	97/50a Onboard micropalaeontology
97/40 Gravity & Magnetic studies			97/50b Secure Webserver & Comms link
98/1 Gravity & Magnetic studies and 2D / 3D interpretation			97/51 Sea bed data assessment
			98/11 TTR7 Cruise license
			98/19 Sea bed sampling using ships of opportunity
			98/20 Geotechnical sample analysis
			98/21 Geochemical sample analysis

Table 3 RSG Projects categorised by technical discipline

The projects can also be classified along the following lines: New Data Acquisition, Data Processing and Interpretation, Data Compilation / Meta data and Methodology / Research (Table 4)

New Data Acquisition	Data Processing & Interpretation	Data Compilation / Metadata	Methodology / Research
97/11 RAPIDS 3	97/8 Fluid Inclusion studies of deep borehole cores	97/3 Structural elements nomenclature	97/2 Statistical characteristics of reflectivity patterns in deep seismic profiles
97/14 TOBI acquisition	97/34 High resolution biostratigraphy	97/29 A Met-ocean strategy for the Rockall Area	97/21 Seismic imaging study
97/50 Atlantic margin drilling	97/28 Sedimentological analysis of deep borehole cores	97/51 Sea bed data assessment	97/40 Gravity & Magnetic studies
98/6 Cetacean and sea bird monitoring	97/50a Onboard micropalaeontology	97/52 Environmental data gathering	
98/19 Sea bed sampling using ships of opportunity	98/20 Geotechnical sample analysis	98/11 TTR7 Cruise license	
	98/21 Geochemical sample analysis	98/18 Database Review	
	97/14a TOBI – processing & interpretation	97/16 Crystalline basement study	
		98/1 Gravity & Magnetic studies and 2D / 3D interpretation	

Table 4 RSG Projects categorised by project type

2.2 Overview of RSG Projects

Table 5 below provides a brief description for each of the RSG projects

New Data Acquisition	
97/11 RAPIDS 3	Shooting of seismic lines across the Rockall Trough to provide a 3D constraint on crustal and broad sedimentary geometries.
97/14 TOBI acquisition	TOBI (Towed Off-Bottom Imager) will be used to produce an image mosaic of certain areas of the Rockall Trough sea floor. The imagery provides useful information on the sedimentary, slope stability, geological development and environmental setting.
97/50 Atlantic margin drilling	2 aspects of the project, an initial site survey program (seismic & gravity cores) followed up in 1999 by a deep borehole drilling program. The cores produced by this latter program will be used by project teams from other projects (see Figure 2).
98/6 Cetacean and sea bird monitoring	Cetacean and seabird monitoring program in the Rockall Trough area.
98/19 Sea bed sampling using ships of opportunity	Sea bed sampling using ships of opportunity

Data Processing & Interpretation	
97/8 Fluid Inclusion studies of deep borehole cores	Fluid inclusion studies of samples from the deep borehole cores provided by project 97/50. The project is closely linked to 97/28.
97/34 High resolution biostratigraphy	High resolution biostratigraphy at the margins of the Rockall Trough. This project is closely tied with project 97/28. The cores from 97/50 will be used as source material.
97/28	Sedimentological analysis of the cores provided by project 97/50. The project is

Sedimentological analysis of deep borehole cores	linked to project 97/34.
97/50a Onboard micropalaeontology	Micropalaeontology study carried out on board the drill vessel used for the deep sea coring project 97/50.
98/20 Geotechnical sample analysis	Geotechnical sample analysis of gravity cores required by project 97/50.
98/21 Geochemical sample analysis	Geochemical sample analysis of gravity cores acquired by project 97/50
97/14a TOBI – processing & interpretation	Processing and interpretation of the TOBI imagery acquired during project 97/14, comparison with existing geological & imagery data.

Data Compilation / Metadata

97/3 Structural elements nomenclature	Structural nomenclature of elements in the Rockall Trough (Porcupine) area, standardising names used and locations of features.
97/29 A Met-ocean strategy for the Rockall Area	A Metocean strategy for the Rockall Area, includes data compilation and review.
97/51 Sea bed data assessment	Seabed data gathering and preliminary assessment.
97/52 Environmental data gathering	Environmental data gathering and preliminary assessment
98/11 TTR7 Cruise license	CD-Rom of sedimentological, geochemical, biological and geophysical data collected by a vessel for UNESCO – IOC
97/16 Crystalline basement study	Investigation of reactivated structures in the crystalline basement of the Rockall Trough.
98/1 Gravity & Magnetic studies and 2D / 3D interpretation	Gravity and magnetic compilation and 2D/3D interpretation of the Rockall Trough using existing data sets

Methodology / Research

97/2 Statistical characteristics of reflectivity patterns in deep seismic profiles	Development of algorithms to improve interpretation of reflectivity patterns in deep seismic profiles.
97/21 Seismic imaging study	Character matching techniques for imaging through high impedance heterogeneous layers (basalt)
97/40 Gravity & Magnetic studies	Gravity and Magnetic studies which provide a follow up to work carried out in project 98/1

Table 5 Brief descriptions of each of the RSG projects

There is strong interdependency between projects that is highlighted in Figure 2 on the next page. In reality there are 3 core projects; 97/11, 97/14 and 97/50 from which there is strong interdependency for other projects. This makes project management, particularly important as other project schedules will be put in jeopardy if delays are occurred in other projects.

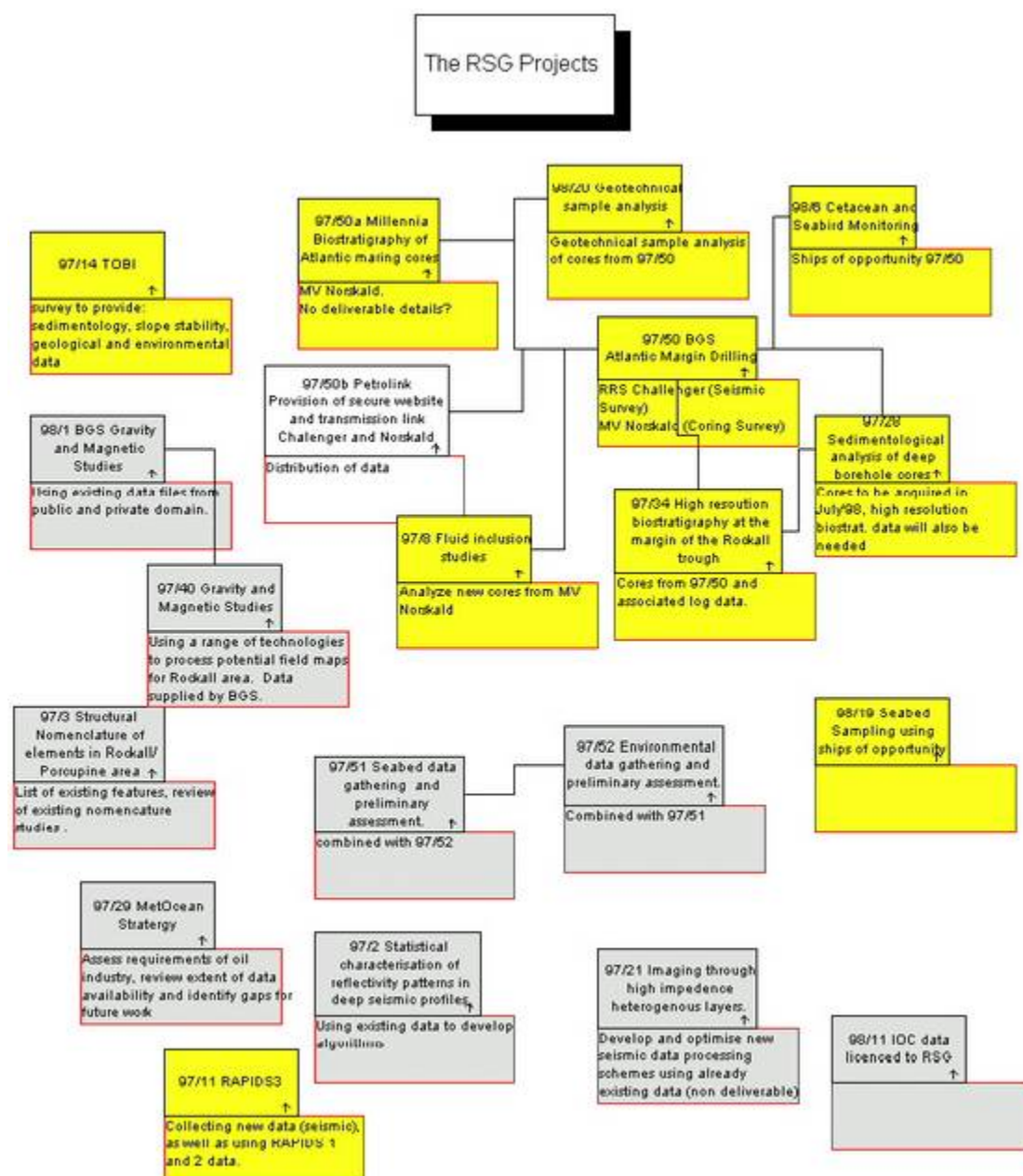


Figure 2 Interdependency Diagram for the RSG Projects

2.3 Information Gathering Methodology

Information about the projects including; deliverables, standards used, quality assurance, confidentiality, geospatial accuracy is necessary to get a broad view of the diversity of the various projects and their deliverables. In order to ascertain this information a set of three questionnaires was designed and distributed to key members of each of the projects. Interviews were also used where possible. Designing a standard questionnaire that would suit the various project classifications shown in Table 4 is impracticable as the questionnaire would be too long and would have too many irrelevant sections, so three questionnaires were designed for; New Data Acquisition, Data Processing & Interpretation and Data Compilation / Methodology (see Appendix 1).

The questionnaires were initially reviewed by the technical committee chairpersons and Bronwyn Cahill (IMDC), Martin Davies (CSA Oil & Gas) and Dave Naylor (ERA) were interviewed both about

their projects and also for their experience and advice regarding the layout and design of the questionnaires. Thanks must go to them for their help and their contributions.

Table 6 below lists the projects, the type of questionnaire sent and to whom and their response.

2.4 Response to the Questionnaire Survey

Project	Questionnaire Sent	Person sent the Questionnaire	Response
97/11 RAPIDS 3	Acquisition	Pat Shannon (UCD) Brian Jacob (DIAS)	No questionnaire submitted
97/14 TOBI acquisition	Acquisition Interpretation	Neil Kenyon (SOC) Brian Jacob (DIAS) Quentin Huggett (GEOTEK)	Interview No questionnaire submitted
97/50 Atlantic margin drilling	Acquisition	Alister Skinner (BGS) Ken Hitchen (BGS) Nigel Fannin (BGS)	Completed questionnaire submitted
98/6 Cetacean and sea bird monitoring	Acquisition	Emer Rogan (UCC) Mark Tasker (JNCC)	No questionnaire submitted
98/19 Sea bed sampling using ships of opportunity	Acquisition	Alex Jones (Phillips)	No questionnaire submitted
97/8 Fluid Inclusion studies of deep borehole cores	Interpretation	John Parnell (QUB) Martin Feely (UCG)	Questionnaire completed jointly
97/34 High resolution biostratigraphy	Interpretation	Ken Higgs (UCC) Jake Jacovides (Millennia)	No questionnaire submitted
97/28 Sedimentological analysis of deep borehole cores	Interpretation	Peter Haughton (UCD)	No questionnaire submitted
97/50a Onboard micropalaeontology	Interpretation	Jake Jacovides (Millennia)	No questionnaire submitted
98/20 Geotechnical sample analysis	Interpretation	Tim Paul (JBA) Barry Lehane (TCD) Mike Long (UCD)	Questionnaire completed jointly
98/21 Geochemical sample analysis	Interpretation	Nigel Goodwin (LGC)	No questionnaire submitted
97/14a TOBI – processing & interpretation	Interpretation	Brian Jacob (DIAS) Peter Readman (DIAS) Keith McGrane (DIAS) Pat Shannon (UCD)	Interview with Quentin Huggett
97/3 Structural elements nomenclature	Compilation	Dave Naylor (ERA)	Interview Completed questionnaire submitted
97/29 A Met-ocean strategy for the Rockall Area	Compilation	Bronwyn Cahill (IMDC)	Interview and written submission
97/51 Sea bed data assessment	Compilation	Martin Davies (CSAOG)	Interview Completed questionnaire submitted
97/52 Environmental data gathering	Compilation	Martin Davies (CSAOG)	Interview Completed questionnaire submitted
98/11 TTR7 Cruise license	Compilation	Neil Kenyon (SOC)	Viewed CD-Rom deliverable

Project	Questionnaire Sent	Person sent the Questionnaire	Response
97/16 Crystalline basement study	Interpretation	Stephen Daly (UCD)	No questionnaire submitted Discussed project in conversation
98/1 Gravity & Magnetic studies and 2D / 3D interpretation	Interpretation	Richard Carruthers (BGS) Phil Houghton (ARK)	Completed questionnaires submitted Discussed project in conversation
97/2 Statistical characteristics of reflectivity patterns in deep seismic profiles	Interpretation	Paul Ryan (UCG)	No questionnaire submitted Discussed project in conversation
97/21 Seismic imaging study	Interpretation	Chris Bean (UCD)	Completed questionnaire submitted
97/40 Gravity & Magnetic studies	Interpretation	Andrew Brock (UCG)	Completed questionnaire submitted

Table 6 Responses to the RSG Project Questionnaires

The gathering of information on the RSG projects through interviews and questionnaires was useful in acquiring information that was not included in the proposals or contracts. It also helped this project team to acquire a better understanding of what the individual projects were trying to achieve, the methods they use and an idea of what they intend to deliver.

The information that was gathered is not enough to make a comprehensive metadata dictionary. Deliverables from many of the projects are unclear and many of the projects have not corresponded that they intend to deliver any digital formats but to deliver entirely paper based deliverables. Clearly paper deliverables are unusable in a database option unless there is time and expense made in reproducing digitally the work. As discussed later (section 3), an RSG data dictionary framework has been developed. For it to be in full working order, it will be necessary for the individual projects to come back with much clearer information as to deliverables.

2.5 Scientific Diversity

Scientific disciplines vary from geophysics to wildlife watching. Table 7 below lists some of the disciplines covered by the RSG projects.

Geophysics	Deep Seismic Profiles TOBI imagery Seismic Interpretation New algorithms for solving seismic reflectance problems
Geological	Core Drilling Palaeontology Sedimentology Sea bed sampling Structural nomenclature Fluid Inclusions Geochemistry Study Geotechnical Study
Biological / Environmental	Cretaceous and Sea Bird Monitoring Met-Ocean Strategy Sea Bed Imagery Environmental Data Compilation Study

Table 7 Scientific Disciplines covered by the RSG projects

2.6 Collection Standards and Quality Assurance

Part of the questionnaire addressed the issue of standards and quality assurance to be used by the Project Teams in carrying out their tasks.

Some of the RSG projects use international standards for the collection of data. For example, Project 97/50 uses UK00A standards for navigation. Many of the projects use their own organisations standards. For example, the BGS have their own internal standards for quality assurance both in terms of collection of data as well as in analysing and processing data. JNCC likewise has well regarded internal standards for the collection of data regarding sea-bird and cetacean monitoring.

With regard to the collection and analysis of scientific data it is often the reputation of the scientists and experts involved which is reflected in the degree of confidence in a particular data set. Often this is the case where there are no international operational standards. An example of such would be the collection and processing of the TOBI imagery (Project 97/14), where the reputation of the team involved and also the quality assurance by Quentin Huggett is the main guiding factor as to the confidence of quality assurance.

Probably the most important aspect of any quality assurance and operational standards is the degree of documentation that is provided with any data set. It cannot be over emphasized the importance of accurate recording of any procedures, corrections or modifications carried out on a data set. An example, might be the decrease in resolution of an image data set in order to reduce file size. This may have repercussions for a user who is unaware that there is a more detailed original data set. This sort of information should be recorded in any associated accompanying metadata.

2.7 Data Formats

Data formats and the translation between software packages is a bug bear of the IT sector. A lot of time and expense can be spent in solving translation problems between different software packages and of particular relevance to this project, the integration into a database.

Seismic imagery for example is commonly transferred using SEG-Y formats. However, there are a number of different SEG-Y formats and it is important that within any associated metadata the exact format type is recorded.

Graphics and reports are often easier to translate into other packages because of the rapid progress that has been made by the IT industry. However, there are formats out there which can be difficult to translate. It should also be noted that not all of the projects mentioned in their questionnaire answers, responded that they were going to supply a digital copy of their thesis or report.

2.8 Deliverable Formats

Deliverables from any of the RSG projects are the only tangible information for future users interested in the projects. What is delivered, whether it is a thesis or an algorithm should be clearly identified and described before any database option is considered. Projects where even the data sources are undecided (primarily the research projects) let alone the deliverable format obviously has an impact on any organisation tendering for a databasing contract.

Table 8 lists in as much detail as possible from the information available, the planned deliverables and formats. The last column highlights the main data set that will become available from each project.

Project	Deliverable Description	Formats	Digital Data sets
97/11 RAPIDS 3	Raw Seismic Data, Processed Seismic, Seismic Models, Cruise Report	Optical disk Q files from Seismic Handler on CD-ROM (SEG Y available on request)	Deep Seismic Image Sections through the Rockall Trough

Project	Deliverable Description	Formats	Digital Data sets
97/14 TOBI acquisition	Final Report Raw navigation, Raw imagery, Mosaiced Image, Cruise Report Cruise Logbook	Navigation Text file, Optical Disk	
97/50 Atlantic margin drilling	Digital Airgun and Sparker data, Seismic tape logs, Analogue seismic data, Raw navigation data and log sheets, Cleaned navigation data, Gravity Core locations Qubit navigation print- out, Echo-sounder print-out, Oceans system Sound Velocity Probe data, Certificate of calibration for SVP, Ship board laboratory log book, Track chart for each site Gravity Core Drill Core	Coda format on Exabyte tapes, Prints and PC disk Hard-copy P1/90 format Excel Spreadsheet Hard-copy Hard-copy Floppy disk Hard-copy Hard-copy Hard-copy	Local Seismic Surveys,
98/6 Cetacean and sea bird monitoring	Seabird and Cetacean Report Cruise Report	Hard-copy Database (?)	
98/19 Sea bed sampling using ships of opportunity	Cruise Report Core Descriptions	Gravity cores Box cores (?) Piston cores (?)	
97/8 Fluid Inclusion studies of deep borehole cores	Fluid inclusion petrography final report	Hard-copy and digital copy	
97/34 High resolution biostratigraphy	Biostratigraphic charts, paleoenvironmental charts, Interpretation report	Large paper charts compiled from StartaBugs software Hard-copy and digital copy	
97/28 Sedimentological analysis of deep borehole cores	Core summary sheets, Core sedimentology logs, Clay mineralogy, Photographic inventory, Well summary log.	Hard-copy Hard-copy and digital copies of the logs (format ?) Excel database Photographs Logs produced in Canvas software	Well Summary Logs
97/50a Onboard micropalaeontology	Onboard biostratigraphy report	Specialist software (StrataBugs, Ragware or Checklist II) Data files in ascii format	
98/20 Geotechnical sample analysis	Geotechnical test results,	Hard-copy report Excel files	

Project	Deliverable Description	Formats	Digital Data sets
	Geotechnical characteristic report		
98/21 Geochemical sample analysis	Interpretation report Data Charts	Hard-copy report	
97/14a TOBI – processing & interpretation	Clean navigation, Corrected imagery, AO Maps, Final report	Text file on CD-Rom Imagery Erdas Imagine	TOBI Image Mosaic
97/3 Structural elements nomenclature	Nomenclature map, geological sections and report	Maps and sections in ArcView format Hard-copy report	Geological Nomenclature
97/29 A Met-ocean strategy for the Rockall Area	Metocean data compilation report	Hard-copy report Access database on floppy disk	Metocean Data Inventory
97/51 Sea bed data assessment	Seabed data compilation report	Hard-copy report with Excel spreadsheets	Seabed Data Inventory
97/52 Environmental data gathering	Environmental data compilation report	See 97/51	Environmental Data Inventory
98/11 TTR7 Cruise license	CD-Rom of samples collected by TTR7 cruise (R/V Professor Logachev)	CD-Rom with Acrobat multi-media viewer (photos, charts, core logs, geochemistry data)	
97/16 Crystalline basement study	PhD Thesis	Hard-copy maps and report	
98/1 Gravity & Magnetic studies and 2D / 3D interpretation	Digital Atlas of 23 of conventional and shaded relief maps at 1:100,000 scale. Set of digital grids 3D whole crust model Set of digital grids of 3D model surfaces Final and interim reports	Hard-copy and CGM format on CD-Rom Zycor, ascii or other format on CD-Rom Hard-copy and CGM format on CD-Rom Zycor, ascii or other format on CD-Rom Hard-copy	Image Atlas, Gravity/magnetic compilation, 2D and 3D Modelling Maps
97/2 Statistical characteristics of reflectivity patterns in deep seismic profiles	Basement Geological Maps and sections Software designed PhD Thesis	 Exabyte tape with software on it Hard-copy	
97/21 Seismic imaging study	Final Report PhD Thesis	Hard-copy	
97/40 Gravity & Magnetic studies	Maps & Data files PhD Thesis	No information on. Hard-copy	

Table 8 Deliverables from the RSG projects

A strong recommendation to the RSG management committee is that a Data Handbook is put together to give some guidance on what formats, media types, key words and metadata necessary (see section 7.2). Without such information, it leaves the door open for project teams to deliver what they see fit and may lead to integration problems later and in the worst case scenario unusable data.

3 The Data Dictionary

3.1 Introduction

Most of the RSG funded projects are data-intensive but because delivery of information appears to be controlled by unrelated teams assembled to work on each project the task of achieving overall RSG information automation is complicated. While there is no lack of technology and application systems to serve the RSG projects, without some level of collective planning and control each of these individual systems are likely to be created and implemented in isolation with little regard for other project systems or project requirements.

Data consistency is a primary need of most data users and the principal goal of a data dictionary. Oil companies, oceanographers, environmentalists and others working in RSG related sectors rely on information from multiple sources and consequently where possible consistency is required to make judicious, well informed decisions. Even though there are a number of data standards to facilitate communication between different groups of data users working in the oil and oceanographic sectors it would appear that there is no single set of standards that would cater for the diversity of data collected as part of the RSG. As one of the possible steps towards addressing the problem and thus enhancing the utility of the RSG results, a data dictionary could be established.

This type of data dictionary will describe the data parameters used by a project or in this case a collection of projects. This information in the dictionary will be described as a hierarchical structure. Individual data parameters exist at the most detailed level (e.g. temperature, salinity, etc). Data parameters are then aggregated into categories (e.g. geophysics, hydrography, sediments etc.). In such dictionaries it is possible to have multiple levels of categories. A data dictionary should not be confused with a metadata directory which is a description of the specific data collected on a project. Metadata will typically refer to the scientists, the specific cruise, the where, when and how specific data was collected.

3.2 Developing a dictionary

Data dictionaries have been described in the past as sharing two characteristics with cathedrals they are usually large and complex structures, and they rarely seem to be completed. It is little wonder that organizations dealing with diverse data are reluctant to initiate a data dictionary project and yet, as with many major undertakings, the data dictionary begins with relatively simple steps:

1. **Define the scope of the project.** In this situation the scope includes all RSG funded projects.
2. **Consider the experiences of others.**
3. **Decide the structure of the dictionary.**
4. **Decide on the level of technology to be used in maintaining the dictionary.** In this situation the dictionary will be maintained in MS Access.
5. **Start defining the entities.** It is at this step that the RSG must simply jump in and start describing its parameters.

It is important to understand that a data dictionary will require ongoing maintenance and especially so in the RSG-type situation where projects cannot be definite about the specifics of the parameters they will gather and deliver. It is likely that most of the dictionary content will be defined as the projects deliver data.

3.3 An existing dictionary model

Having investigated existing dictionary models it would appear that the dictionary developed by the British Oceanographic Data Centre (BODC) offers the RSG with a good reference point for designing an RSG specific dictionary. This BODC dictionary structure is relatively simple and offers the user great flexibility. However, while this BODC dictionary comprises 2,700 terms, it was developed for oceanographic research in mind and consequently it does not include a significant number of the parameters required to completely describe all the data from the RSG projects and it also includes a

large number of parameter terms not required for the RSG. This data dictionary is structured with the following fields:

Parameter name – this is the detailed level of the description and in the BODC dictionary this field contains 2700 entries. It is possible that the entries in this field are repeated and the user needs to consider the information held in the description field to distinguish one entry from another. For example it is possible to have several entries for the parameter “Sea Temperature” but the entries in the description field will distinguish between temperature measured with a thermistor chain, reversing thermometer or from a CTD.

Category – under this field parameters are grouped. For example the category called “Hydrography” includes parameter entries for temperature, salinity, attenuation, sechi depth, pressure, sound velocity, etc.

Description – as mentioned this field describes subtle difference between parameters. For example it will distinguish between Corrected AMS 14C sediment age (foramenifera tests) standard error and Corrected AMS 14C sediment age (Globigerina bulloides tests) standard error

Method – describe briefly how parameters are measured. This field will for example distinguish between temperature measured with a hand-held digital thermometer and mercury in glass thermometer.

Units Description – gives units of measurement

Each parameter in the dictionary is given a unique code. An extract from this dictionary is shown in appendix 2.

3.4 The RSG dictionary

It is proposed that the RSG dictionary structure follows that of the BODC and therefore will include:

- A RSG parameter code
- A parameter name
- A parameter description
- A category
- A method
- Units measured

It is proposed that the parameter categories will include: Bathymetry, Benthic fluxes, Nutrients, CO2 system, Currents, Dissolved gases, Geophysics, Halocarbons (including freons), Hydrocarbons, Hydrography, Isotope chemistry, Metals, Meteorology, Microplankton, Navigation, Non-metallic element chemistry, Optics, Particulate load, Phytoplankton species, Pigments, Plankton production, Sediment biogeochemistry, Sediment properties, Sediment trap, Waves, Zooplankton.

While a number of basic parameter categories can be defined at this stage it is proposed that the detail of the directory can only be established with the involvement of the scientist collecting data and as or after they collect the data. The current project proposals and feed back from scientists does not have sufficient parameter-level detail to populate a dictionary. Consequently, a skeleton RSG directory structure with a number of relatively standard entries is given in Table 9 below. This directory will need to be elaborated and populated as data deliverables are received.

Code	Category	Parameter	Description	Method	Units
BAT001	Bathymetry	Bathymetric Depth		It will be necessary to see what is being delivered	
SBI001	Seabirds		The detailed parameters under these groups will need to be defined		
SBI002	Seabirds				
SBI003	Seabirds				
COR001	Cores	Biostratigraphy	The detailed parameters such as species, geological period, etc under these groups will need to be defined and replace the parameter groups listed		
COR002	Cores	Fluid Inclusion			
COR003	Cores	Mineralogy			
COR004	Cores				
COR005	Cores				
COR###	Cores				
CUR001	Currents	Horizontal Current Direction	Depends on the type of instrument used - it is likely that multiple parameter entries will be required with different descriptions for different techniques.		
CUR002	Currents	Horizontal Current Speed			
CUR003	Currents	Turbulence Intensity			
CUR004	Currents	East-West Current Velocity			
CUR005	Currents	North-South Current Velocity			
CUR006	Currents	Vertical Current Velocity			
CUR###	Currents				
CUR###	Currents				
			E.g. there will be one entry for Horizontal Current for ADCP and another for Electromagnetic Current meters		
GPH001	Geophysics	Magnetic Anomaly	The parameters under these groups will need to be defined		
GPH002	Geophysics	Gravity Anomaly			
GPH003	Geophysics	Boomer			
GPH004	Geophysics	Sparker			
GPH005	Geophysics	SideScan sonar			
GPH###	Geophysics				
HYD001	Hydrography	Sea Level			
HYD002	Hydrography	Electrical Conductivity			
HYD003	Hydrography	Salinity			
HYD004	Hydrography	Sea Surface Temperature			
HYD005	Hydrography	Sea Pressure			
HYD006	Hydrography	Sea Temperature			
HYD007	Hydrography	Sound Velocity			
HYD###	Hydrography				
MAM001	Mammals / Cetaceans	Species			
MAM002	Mammals / Cetaceans	Numbers			
MAM###	Mammals / Cetaceans				
MAM###	Mammals / Cetaceans				
MAM###	Mammals / Cetaceans				
NAV001	Navigation	Roll Angle			
NAV002	Navigation	Platform Heading			
NAV003	Navigation	Longitude East			
NAV004	Navigation	Latitude North			
NAV005	Navigation	Platform Speed			
NAV006	Navigation	Pitch Angle			
NAV###	Navigation				
OPT###	Optics				
PHY###	Phytoplankton				
PHY###	Phytoplankton				
PHY###	Phytoplankton				
PHY###	Phytoplankton				
SGY###	Seabed Geology		This is closely related to the Geophysics which includes some of the deeper Seismic parameters		
SGY###	Seabed Geology				
SGY###	Seabed Geology				
SGY###	Seabed Geology				
SGY###	Seabed Geology				
SGY###	Seabed Geology				
SED001	Sediment	Magnetic susceptibility	There is likely to be many other sediment parameters and some will be repeated for different sampling techniques		
SED002	Sediment	Dry bulk density			
SED003	Sediment	Mean grain size			

SED004	Sediment	Grain size mode			
SED005	Sediment	Median grain size			
SED006	Sediment	Sediment age			
SED007	Sediment	Proportion of sediment in a size class			
SED008	Sediment	Organic carbon content			
SED009	Sediment	Total carbon content			
SED010	Sediment	Inorganic carbon content			
SED011	Sediment	Total nitrogen content			
SED012	Sediment	Bottom shear velocity			
SED013	Sediment				
SED###	Sediment				
SED###	Sediment				
SED###	Sediment				
WAV001	Waves	Wave Direction			
WAV002	Waves	Average Wave Crest Period			
WAV003	Waves	Average Wave Height			
WAV004	Waves	Significant Wave Height			
WAV005	Waves	Swell Direction			
WAV###	Waves				
WAV###	Waves				
WAV###	Waves				
MET001	Meteorology	Wind Direction			
MET002	Meteorology	Gust Wind Direction			
MET003	Meteorology	Gust Wind Speed			
MET004	Meteorology	Wind Speed			
MET005	Meteorology	Atmospheric Pressure			
MET006	Meteorology	Relative Humidity			
MET007	Meteorology	Specific Humidity			
MET008	Meteorology	Air Temperature			
MET###	Meteorology				
MET###	Meteorology				

Table 9 RSG Dictionary Framework

This directory has been developed in MSAccess 97 and contains a single table for simple maintenance.

4 Data User Requirements

4.1 Introduction

A questionnaire and interviews where practicable, were used as tools to examine the data requirements of the RSG members. Our objective was to form a broad picture of data requirements in terms of; scientific data used by the members, data formats and software used, and also information on use of metadata and data management standards.

The RSG projects as described in Section 2, will produce a diversity of deliverables of different data types, on different media and in different data formats. The degree of synergy between the RSG members requirements and each of the RSG Projects is difficult to estimate as the deliverables are not finally decided upon and the majority of the members did not reply to the questionnaire.

4.2 Data Publication

Any data management option requires some thought as to the; degree of functionality the user requires, ease of use, access speed, data classes to be used and data constraints. The following sections 4.3 to 4.8 are concerned with these publication issues, whilst section 4.9 looks at the results of the questionnaire survey and their relevance to any data management considerations.

4.3 Audience

The RSG members (15 members) will be the key audience for the RSG project deliverables. Other audience participants will include the Petroleum Affairs Division and to an undisclosed extent other scientific agencies and also at some stage the general public.

Within the RSG member organisations who are the likely users of the RSG project deliverables? This is an important question and influences the options for any data management solution.

The following is a list of possible audiences for the RSG Project deliverables:

RSG Member organisations	Corporate Management, Scientists, Engineers
RSG Project Teams	Other Teams who are dependent on deliverables from other projects
Government Departments / Agencies	Petroleum Affairs Division, Department of the Marine & Natural Resources, Geological Survey of Ireland
Universities	Researchers, students
General Public	Researchers, press, general public information
Data Centers	Irish Marine Data Center and other internationally recognised data centres

Table 10 Categories of Audience for RSG project deliverables

Each audience category will have different deliverable requirements. For example an audience category such as corporate management may have a very different requirement (final interpretations, summary maps and reports) to the raw data requirement of a researcher for carrying out processing and interpretation. A decision regarding who will be the target audience for the data management solution should be made before deciding on a publication strategy.

4.4 Classes of Data

As has been shown from Stage 1 of this project, the RSG projects are producing a diversity of classes of data. As mentioned above raw data for example may be of little interest to corporate management but may be extremely useful to scientists. Table 11 below shows examples of who might be interested in which classes of data. Decisions need to be made regarding which classes would be distributed.

Raw data	Scientist's, Commercial agencies, Non commercial agencies, Universities
Processed data	Scientist's, Universities, Commercial agencies, Non commercial agencies General Public
Interpretations	Corporate management, Scientist's, Universities, Commercial agencies Non commercial agencies, General Public
Algorithms	Scientist's, University's
Index's	Corporate management, Scientist's, Universities, Commercial agencies Non commercial agencies, Government agencies, General Public
Reports	Corporate management, Scientist's, Universities, Commercial agencies Non commercial agencies, Government agencies
PhD's	Universities, Scientist's
Geology Sections & Maps	Corporate management, Scientist's, Universities, Commercial agencies Non commercial agencies, Government agencies
Images	Scientist's, Universities, Commercial agencies, Non commercial agencies Government agencies
Photographs	Scientist's, Universities, Commercial agencies, Non commercial agencies Government agencies
Core	Government agencies, Universities
Samples	Government agencies, Universities
Geochemical Sample Results	Scientist's, Universities, Non commercial agencies, Government agencies

Table 11 Classes of Data and the audience who might use them

4.5 Data Distribution Constraints

If audiences such as the general public are chosen, there may be confidentiality problems. Even within the group of RSG members there are problems with some data sources used, where certain member organisations have subscribed to the data source but others have not. Examples of such a situation include the wave height data source used in project 97/29, the use of seismic data in 97/3 and the use of Gloria data in 97/14.

There may also be data distribution or data restriction issues if the data management solution chosen is to be used internally for the sharing of data between the different project teams during the course of their projects.

4.6 Level of Functionality

The audience for any data and information, influences the degree of functionality to be built into any distribution system. For example, corporate management maybe interested in a multi-media type solution that allows the final results to be displayed through an easy to use interface. Scientists meanwhile maybe more concerned with the procedures used in collection of data and therefore more interested in having detailed metadata rather than a fancy interface.

Should the system fully integrate all data or just some of the data? Should the distribution system have some sort of search engine that will search all of the text data for key words? Should the search engine work along the lines of location? There are numerous options, many of which can be ruled out through cost. As a rule of thumb the higher the degree of functionality included in the distribution system the more expensive to develop.

Browsing	<ul style="list-style-type: none"> - Searching using words - Search using a graphics location map screen - Interactive by hyper-linking to other parts of the system
Aesthetics	<ul style="list-style-type: none"> - Degree of intuitively of the interface - Graphics user interface - Command line interface

Table 12 Examples of Functionality

4.7 Performance Issues

Access speed to the data maybe an important issue. Is it enough to have access to the data within a week or is it necessary to have immediate access to the data? Is there a requirement for just an index of what data is available from the projects or is it necessary to have all of the data live and on-line. If the data is to be on-line can a database solution handle the quantities of data involved with some of the RSG projects, in particular the imagery projects (97/11, 97/14, 97/16, 98/1, 97/21, 97/2, 97/40).

4.8 Update-ability

Is there a requirement for the data to be updated. Will the updates be on a regular basis or a continuous process. This may have an effect on CD-Rom solutions where new CD's would be required to be pressed with project data updates.

4.9 RSG Members Feedback

As mentioned in 4.1, a questionnaire was designed with the objective of gathering information from the oil company RSG members with regard to; data types they are interested in, data formats and standards that they use, use of metadata, and the degree of a solution that they require. The questionnaire is included in Appendix 1.

Management Committee Member	Person Sent the Questionnaire	Response
AGIP	Franco Polo Ron Lansdell	
Andarco	Richard Hook	
BG Exploration	Peter Haynes Jon Peachey	
BP	Chris Bird Steve Cawley	Completed questionnaires submitted
British Borneo	John Robbins	
Elf	Anne Schwab	Completed questionnaires submitted
Enterprise Oil	Ciaran Nolan	Interview
Mobil Oil	Gerry Worthington John Wood	
Murphy Offshore	Jeremy Gardiner-Brown	
PAD	Noel Murphy Peter Croker	Interview Completed questionnaires submitted
Phillips	Fergus Cahill John Chamberlain	Interview
Saga	Anton Kjelaas	
Shell	Herve Quinquis	
Statoil	Geirr Hærr	Completed questionnaires submitted
Total – Marine Oil	Tony Rochester	
Arco	Chris Atkinson	
IMDC	Bronwyn Cahill	Interview

Table 13 Responses from RSG members

Six of the members responded to the questionnaire or representatives made themselves available for interview.

4.9.1 Data Types

Table 14 summarises the rankings given by the questionnaire respondents as to which data types they currently use and their importance to that organisation.

Data Type	Data Sub Type	Rank	
Geophysical (averaged on 4 questionnaires)	Side Scan Sonar (LF)	3.25	Medium (3-4)
	Side Scan Sonar (HF)	3.5	Medium (3-4)
	Single Channel Seismic Reflection	3	Medium (3-4)
	CDP Seismic Reflection	5	High (4-5)
	Seismic Refraction	3.5	Medium (3-4)
	Digital Shot Point Navigation	5	High (4-5)
	Bathymetry	4.5	High (4-5)
	Magnetics	4.75	High (4-5)
	Gravity	4.75	High (4-5)
Biological (averaged on 2 questionnaires)	Biostratigraphy	5	High (4-5)
	Birds	5	High (4-5)
	Cetaceans	4	High (4-5)
	Fish	3.5	Medium (3-4)
	Benthic Fauna / Flora	4	High (4-5)
	Plankton	3	Medium (3-4)
Chemical (averaged on 2 questionnaires)	Fluid Inclusions	3	Medium (3-4)
	Water Chemistry	3	Medium (3-4)
	Geochemistry	4.6	High (4-5)
Physical (averaged on 2 questionnaires)	Wave Data	4	High (4-5)
	Meteorological Data	4	High (4-5)
	Currents	4	High (4-5)
	Temperature	3	Medium (3-4)
	Salinity	2.6	Low (< 3)
	Geological Interpretation	5	High (4-5)
	Sedimentology	5	High (4-5)
	Porosity	5	High (4-5)
	Permeability	5	High (4-5)
	Core Description	5	High (4-5)

Table 14 Ranking of Data Types and their use by the RSG members

As can be seen from Table 14, there are certain data sets that the members regard as of high importance. The geological and geophysical (Magnetics and Gravity) as would be expected all come out as very important. The objective of this section of the questionnaire was to gather information on the perceived importance of various scientific data sets to the RSG members so that data sets could be prioritised for databasing. The ranking is from 1 to 5 with 1 being of low importance.

4.9.2 Data Formats and Software used

The data formats are related primarily to the software used for that particular scientific discipline. Some of the members are using comprehensive software systems that can handle a number of different data types, for example Landmark handles both seismic, navigation and geological data-sets. Formats for navigation information include UK00A P1/90 and UK00A P2/91. SEG-Y is a common format for seismic data.

4.9.3 Metadata

All of the questionnaires replied that they had a requirement for metadata. Some organisations have an actual requirement for metadata to be supplied with a dataset.

5 DATA MANAGEMENT CASE STUDIES

5.1 Introduction

Data Management is currently an important key issue for many scientific disciplines, in particular for multi-disciplinary projects involving many different parties and disciplines. The RSG project is not unique and there are some good examples of data management procedures being used in projects of similar scope.

The key aspects of data management that are currently being looked by the industry are;

- 1) **Data Models** enabling easier transfer of data and understanding of the data descriptions used.
- 2) **Data Standards** ensuring that data is collected and managed to agreed international standards.
- 3) **Data Formats** standardisation allowing transparent transfer between software.
- 4) **Data Archiving** and the problems of archive media selection.
- 5) **Communications**, with the advent of the Internet, data can be managed centrally but accessed globally and the Internet is radically changing both areas of project management and also the ease of acquiring and sharing information and data.

5.2 Data Models

A data model is similar in many respects to a conventional dictionary. A data model, however, goes beyond names and definitions and details the characteristics that each entity (item, object) may have. Some characteristics, such as the identifier or description, provide simple identifying and explanatory information about an occurrence of the entity. Other characteristics specify interrelationships between entities. For example, a Well has the characteristic of containing one or more related Wellbores, each of which is specified as another entity in the model.

Perhaps the biggest difference between data models and conventional dictionaries is that data models are designed primarily to enable the storage of electronic data in computers. Many details of the model may be of little interest, or even confusing to the casual reader, but these details are required by computer systems.

There are currently two key contenders for developing a standard data model for the oil exploration and production sector (McNaughton 1995); POSC (Petrotechnical Open Software Corporation) both the RDBMS and object orientated database options and PPDM (Public Petroleum Data Model). Both organisations are non-profit making.

1) POSC's Epicentre Data Model¹

<http://www.posc.org/>

POSC is a collection of 100 oil companies, consultants and organisations whose key contribution so far is the Epicentre data model which is rapidly becoming the accepted standard for exchange of data and information among oil companies, oil field service companies, governments and other players in the industry. POSC's members pay an annual subscription ranging from 2,000US\$ to 125,000US\$ depending on the size of the company. 9 of the RSG members are members of POSC.

The Epicentre model is complicated and is based on the EXPRESS² language and requires a high standard of database knowledge before the model can be used.

In Epicentre, there is an alphabetical listing of names and definitions for more than 750 real-world technical and business objects concerned with petroleum exploration and production. In POSC's data modeling terminology, these objects are called entities.

¹ Source of the following information: The following sections are abbreviated from the POSC web-site and POSC specifications CD-Rom Version 2.2.

² EXPRESS is a language used to describe information in an object orientated database, developed by the ISO standard for the exchange of Product Model Data.

A critical point to understand about Epicentre is that it is a logical data model, and as such, it is not directly implementable as a physical database. Epicentre is documented precisely in the EXPRESS language, but EXPRESS is not the equivalent of a data definition language (DDL), such as Standard Query Language (SQL) DDL³ used in many database applications. To build a POSC data store, it is necessary to transform Epicentre EXPRESS into a set of DDL statements using rules consistent with the target data store's database management software. POSC refers to this process as projection.

Before going any further it is perhaps worth mentioning the two most important types of database on the market; RDBMS and Object Orientated. RDBMS is a database where the data is expressed in terms of rows and columns. It is a very common structure for databases but it does have problems in flexibility once a database has been designed and implemented. RDBMS also cannot handle vector data within its structure. Object Orientated databases use data models to describe objects. The database has no predefined structure and individual data types can be grouped into class structures using natural real world situations. Object orientated databases allow for much more flexibility in design and being able to modify databases. Epicentre has both an object orientated data model and an RDBMS option.

The key to success in integrating the requirements of the various disciplines of E&P is through the use of a single coherent data model with an integrated architecture. Epicentre's integrated architecture is based on the distinction between objects, the properties or characteristics of objects and the activities that utilize objects and determine their properties. This separation of information supports the business requirement that the properties of an object have multiple versions or descriptions each clearly associated with its own documented reason for existence or process history.

The architectural principle of separating objects, properties and activities is applied throughout Epicentre. This gives Epicentre a different character from many data models currently in use. Thus, items one sees as simple attributes of entities in some data models appear in Epicentre as entities in their own right.

Another fundamental part of the Epicentre architecture is the concept that many entities have characteristics of spatial representation. In exploration and production, much of the information recorded is composed of coordinates that describe an object's location relative to the earth. To facilitate the general use of spatial data across the model, POSC have defined a set of generic spatial objects and spatial relationships. Each of the various geometrical business objects in different parts of the model may be connected to the earth through relationships with one or more generic spatial objects.

The Epicentre model currently has been set-up to run on Informix, Oracle and Sybase database systems.

Example of Epicentre Data Entity Format

Entity Name : ***SEISMIC_ACQUISITION_ACTIVITY***

An activity associated with acquiring seismic trace data, for example, permitting, shot hole drilling, surveying/positioning, recording, navigation processing, overseeing, etc.

Some of the Attributes associated with the above entity

name (O, K, I: ndt_name)

The name or label given to the activity. Inherited from activity.

ref_existence_kind (M, K, I: ref_existence_kind)

The lifecycle kind of the activity e.g. actual, planned, required or predicted. Inherited from activity.

typical_activity (O, K, I: typical_activity(activity))

Gives the typical activity which acts as the template or design for this activity. Each seismic_acquisition_activity may be an occurrence of one typical_activity. Inherited from activity.

containing_activity (O, K, I: activity(contained_activity))

³ SQL is a language used as a means of defining, accessing and modifying a relational database (RDBMS).

Specifies the containing activity. The containing activity indicates the activity for which this activity is a component sub-activity. Constrains the time of the contained activity to be during the containing activity. Each seismic_acquisition_activity may be contained by one activity. Inherited from activity.

seismic_geometry_set (O, K: seismic_geometry_set(seismic_acquisition_activity))

The primary geometry set with which this activity is identified. This will typically be the "survey". Each seismic_acquisition_activity may be identified by one seismic_geometry_set.

cost (O, I: ndt_money)

The cost of the activity. Inherited from activity.

2) PPDM

<http://www.ppdm.org>

PPDM was a model instigated by a group of Canadian companies and has gained widespread acceptance within the E&P industry. The model uses a relational database structure (RDBMS) and the Structured Query Language (rather than Express used by POSC). It forms the basis of several vendor's products (Landmark, Schlumberger) and is used by many oil companies in-house. Since 1993, there has been talk of merging PPDM and POSC.

3) OPEN SPIRIT Initiative

The Open Spirit Initiative was instigated by Shell, and is not really a data model rather software translation options for integrating different E&P software packages. Rather than defining new data models it takes current software vendors data products and sits software on-top making translation of data between packages as seamless and transparent as possible.

3) MAST Data Management Procedures

<http://www.marine.ie/datacentre/projects/edap/>

The Irish Marine Data Centre developed the EDAP Document (Guideline on Electronic Data Publishing for MAST Projects), as a resource to maximise the re-use of marine data. Also as a guide to those generating data in producing electronic products of marine data and information that can be viewed, navigated and distributed electronically using such media as CD-ROM or World Wide Web (WWW). The guideline is aimed at those who fund and participate in marine science and technology projects including project co-ordinators, participating scientists, data managers, developers, publishers and policy makers.

The following is some of the procedures used for the data model.

Reporting on data collection in MAST projects

1. A copy of the cruise summary report should be sent to the project coordinator within one week after the field experiment ends.
2. A copy of the experiment report (cruise report) including station lists, etc., should be sent to the coordinator within one month after the field experiment ends.
3. Inventories of continuous observations should be updated regularly and copied to the coordinator.
4. The coordinator adds the copies of the cruise summary report forms and inventories to his/her regular management reports.

Managing data handling issues in MAST projects

1. The leading scientist within a project (e.g. the coordinator or the steering group) has to identify persons and/or institutions who have the joint duty
 - to receive the raw data including its documentation, to correct for instrumental errors and to safeguard the corrected and documented data as close as possible to internationally agreed standards as far as they exist,
 - to take care of quality assurance and quality checking of the data (all levels of processing),
 - to take care of preliminary banking of data for project use,
 - to take care of final banking or publishing of the data for public use.
2. A data management plan has to be a part of the task description and the financial planning of the project. The data management plan shall describe work and responsibilities concerning: collecting, quality checking, banking or publishing of data. Institutions involved in the banking and quality checking but not participating in the projects have to be contracted formally.
3. The leading scientists of the project (e.g. the coordinator or the steering committee) decide not only on technical matters of data handling (e.g. time delays) but take care of the interests of the

data originators. In particular, the leading scientists ensure that the people and institutions in charge of handling the data serve the needs of the data originators and project scientists.

4. The inventories of the preliminary banked quality checked data are updated every 6 months and copied to the coordinator who adds them to his/her regular management report. The final project report will state that the banking of the data is in accordance with the data management plan. EDMED forms are included in the report describing the final data sets produced by the project.

Property rights on data

1. Data originated within MAST are understood to be a property having the legal status of “foreground information” as this term is defined in the contract. In that sense the contract sets the basic rules for sharing of data within the project, among different MAST projects, with other community undertakings and with third parties.
2. Data disposed for final banking are flagged for the limited access for a period preferably not longer than 6 months after the formal end of the project. The access limitations cannot be more restrictive than those on the sharing of “foreground information”.

Other web-sites of interest:

1. Geoshare ; <http://www.geoshare.org>
2. CDA(I) Common Industry Data Access (Initiative); <http://www.cdal.com>

5.3 Operational Standards

Unification of operational and data standards is one of those dream goals that academics and researchers are striving for. Standards vary from country to country, discipline to discipline and organisation to organisation. The E&P sector is no different to any other and is grappling to develop some sort of cohesive effort to define international operational standards.

The questionnaire survey of both the project teams and the RSG members highlights the lack of clear operational standards. As mentioned in section 2.4, reputation and good accurate documentation is as good a standard as any other. Saying that you comply to such and such a standard means nothing if it is not verifiable and checked.

Other web-sites of interest:

1. UKOOA UK Offshore Operators Association Ltd ; <http://www.ukooa.co.uk>
2. IADC International Association of Drilling Contractors; <http://www.iadc.org>
3. SEG Society of Exploration Geophysicists; <http://www.seg.org>
4. IP Institute of Petroleum; <http://www.petroleum.co.uk>
5. E&P Forum; <http://www.nts.no/epforum>
6. NORSOK Norwegian Oil and Gas Standards; <http://www.nts.no/norsok>
7. API American Petroleum Institute; <http://www.api.org>

5.4 Data Format Standards

The degree of data format standardisation principally depends on whether the software is specialist or proprietary. The more popular and universal the software the more likely it can export data in a universal standard format. In general, software operates using its own formats but with most packages there are export options. Where standardisation has taken place by the industry it tends to be in individual scientific disciplines. For example, SEG-Y is a well-recognised data standard for seismic data and most if not all software that uses seismic data will accept this format. However, there are different varieties of SEG-Y format so the matter is not entirely clear cut. SEG-Y standards are completely irrelevant to any other scientific discipline.

Popular software packages such as Word, Excel, CorelDraw etc accept many different formats and there is much more transparency now between such software for sharing data. Graphics formats have also become more standard, but there are still the odd exceptions.

5.5 Archiving Media Standards

Archiving is a key topic of debate at the moment. In the rapidly changing world of technology it is important to archive valuable data assets on media that is likely to last for the foreseeable future and at least be on a media that will allow for translation to other media. There is currently a lot of confusion

about what media types are the most suitable for large data-sets such as seismic data and what media types will still be around in a few years time.

For example, if you discovered that you had some documents archived in the early 80's on a word processor cartridge you may have extreme difficulty finding someone who will be able to translate the file now.

The CD-Rom is very much in favour at the present time, but the new DVD tapes hold out huge potential to the E&P industry as they can store many magnitudes greater than the 600 Mbytes that are currently possible to be stored on a CD-Rom. At present, though there is no single DVD standard and until the industry sorts out which DVD becomes the standard it is best to wait.

5.6 Communications and Internet accessible databases

The Internet is probably the single most important development in the last 5 years for realising the power of good data management. Databases on the Internet allow access globally to an organisation's data resources. Progress in solving security issues has been made and improving all the time. The benefits of Internet solutions cannot be over emphasised:

- Real time access to large corporate databases
- Single distribution cost for development and maintenance
- Truly global access from an oil rig to a jungle station
- No problems with restricting access and protecting the corporations assets
- An invaluable aid for project management of multidisciplinary and international projects

Currently the main problem today with using the Internet for data transfer is bandwidth but this is expected to rapidly improve. As much of the RSG data will be very large file sizes, there are technical problems with transfer of data sets such as the TOBI imagery, via the Internet.

There are several good working examples of electronic data submission/query forms (metadata) on the Internet at present. These forms can be downloaded from the Internet and the query sent back to the data centre by post (e.g. EDMED, <http://www.marine.ie/datacentre/projects/edmed/>), or the form can be viewed and filled in and sent back to the originator directly on the internet (e.g. Reids, <http://www.informatic.ie/cds>). Data Submission forms for both examples are included in appendix 3.

Listed below are some examples of where the Internet is being used to allow global access to large data resources.

1. NGDC National Geophysical Data Center: <http://www.ngdc.noaa.gov>

The National Geophysical Data Center (NGDC) manages geophysical data in the fields of marine geology and geophysics, paleoclimatology, solar-terrestrial physics, solid earth geophysics, and glaciology (snow and ice). In each of these fields NGDC also operates a World Data Center (WDC-A) discipline center.

Although not all of their data holdings are available through NGDC's Geophysical on-line Data (GOLD), new data, meta-data, and information are continually being added. Data and inventories in many disciplines are fully searchable and selected listings, data, and images can be downloaded.

2. NODC National Oceanographic Data Center: <http://www.nodc.noaa.gov/>

The National Oceanographic Data Center (NODC) is the U.S. repository and distribution facility for global ocean data. The NODC ensures that oceanographic data collected at great cost are preserved and maintained in a permanent archive where they are available for use by scientists, engineers, resource managers and planners, and others.

3. World Data Centre: <http://www.nodc.noaa.gov/NODC-wdca.html>

World Data Center A (WDC-A) Oceanography is one component of a global network of discipline subcenters that facilitate international exchange of scientific data. Originally established during the International Geophysical Year of 1957, the World Data Center System functions under the guidance

of the International Council of Scientific Unions (ICSU). WDC-A, Oceanography is collocated with, and operated by, the U.S. National Oceanographic Data Center (NODC).

In accordance with principles set forth by ICSU, WDC-A, Oceanography acquires, catalogues, and archives data, publications, and data inventory forms and makes them available to requesters in the international scientific community. To protect against catastrophic loss and to improve user access, WDC-A provides copies of data it receives to its counterparts, World Data Center B (Obninsk, Russia) and World Data Center D (Tianjin, China). Oceanographic data contributed to WDC-A become automatically available to scientific investigators in any country. Thus, there can be no restrictions or limitations placed on data exchanged through the WDC system. For certain types of data, the exchange of inventories of available data in a WDC subcenter may be considered acceptable in lieu of the transfer of the actual data sets.

4. **IMDC:** <http://www.marine.ie/datacentre>

The Irish Marine Data Centre, an Integral Part of the Irish Marine Institute, has a team with diverse skills, ranging from oceanography through geology to information technology.

It provides solutions to Irish marine data management problems, whilst also actively providing services to a variety of international groups. As part of its solution the Data Centre provides easy-to-use, sophisticated graphical database applications that allow the user to easily retrieve information.

In addition, the Data Centre receives, processes and quality controls the information used to populate these databases.

5. **BODC:** <http://www.nbi.ac.uk/bodc/>

The British Oceanographic Data Centres provides data management support for the UK marine sector. The BODC collaborates on behalf of the UK, in the international exchange and management of oceanographic data. An example from the BODC data dictionary is included in Appendix 2 and has also been used as a foundation for the RSG Data Dictionary outlined in section 3.

6. **AODC:** <http://www.aodc.gov.au/>

The Australian Oceanographic Data Centre (AODC) was established in 1964 within the Royal Australian Navy (RAN) as a result of an agreement between the CSIRO Division of Fisheries and Oceanography, the Bureau of Meteorology and the Department of Navy. The aim of this agreement was to improve the communication of oceanographic information and data within the Defense and civil communities. AODC was established within the Hydrographic Service of the RAN and from 1965 to 1982 consisted of a single civilian officer with administrative support provided Hydrographic Office personnel.

During the 1980s the AODC steadily grew with an increase in personnel and other resources to keep pace with rapidly growing demand for oceanographic information. In 1993, the AODC separated from the Hydrographic Services and relocated to Maritime Headquarters within the Operations Division of the Maritime Command.

Today, the AODC is recognised internationally by the IOC as the national data centre for the acquisition, archival and management of physical oceanographic data in Australia and the focal point for international data exchange. The Head of Marine Agencies (HOMA) Committee and its associated agencies; the Commonwealth Spatial Data Committee's (CSDC) and HOMA's Marine Data Group (MDG) and the ORV FRANKLIN Steering Committee also recognise AODC as the national oceanographic data management agency.

5.7 *Relevance of the case studies to RSG Data Management Requirements*

Data models such as POSC or PPDM, unfortunately do not cover all of the scientific disciplines covered by the RSG projects. POSC in particular is more concerned with oil well drilling and seismic and has yet to look at environmental or metocean data. Also, only 9 RSG members are members of POSC.

Both the POSC and the MAST data dictionaries are a good basis for the designing of any data management system and this idea is used in our recommendation (see section 7).

The Internet Database Centers such as NODC are good examples of how large data sources can be shared internationally. One comment that should be made at this stage is that many of these database centers are not particularly easy to use and are geared to the scientific and academic communities. For publishing of information they are not particularly great sources. The work by the IMDC on the EDAP project gives good examples of data management options for different audiences. The Pirate multimedia option is a very good option for provision of information to a wide audience, whilst the CD-Rom option with just data files is more applicable a solution for a research audience. The Internet is of growing relevance and importance for the distribution of information and eventually when bandwidth allows for data transfer. A complete Internet based solution forms part of our overall recommendation.

Data formats and standards, are import issues as it may lead to data being unusable if the end user cannot decipher the format used. For projects like the RSG where there are a number of different end users involved, it should be where practicable, for the projects to deliver their deliverables in formats that are the most practicable and “common” in usage. We have taken this on-board and we recommend the publication of a Data Handbook to provide guidelines to the RSG project teams.

6 Options

6.1 Introduction

Section 2, highlights the diversity of the RSG projects both in terms of scientific disciplines, data types, data formats operational standards and deliverables. Without some guidance, project teams may submit data on media and in formats that may make the data unusable to others. Any deliverable (in particular data sets) should have accompanying documentation (metadata) so that the end user can fully utilise and be aware of any constraints or limitations. Section 7.2, details the requirement of a Data Handbook, which provides guidelines for project teams for the distribution of data and other deliverables. Section 7, is a detailed recommendation for an Integrated System for the RSG (IPIPS).

Before choosing any data management option, it is important that the publishing requirements highlighted in Section 4 are used as a basis for any decision. These requirements include deciding on the audience, data classes for distribution, constraints on individual data classes, functionality issues, performance issues and updateability.

6.2 Direct Dissemination – “The Cardboard Box Option”

One option is to involve no database solution what’s so ever, and just reproduce the final deliverables where possible and distribute to each of the RSG members. Any data or deliverable that is not reproducible may be stored at a central repository such as the IMDC or the PAD. A simple deliverable index listing all items delivered by the projects would be maintained by the secretariat for the duration of the RSG project and supplied to parties on request.

6.3 Metadata Index – “Data Inventory”

The diversity in nature of the projects and the number of teams and individuals involved ensure that there is a requirement for a metadata index to record information pertaining to project deliverables. The index can be created either in a database system such as Access or as a hard copy print out. The key to the index is to make it as comprehensive as possible. Database centers such as the BODC and the IMDC (see section 5) have established metadata requirements for any data set sent to them for storage on their systems. The metadata is typically filled in on a form provided by the data centre. Key sections of the form are mandatory (name, location etc) whilst other sections are voluntary. An example of the EDMED metadata form and the RIEDS metadata form is enclosed in Appendix 3.

The data dictionary (Section 3) can be incorporated with the metadata index to provide a searchable database system for information on data, sources of information and projects.

The metadata index could be distributed as a digital file (Access, Excel etc.) by post or on-line on a web page. The advantages of the web page are; the ease of use to the user, ease of updating, global access and security access management.

6.4 Database – “Data distribution”

The data from the RSG projects does not readily itself for databasing, as the deliverables will be in a variety of data formats (hard copy reports to maps and images). There are a number of options that can be used for data distribution.

Data Storage Options:

1. Direct dissemination to all RSG members – no central storage except for archive
2. Central Storage at a Data Centre such as the IMDC
3. Central Storage at a number of different locations for individual scientific disciplines

Data Access Options:

1. Direct dissemination of all of the data to each of the RSG members
2. Central distribution on request
3. On-line solution via the Internet

Data Format Options:

1. Data distributed in its proprietary format
2. Data translated from proprietary formats to formats suitable for databasing

Data Media Options:

1. Tapes (Exabyte, DAT)
2. Zip Disks (up to 120 Mbytes)
3. CD-ROMs
4. DVD tapes

Database Options:

1. RDBMS (Access, Oracle, SQL Server, Informix)
2. Object Orientated (Oracle 8)
3. Proprietary Software (Finder)

Translating data from a proprietary format to another format, for example, a seismic image in SEG-Y to a PDF graphics format, may lead to a reduction in resolution, loss of data, loss of colours, loss of accuracy etc. The benefits of translation are the ease of use to other users who do not possess software that will read the proprietary format.

7 Proposed Solution

7.1 Introduction

The overall proposed system, based on the options described above, is provisionally referred to as IRSGS (Integrated RSG System). It has three software components, each of which may be implemented independently but, resources permitting, together will provide a strong technical foundation on which to build a complete integrated system for all RSG type projects. The three components are the RSG Data Inventory, the Generic RSG Information System and the RSG Web Page.

As part of our preferred solution, we strongly recommend that the RSG design and implement a Data Handbook. These are used commonly now on similar multi-disciplinary projects.

7.2 The RSG Data Handbook

The RSG Data Handbook will describe data management procedures for the RSG projects, including how to submit, access, retrieve and request data. It will give guidance on data formats and other data management practices and policies. The handbook will not necessarily supercede already existing procedures, but will ensure that all the project data are described and controlled to an agreed standard. The procedures outlined in the handbook, should be based on procedures commonly in use by many scientists, particularly in the case of data submission and cruise reports.

7.3 The RSG Data Inventory (RDI)

The objective of the RDI is to describe data sets collected by the various RSG funded research projects. It is intended that this component of the overall IRSGS will not contain any data but instead will be used to direct users to the many sources. The data required to populate the inventory will be supplied from RSG project partners. The data should not be complex and key fields should probably include:

- Data set title
- Data originator/source with contact details
- Data set description
- Data Classification (according to the RSG data dictionary)
- Publication details arising from data
- Availability of data and format details
- Brief notes on quality and methodology
- Parameters included
- Data set coverage

All data sets will be reported on a RSG data inventory form (RDI Form) which should not exceed a maximum of four A4 pages in length. The proposed method for compiling the initial data set with which to launch the inventory is to conduct a series of interviews or workshops with key supporters of the project. It is expected that if significant support can be shown for this inventory at an early stage, even by a relatively small number of key players, then the RDI project is likely to have a far greater chance of success.

The RDI system should be developed to run on the internet to allow easy access for its audience. The system will have a very user friendly interface, it will be based on MS-Access database initially but must be portable in the future to any of the well known client server RDBMS's (Oracle, SQL Server, Sybase).

It is expected that the RDI system should be operational, with the first populated data records, within six months of project start.

It is possible that the RDI could be hosted at either the PAD, Irish MDC, CSA, ERA or any of the oil companies or by some contractor specifically charged with its maintenance.

The benefit of this inventory is threefold.

1. If the RSG is to move more towards a complete integrated management of the data collection it funds, this inventory will easily identify data sets with potential for inclusion and in doing so, will help identify gaps that need to be addressed.
2. It will give key users (e.g. the RSG oil company members) easy access to data sources and should lead to a greater utilisation of RSG data.
3. It will be relatively easy to implement and maintain and it will involve almost all data providers and therefore it will very quickly raise the profile of the RSG database project.

7.3.1 Outline specification of RDI

General:

Preferably the system will be developed to operate on the Microsoft NT Web Server and should be proven to work with at least MS Explorer and Netscape Navigator. The design of the system will allow for relatively easy future expansion. It will have a user friendly interface and it will allow customisation by an expert system administrator. The graphic user interface will most probably be developed primarily using Java and Active Server Pages. At the end of the project, all applications and applets and the fully documented software code should be made available so that the long term project team can modify or expand the system in the future without necessarily referring back to the original developer.

It is proposed that a GIS component be included which will allow the user search for data or display data using the map. It should be possible for the operator to change base maps, it will allow multiple area searching, basic thematic mapping and must be capable of operating effectively over dial-up access. This system will be developed to interface with the RDI MS Access tables. It should not be difficult at a later stage to upsize to a client server system such as Oracle or SQL server.

Features & Functions:

Although a final technical specification will not be agreed until the project starts, it is proposed that the following should be included.

1. **Searching:** In addition to a powerful search interface which will allow the user build relatively complex queries, the system will also incorporate a menu of prepared queries for the non-expert users. This set of menu queries will be maintained in an access table which can be customised by the system administrator as required. The system will allow searching using the maps.
2. **Reporting:** The system will incorporate a variety of reporting options which will be output in HTML to the user's web browser. This can in turn be cut and pasted into MSWord, Excel or almost any other windows applications. The exact structure of these reports will be agreed in discussion with the RSG secretariat and if additional reports are required, it should be possible for the long term systems administrator to develop them.
3. **Data Entry:** For security, performance and quality control reasons, it is proposed that the core system is not used for data entry. Instead it is proposed that forms are filled out on paper or preferably on disk by the data originator and subsequently are uploaded by the system administrator. It is recommended that an electronic form filler with the PDI form be provided to all RSG data collectors.
4. **User registration:** In order that the IRS GS team can monitor system usage, it is proposed to incorporate a user registry system. Users will be required to log on when entering the system.

5. **Links to data holders** : The system will allow internet links to data holders as appropriate

7.4 The Generic RSG Information System (GRSGIS)

This represents the core of the long term integrated information system for RSG data. It will be developed so that it has three levels of information. The top level will comprise of “metadata” or the data set description, which will be largely based on the RDI record structure described above. The middle level will be used to technically describe the structure of the data set in the system – the variables, units and possibly geographic references for the data. This middle level should probably be hidden from the normal user. The bottom level will contain the RSG data sets themselves in what could include almost any table structure or even data objects such as images. Data objects other than images (e.g.. GIS files) could be launched with their original application. In other words if a record referred to a MapInfo data file then the system could launch MapInfo with the data file if MapInfo existed on the users computer.

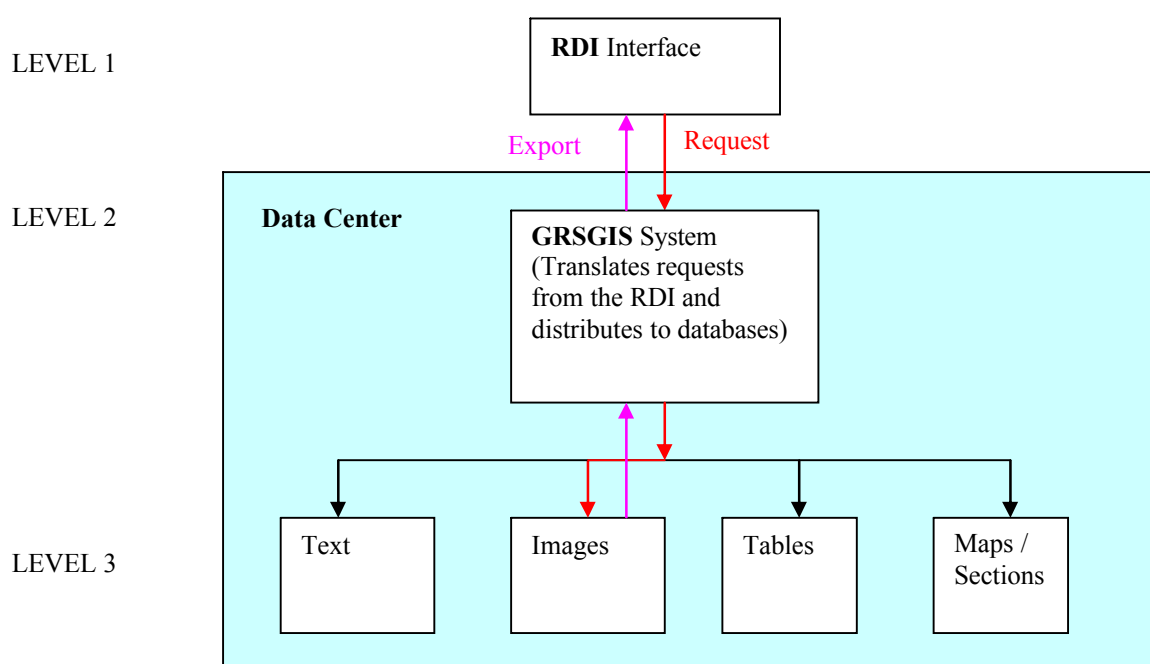


Figure 3 IRSGS Operational Flowchart

It is proposed however that administrative controls be put in place to achieve a convergence rather than divergence of data structures and therefore move closer to data comparability and integration. It will be important in these procedures to address the definition of variables. These procedures and controls will be documented as part of the development phase and certain controls may be coded in the application.

Once operational the administrator or a team manager will be able to create new instances of tables for different data collection scenarios – as an example of a bottom level template an administrator may create a table for recording the various relevant data to comply with the requirement of a coring project and this template may then be implemented as separate instances for several different sites.

It is expected that this component of the overall system will have three levels of user. There will be a requirement for an administrator with good technical expertise and once the system is operational this person will act as floating support for all organisations that implement the system. There will be a senior operator or manager at each organisation implementing the system and this person will be in a position to create new instances of a template but cannot edit the template database structure. The third level of user is the operator who will require much less technical IT skill will use the system to input and retrieve data as required.

7.4.1 Core content for the GRSGIS

This study revealed that there will be many diverse data types collected in RSG projects and therefore it is difficult in a short period to clearly identify all of the key data sets for inclusion in the proposed system. Furthermore, even if it was possible to list and evaluate all of the important data sets, it is considered likely that recommendations made here will require periodic review in light of changes to project design as they are implemented.

It is proposed that when selecting data for inclusion in the system a mechanism should be agreed for prioritising data sets. One such system would involve scoring each dataset on a scale 1 to 5 under six headings as follows:

Data availability: If data from a project was well managed and easily available from one source, then it would score 5. If a data set was not particularly well managed, but available from one source, then it scored a 4. If data was distributed among many different sources (i.e. all the partners in a project had a bit of the data) then it scored lower until a score of 2 is reached. Data with very restricted access gets a score of 1.

Relative Effort in Processing and Maintaining data: If a data set structure was well defined and of good quality or easy to check it scored highly. If a data set was very specialist, complicated and difficult to process (e.g. it is recorded on paper) then it scored low.

Estimated Quantity of Data : If it was estimated that a large quantity of a particular type of data existed, then the investment involved in setting up a system to manage it would achieve greater usage and therefore scored highly. A data set scored low if it was estimated that very little existed

Demand for data: Estimates of demand are based largely on the consultative process in stage one of the project. If there appeared to be a significant demand then it scored highly. If the demand was low, then the score was low.

Perceived Benefit of data: Depending on the use of a data set, its impact may have a varied perceived benefit. For example, a data set which may influence a decision regarding a multi-million pound investment will have a greater benefit than one that is collected for academic research and therefore will score higher.

Potential for Integration with other data: Although very difficult to assess without more in-depth analysis, a data set which could more easily be incorporated into an integrated system will score higher than one that is completely stand alone.

7.4.2 Outline functional specification for GRSGIS

General: Considering that the project resources are not unlimited and more importantly, that time is limited, it is proposed that the first version of this software should be developed for use on a standalone PC or Local Area Network (LAN). The front-end system should operate on a Windows 95 or NT platform and if on a network, then the database back-end should be capable of residing on NT Advanced Server Version 4 and Novell 3.12 or better. It is also proposed that at this stage MS Access is a viable and cost-effective option for the database since the volume of data is relatively small in most cases or is held as compact data objects.

Features & Functions: In the interest of keeping the interface consistent between the two modules (this and the RDI), it is proposed that much of the end user functionality of the RDI will be included in the GRSGIS. The key exception is that GIS mapping is not proposed for inclusion in the system. It is proposed that although the RDI system will probably be built using Java and the GRSGIS with some other language, not yet specified, both user interfaces should be very similar. It will also be important that the GRSGIS system can export data.

7.4.3 The RSG Web Page

As mentioned in section 5.5, it is widely recognised that the World Wide Web is becoming a very cost effective medium for finding and disseminating information. Consequently it is proposed that as the third component of this information system, a supporting web page be established. While it is accepted that the web content should be agreed between a steering group made up of key experts from the relevant organisations, the current RSG page is modified to contain the following additional pages;

1. GPIPIS Introduction
2. PDI Database Application
3. GPIPIS Data Sets
4. Publications / Standards etc.
5. Links to other sites

It is expected that this page will have modest content for the start-up phase.

7.5 *IRSGS Implementation*

7.5.1 Introduction

The implementation plan for the RSG project will require nine months to complete. The principal activity in the early stage is the development of the technical specifications, followed by data gathering and system development. It is considered crucial that a strong team leader is appointed and secondly that the consultation with end users is developed and maintained.

7.5.2 Developing and Hosting System

Each of the system's three components (the RDI, GRSGIS and the Web Page) will require a different level of expertise and resources to maintain. From a technical point of view hosting a RSG world wide web page, which is the simplest component of the three, should not be a problem for most organisations. However, the task of maintaining and further developing the RDI and the GRSGIS databases will require specialist IT expertise.

Based on consultations and considering that an organisation with multi-sectoral interest to be more appropriate, it would appear that the Irish Marine Data Centre, CSA or Marine Informatics are those most likely to successfully host and manage the system. While there are many organisations with technical IT skills, each of those mentioned above is already familiar with the RSG and all have a strong IT Unit.

There is however the question of long-term ownership and after the RSG has ended and even if the system is developed and implemented by a commercial organisation, it is proposed that it should eventually be transferred to a public body such as the Irish Marine Data Centre. Using a commercial company(s) for development and implementation followed by transfer to public body has the advantage that during the development phase, the burden on a public body such as the IMDC or PAD associated is reduced and all that remains is the task of ensuring that the overall quality is maintained.

The decision on which organisation or organisations should be responsible for the system will require further discussion between the various interested parties.

7.5.3 Schedule

It is important not to underestimate the task involved in establishing the RSG database. In addition to the time required to develop the software system the process will require time to achieve active support from many individuals, it will need time for an iterative, detailed, technical design process. It will involve training and it will require time for pre-operational testing. While the availability of resources and the time taken for implementation are related, there is a limit to the rate at which a quality solution can be produced. Any proposed solution must therefore consider this as a key issue as it would appear that the RSG will end in 2001 and therefore it is proposed that there only be a six to nine month time period for the pre-implementation development. It is proposed that the following time frame is achievable:

Timescale in Months

0 – 1	Preparation of the RSG Data Handbook
0 – 1	Technical specifications for the RDI software
0 – 3	Technical specification for the GRSGIS software
0 – 6	Initial Data Collection for the RDI
0 – 6	Development of RDI software and implementation
0 – 9	Development of GRSGIS System
0 – 8	Initial Data preparations for GRSGIS

(0 = project kick off)

Table 15 Timescale for Recommended Options

7.5.4 Funding

Although provision has been made under the technical assistance measure of the PIP, it is important that one realises that when considering the cost of a RSG database, there is more involved than just the cost paid to produce the software system. The total cost is better considered as the combined pre and post operational cost. The pre operational cost includes the technical design process, development, the cost of training and initial implementation. The post-operational cost includes the cost of ongoing system maintenance and support. The proposed solution will consider the total cost and make recommendations to maximise value for investment.

The budget cost (excluding VAT) of the entire system proposed is summarised as follows and elaborated below:

Development of the Data Handbook	£10 - £12K
Development of the RDI software	£30 – £35K
Development of the GRSGIS	£65 - £85K
Initial Data Collection and processing for the RDI	£15 - £20K
Initial Data preparations for GRSGIS	£25 - £30K

All prices in Irish Pounds

Total Cost £145 – £187K

Table 16 Recommended Option Costing

It is of course possible that only certain items from the above list are selected for implementation. For example the RSG may decide to implement the Data Handbook, the RDI and the Web. The cost of this including development and implementation would be between £55,000 and £67,000.

The development of the RDI involves 3 stages. Firstly, the development of the technical specifications which includes detailed database design, user interface design, and communications protocol design. The second stage involves the software coding, testing and implementation and the third stage involves system documentation. It is estimated that in total this development work will require an experienced expert development team for approximately four full time equivalent (FTE) months to complete or perhaps six months for a less experienced team. Taking into consideration current competitive market rates for this type of development a budget of £30 - £35K is proposed.

The development process for the GRSGIS is similar in structure to the RDI. There are however two significant differences. Firstly, this application will be significantly larger in size than the RDI and secondly it is not proposed to develop this for the internet and so development productivity is higher. It is estimated that this component of the IRSGS will take more than twice as much effort to develop as the RDI and therefore cost in the region of £65 -£85K.

The population of the RDI requires a data inventory population from RSG data sources. Experience has shown that when launching an inventory such as that proposed, it is wiser to actively interact with those providing the information and therefore face-to-face contact is recommended. The cost estimated above included approximately two months of data gathering on-site at key locations and one month of data entry into the DTI system. The cost of staff to carry out this work is less than that for software development but this task does have the overhead associated with site visits. These site visits also

serve as an opportunity to market and gain support for the project. The estimated budget for this component of the work is £15-£20K. The data gathering for the GRSGIS includes the compilation of fey data already in an electronic form. The digitizing and inclusion of raw, high volume data is likely to be significant and would be costed separately depending on the volume and nature of the data.

In order to develop the system to its maximum capacity and efficiency, two further options might be considered.

1. Upsizing the GRSGIS database: In order to exploit the data to the maximum possibility, it is recommended that this database be upsized to a Client-Server RDBMS such as MS SQLServer. This will vastly improve speed and scalability and will enable a much tighter implementation of Data Security. The cost is estimated in the region of £30K - £50K.

2. WEB - Enabling the GRSGIS. This will enable anyone with access to the Internet to gain view the GRSGIS data. The viability of WEB-enabling the GRSGIS is dependent on the types and volumes of data, given that speed of download is a key factor. However, we would like to point to Microsoft's implementation of TerraServer (<http://terraserver.microsoft.com/>), as an example of a High Volume Web Database system. The cost is estimated in the region of £40K - £60K.

If it transpires that the overall cost of the IRSGS system is prohibitive and cannot be allocated at one time for complete implementation, then any of the three components maybe developed in phases over a longer period of time. If this approach is adopted, it is recommended that priority is given to the RDI followed by the Web page and finally the GRSGIS. The reason for prioritising the RDI is that its database will act as a foundation for the development of the GRSGIS databases, it will involve most collectors and users of transport data, it will be readily accessible over the internet and it costs less than the GRSGIS. It is also likely that if the GRSGIS is developed in advance of the RDI, it may cost more than proposed above as the developer will have to design and implement a new top level of the GRSGIS which, as proposed above, is based on the RDI. Until the RDI or GRSGIS is developed there is little point in developing the Web Page.

7.5.5 The IRSGS Team and their functions

The implementation and operation of the system should be directed by a person with a good knowledge of information technology and RSG data. Additionally this person should have the ability to organise and motivate the various individuals at the agency and academic organisations that will need to participate in the project. It is envisaged that the team with responsibility for establishing the system will comprise a small expert group (probably 2 or 3) dedicated to the project, with input, some of it substantial, from those individuals already responsible for data compilation at the various RSG related organisations.

Once the system has been established in its initial form, it is likely that the core team could be reduced to a single expert. It is recommended that this expert, if not the same as above, should be expert in IT and have a good knowledge of RSG statistics.

8 Summary

The diversity and number of project teams involved in the Rockall Studies Group entails that there is a requirement for a data management solution to the distribution and sharing of data from the project teams. Any solution must address the issues of audience, functionality requirements, speed of data access, updateability and performance issues.

The solution proposed is a staged solution (IRSGS), allowing the RSG management committee to opt for a data management solution that builds up over time. The first stage is a Data Handbook which is strongly recommended to provide guidelines for deliverables and sharing of data. The second stage is a web based index system (RDI) which will allow users to be directed to various data sources. The third stage is the development of a Windows based system (GRSGIS) which will allow a similar interface to RDI to browse and access data from a collection of CD-Roms.

REFERENCES

- ALLEN, G. S. 1995, *CD-ROM and its application to the petroleum industry*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- ALGAN U. 1998, *Exploration data management beyond the millenium*, PETEX 98, Expanded Lecture Abstracts
- BLAKE N. et al. 1998, *Exploration data management made easy*, PETEX 98, Expanded Lecture Abstracts
- BOWIE, R.C. 1995, *Data Management in the National Geological Records Centre*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- CHEILEACHAIR, O. 1994, *Project data management is a value adding activity – The EDAP Experience*, Paper produced by the Irish Marine Data Centre.
- CHEW, K. J. 1995, *Data modelling a general purpose petroleum geological database*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- DULLER, R. P. 1995. *The quality assurance of geological data*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- ELEANOR, J. 1998, *The future is random*, PETEX 98, Expanded Lecture Abstracts.
- HENLEY, S. 1995, *Project Databases: standards and security*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- GILES, J. R. 1995, *The what, why, when, how, where and who of geological data management*. From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- GRAHAM A. et al. 1998, *Delivering highest quality data to the end user*, PETEX 98, Expanded Lecture Abstracts
- IRISH MARINE DATA CENTER 1994, *A guideline document on electronic data publishing for MAST research projects*, Document published by the Irish Marine Data Center.
- LOWE, D. J. 1995, *The geological data manager: an expanding role to fill a rapidly growing need*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- MCNAUGHTON N. 1995, *A survey of data management in the E&P business*, published by the Data Room
- RASMUSSEN, K. 1995, *An overview of database analysis and the design for geological systems*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.
- RING M. J. 1998, *Harmonising People, Technology and Process in the Exploration & Producing Industry*, PETEX 98, Expanded Lecture Abstracts
- SAUNDERS R. M. et al 1995, *Improving the value of geological data: a standardized model for the industry*, From Giles, J.R.A. (ed.) 1995, Geological Data Management, Geological Society Special Publication No 97.

9 Appendix 1 – Questionnaires

1. New Data Acquisition Questionnaire
2. Data Interpretation Questionnaire
3. Guideline for Interview of Data Compilation and Metadata Projects Questionnaire
4. Guideline to Interview with PIP RSG members Questionnaire

10 Appendix 2 – British Oceanographic Data Center (BODC) Data Dictionary Example

11 Appendix 3 – Metadata Form Examples

1. EDMED Metadata Form
2. RIEDS (Register of Irish Environmental Data Sources Form