

MERMAID

EU Project No: IST-1999-10637

Final User Requirements Specification

Deliverable Nos: D1.1, D1.2, D1.3, D1.4, D1.5
(Rev 0)

August 2000



Report Title:	Mermaid Final User Requirements Specification
Customer:	European Commission, Directorate-General Information Society, IST Programme
BMT Report no:	13300/D/02
Deliverable nos:	D1.1, D1.2, D1.3, D1.4, D1.5
Report status:	Rev 0
Date:	03/08/2000
Contact details:	BMT MARINE INFORMATION SYSTEMS LTD Grove House, 7 Ocean Way, Ocean Village, Southampton, Hampshire, SO14 3TJ. United Kingdom. Tel: +44 (0) 2380 232222 Fax: +44 (0) 2380 232891 e-mail: mis@bmtmis.demon.co.uk Website: http://www.bmtmis.com

	Name	Signature	Date
Author:	Paul Taylor		
Approved by:	Dr Andrew Tyler		

This report is commercial-in-confidence. Any information contained herein should not be communicated to any third party without prior permission of BMT Marine Information Systems Ltd.

Contributors

Company	Name
BMT	Dr. Andrew Tyler
	Paul Goddard
TXT	Matteo Villa
CEDRE	Camille Lecat
Met. Office	Karl Kitchen
	Chris Little
	Joena Peulevé
IMGW-OM	Wlodek Kryzminski
NC	Koen DeHulsters

Distribution List

Company	Name	Number of Copies
EU Commission (DG-XIII)	Jean-Jacques Lauture	5
BMT	Paul Taylor	1
TXT	Matteo Villa	1
CEDRE	Camille Lecat	1
Met. Office	Colin Gray	1
IMGW-OM	Wlodek Kryzminski	1
NC	Koen DeHulsters	1
MCED	Ellik Adler	1



CONTENTS

1	INTRODUCTION	2
1.1	Project Summary	2
1.2	Project Objectives	2
1.3	Project Requirements Analysis	2
2	SECTION A: USER REQUIREMENTS	2
2.1	Introduction	2
2.2	User Requirements Conclusions	2
2.2.1	General	2
2.2.2	Data Holdings	2
2.2.3	Data Formats	2
2.2.4	Data Supply	2
2.2.5	Quality	2
2.2.6	Security	2
2.2.7	Pricing	2
2.2.8	Payment Methods	2
2.2.9	Administration Fee	2
3	SECTION B – USER REQUIREMENTS ANALYSIS	2
3.1	Introduction	2
3.2	Graphical User Interface	2
3.2.1	Home Page	2
3.2.2	Registration Page	2
3.2.3	Metadata Catalog Page	2
3.2.4	Search/Extract/Request Page	2
3.2.5	Pricing Page	2
3.2.6	Purchasing Page	2
3.2.7	Broker Administration Page	2
3.3	Data Structures	2
3.3.1	Metadata Catalog	2
3.3.2	Data Formats	2
3.4	Data Warehouse	2
3.4.1	Delivery Mechanisms	2
3.5	E-commerce	2
3.5.1	Terms and Conditions	2
3.5.2	Security	2
3.5.3	Pricing	2
3.5.4	Payment Methods	2
3.5.5	Broker Administration Fee	2
4	EVALUATION PLAN	2
5	MOCK-UP APPLICATION	2

APPENDICES

1 INTRODUCTION

The Marine Environmental Response data Management and Acquisition using Internet Data brokerage (MERMAID) project, is a 30-month project with seven partners and a budget of over £1.2 million. The funding for the project is being provided by the European Union under the new Framework V Information Society Technologies Programme. The central objective is to provide a seamless, minimum intervention link (Data Broker) via the Internet, to allow end users to rapidly access and use large, distributed marine environmental datasets. The project began in January 2000, and will be completed by June 2002.

1.1 Project Summary

Over the past decade numerous national, European and international programmes have focused on the generation of data on the natural environment. Very little of this wealth of data is actually utilised by the ultimate end users who provide services to the citizen, for example in support of maritime emergencies (marine pollution, search and rescue *etc.*). The central objective is therefore to provide a seamless, minimum intervention link (the Data Broker) to allow end users to rapidly access and use large distributed environmental datasets. Through the development of an Internet-based Data Broker capable of cataloguing, storing/referencing and accessing these datasets, the user will be able to search for, choose, purchase and download data subsets for their specific and immediate data requirements. Using the latest Internet and database tools and complying with international data standards, the Data Broker technology will be designed as an 'open to all' service for data providers and users. The marine emergency application domain has been chosen to demonstrate the technology due to its demanding requirements for timely data. However, the technology is widely applicable and the Broker will facilitate a very wide range of applications requiring marine environmental data

1.2 Project Objectives

The overall objective is the development of a seamless, minimum intervention link to allow end users working in the marine environmental emergency application domain to access and use large distributed datasets of environmental parameters.

The primary user objective is the development of a major virtual shopping centre for environmental data providers and users providing near real-time user access to major international datasets with inherent support at user sites for sophisticated end user applications, and simple web browser-based data reader/viewer applications.

The key technological objectives are the development of web-enabled neutral formats for environmental data transmission and exchange based on existing

standards and improved for temporal knowledge representation and management, combined with the development of web-enabled methods (incorporating E-commerce) to search, extract, compress and transmit the variety of data types routinely encountered.

1.3 Project Requirements Analysis

The first phase of the project constituted the research and analysis of the principal user requirements. This document outlines the user requirement specifications for the development of the project. The document is divided into two sections. Section A details the objective user requirements, concluded from the extensive external user questionnaire survey, the first user workshop and from discussions with the internal user partners. Section B then provides the analysis of these user requirements, which specify how the proposed system will meet the requirements.

It became apparent that there will always be some extremes in what users require, and that catering for all of these would result in an unusable, over-complicated end product. The aim of the project is to develop a realistic, viable and user-friendly service, and therefore the analysis of the user needs has resulted in a feasible and flexible specification of user requirements.

The project is seen as having the potential to develop a fully viable, commercial end product. This service could easily be developed within an almost unlimited scope. However, in order to keep the project manageable, the complexity of the development will be kept within fixed goals, defined by the marine emergency response domain, for which the project is initially targeted. However, the service will potentially be applicable on a much wider level, and therefore, where possible, the system will be designed with a flexible approach, in terms of technology, in order to allow the possible further development of the service in the future.

2 SECTION A: USER REQUIREMENTS

2.1 Introduction

These user requirements have been concluded objectively, from the external user questionnaire survey, from the following user workshop, and from discussion with the internal user partners. The response from external parties to the questionnaire survey and to the workshop was excellent, and better than expected.

It should be noted here that the term 'user' refers to all potential users of the service, rather than simply users of data. This therefore includes both data providers and data consumers. It also includes the final administrator of the broker, who will, by necessity, have a set of requirements in order to maintain and administer the service efficiently once in operation.

The external questionnaire survey was extensive, and more than 120 organisations from across the globe were sent the questionnaire directly. These organisations were selected individually by each partner, and therefore form a wide-ranging cross-section from the marine and environmental domains. The majority of responders were commercial organisations, although there were responders from government bodies (both regulatory and research), non-profit making organisations and academic institutions. To date, seventy organisations have returned a completed questionnaire.

The questionnaire was designed to be generic and objective, and is available from the project web site. It will remain on-line throughout the duration of the project, in order to collect a larger dataset, and to keep a check on any changing requirements. A copy of the questionnaire is included in Appendix 1 of this document, together with a list of those organisations that have returned the questionnaire. The analysis of questionnaire results is also provided in this appendix.

The first user workshop followed up on the questionnaire, and was held at the UK Meteorological Office (Met Office) college, in Reading. Representatives from ten major, market-leading organisations, both data providers and consumers, attended the workshop. In addition, all of the project partners were represented. The workshop was based on two parallel discussion sessions. Providers and consumers were separated, and each group's discussion was based around the same key issues. These were titled data formats; data types; procurement; and data ownership. The agenda and minutes of the workshop are provided in appendix 2.

This document emphasises the needs of the external users, and these are seen as crucial to the success of the final product. However, it should be noted that the inclusion of the internal partner's user requirements is also of great importance to the development of the project. Each internal partner completed an 'external' questionnaire, which were included in the analysis, and were also involved in the user workshop discussions. In addition, each project partner has completed a much more detailed and specific 'internal'

questionnaire, and was involved in continual discussions during the evolution of this user requirements specification. The internal partner's questionnaire is included in appendix 3, together with the completed questionnaire from each partner.

By necessity, many issues have been decided in principal at this stage, and final decisions on details, which will be made during the design phase during the next eight months, will be made in close consultation with all the project partners.

Although the requirements of providers and consumers were collected and analysed separately, they are presented here together, logically grouped into different categories for easy comparison. The results have been summarised here but the detailed statistical analysis is provided in appendix 1.

2.2 User Requirements Conclusions

2.2.1 General

It was clearly evident that the Internet is an important tool for all organisations, and almost all providers and consumers that responded currently have access to the Internet. The majority of organisations consider it to be business critical, and almost every organisation (both providers and consumers) is interested in using the Internet to either distribute or acquire their data on-line.

There is a distinction that must be made between datasets produced on a one-off or irregular basis, and data 'streams' produced at a regular frequency, such as daily weather forecasts. Consumers often require archived datasets as well as current or forecast datasets and data streams on a regular basis for a period of time. The inherent differences between datasets and data streams will lead to differences in the way these will need to be handled by the broker.

Many providers need to raise their profile and that of their datasets and wish to increase the market awareness of their products. Equally, consumers often require a quick and easy reference of what datasets are available, and how to acquire them.

2.2.2 Data Holdings

Size

The size of datasets produced by providers range fairly evenly from less than 10 Kb to more than 10 Gb. This size distribution spectrum quite logically matches the range of dataset sizes required by consumers. However, although some consumers currently acquire datasets that are greater than 10 Gb, often this is only because a smaller dataset is not available.

In addition to the range in the size of datasets, the data produced varies enormously, from varying in four-dimensions (spatially and temporally) on a regular grid, to irregular, non-geo referenced discrete samples. However, the majority of datasets produced are regular or irregular gridded data that vary in both space and time. This type of data can be expressed either explicitly (*i.e.* a value for position, time, and the parameter for each point is given) or implicitly (*i.e.* an equation is provided to define the parameter at each point in time and space, together with the structure and boundaries of the grid). Implicit data is much more space-efficient, but requires knowledge of the algorithm to subsequently calculate the data.

Storage

There are many providers who would not be happy letting a third party hold their data for them without knowledge and assurance of the security mechanisms that would be used to protect the data. However, there are also

many providers (more than half) who would readily welcome a data warehouse that could hold their data.

Metadata

There is no single industry-standard metadata format that is used to describe datasets and in fact providers currently use a variety of metadata standards. These vary depending on the provider and the nature of the data. Providers like to give a varying degree of detail in the descriptions of their data. Some providers only wish to give simple information about their data, while others wish to provide complex descriptions.

Some organisations wish to be able to provide their data descriptions and their data, or to search for datasets, in languages other than English.

Many consumers would like the ability to preview (to some extent) the dataset before they purchase it, rather than merely relying on the metadata description. In particular, simple maps to graphically show the location of the data would be welcomed.

2.2.3 Data Formats

There are no universal, industry-standard formats currently in general use, although there may be standards used for particular sub-sets of data (such as the World Meteorological Organisation). Usually, the format of a particular dataset will be dependent on the nature and scale of the data, the method of production, and the ultimate application. Marine environmental data is produced in both ASCII¹ and binary formats and in certain situations it would not be possible to change the format currently used. Although the majority of users (both providers and consumers) would be prepared to re-format their data, some users do not have the capability of dealing with a different format to the one they are used to.

Most consumers would prefer not to have to re-format their data after acquisition. However, there are a significant number of consumers who actually prefer to re-format the data themselves.

There is therefore a plethora of data formats currently used within the marine environmental domain, which includes non-geo-referenced data such as bibliographic databases, and also non-digitised data, such as paper charts or geophysical core samples. All of these formats are of importance to users.

¹ There is no longer a definitive ASCII code, and in fact there are several variations of it, using different character sets, and either 7, 8 or even 16 bits. These variations cater for non-English characters, such as accented letters (e.g. ê, ø, and å). This distinction was not made in the questionnaire, or at the workshop.

2.2.4 Data Supply

The majority of users either supply, or acquire their data on an ad-hoc basis. However a significant number of consumers require data on a regular basis, either hourly, daily, weekly or monthly. These are often data streams, although it may also be discrete datasets that are required regularly (such as yearly oceanographic cruise data).

The timeliness of the data requirement (*i.e.* how quickly the data is needed by consumers) varies enormously from within an hour to within one month or more. This depends on the circumstances, and the nature of the dataset. In most emergency response situations, some data is required immediately (such as weather conditions) although other data will continue to be required less urgently, for a period of weeks or even months after the incident. This implies that an incident that may be months old, may still be deemed a current event, and that users may be interested in a varying period when talking about the 'current' time.

Data is provided to consumers by a variety of delivery mechanisms, and on a variety of media. The majority of providers and consumers use CD-ROMs, floppy disks, FTP or e-mail for the transfer of data, although traditional hard-copies, sent by post or courier are still useful methods.

2.2.5 Quality

Most providers produce their data with some form of quality assurance (Q.A), and this is equally important for consumers. The level of Q.A. varies between providers and datasets, but it is important that providers can acknowledge the Q.A, accuracy and control procedures of the data they supply. Providers often give a qualitative description of the accuracy of the data in the description, as well as quantitative information with the actual data itself. This would be apparent in the format of the data. It is imperative that consumers know the level of the quality control and accuracy before purchasing the data, to be sure the data is 'fit for purpose'.

It is important for providers that the consumers of their data apply and interpret it appropriately. They therefore need to be able to provide advice and guidance on its use.

2.2.6 Security

Security is of utmost importance to data providers, and many have concerns over the security of their data if handled by a data broker. The ownership of the data must be protected, and unauthorised third parties must not have access to the data. The transfer mechanisms must therefore be secure. There is also a fear that 'Internet theft' may not be considered a crime by some people, which may lead to an increased risk of theft. However, this concern is continually being reduced as the Internet becomes more and more essential and familiar.

Confidentiality must be kept, and in particular metadata descriptions and the details of who has searched for, or purchased specific datasets must be protected. However, providers will need to know the details of all the consumers who have purchased their data, although the consumer's details must not be accessible to any party other than the providers who they have purchased data from.

The liability and jurisdiction must be clearly known by all parties involved (both providers and consumers, as well as the broker). There are obviously legal differences between countries, and these issues must be considered. The terms and conditions of purchase must therefore be clear and simple, and universally applicable. Consumers will not welcome lots of different, complicated terms and conditions for different datasets, which would all require verification and agreement before any purchase could be made.

2.2.7 Pricing

Many providers have complex and flexible pricing structures, with different agreements for different consumers. Therefore a particular dataset may be sold at different rates to different consumers, depending on the provider-consumer relationship. Some consumers, such as educational institutions, are even given data at no cost at all. In addition, provision must be made to allow publicly funded data to be freely available to the public.

2.2.8 Payment Methods

There are a variety of payment methods currently used by providers and consumers, of which credit cards are increasingly becoming more important. However, it is felt that a purely credit-card based system would not be appropriate. Any kind of voucher scheme is not thought to be beneficial in this situation. Consumers are on the whole open to the idea of an on-line, direct debit scheme for payment, and in addition, some providers would be prepared to offer credit on account, to certain consumers, depending on their relationship.

2.2.9 Administration Fee

Many users are not initially happy to pay an administration fee to a third party data broker. For those that are prepared to pay a fee, most would prefer this to be on a percentage-basis. However, this fee should not be prohibitive. A subscription fee for the service would not be unwelcome, and some providers would be happy to pay a rental charge for the storage of their data by the broker.

3 SECTION B – USER REQUIREMENTS ANALYSIS

3.1 Introduction

The Mermaid Data Broker will principally provide the facility for users to search for, extract and order datasets, and subsets of datasets within the marine environmental domain. The types of data that this domain encompasses are likely to include; marine meteorology; physical, chemical and biological oceanography; marine biology, aquaculture, and fisheries; environmental quality monitoring; marine geology and geophysics, and coastal management and engineering. It should be noted that this list is not exhaustive.

The extraction, purchase and transfer of the dataset, together with the subsequent invoicing, will be automated by the broker. In addition, it will also provide a simple catalog of data that is available, giving references and details of the data source, which will help raise the market awareness of the products that are available.

The Broker will therefore consist of four elements; the Graphical User Interface (GUI), the Metadata Catalog, the Data Warehouse, and the E-commerce Engine.

Data providers and consumers who wish to use the broker will need to initially register with the broker. In order for the broker to be effectively administered, it will be necessary for the broker to perform credit checks when appropriate, issue statements and update records when maintaining the service. Each dataset and data stream will have to be registered separately, with an individual description of the dataset/data stream for inclusion in the Metadata Catalog. A data stream is a dataset that is produced at a regular frequency, such as a daily weather forecast. The structure of the data stream will not change, and therefore a single metadata description will define it. However, the request for and purchase of a data stream will have to be handled in a slightly different way to a normal dataset.

Registering with the broker will NOT make all of the data that is owned by the provider accessible.

The service will be as unrestrictive as possible. The philosophy that will be adopted is that the service should, as far as possible, be market-driven. For this reason, the aim is to be flexible and open, wherever possible. For instance, there will not be limits set on the type, size, or format of the datasets that can be registered and it will be possible to provide metadata descriptions in any language. The choice will therefore be left to the provider, and ultimately this will be determined by what is marketable. However, in certain areas, total flexibility will not be practicable, and it will be necessary to be prescriptive. The metadata format that providers *must* use is such an example.

3.2 Graphical User Interface

The Graphical User Interface (GUI) will allow access to the service, via a web site on the Internet. This site will consist of several different areas, or pages. The design of the site will take place during the next phase of the project.

The key information on each page of the site will be provided in all of the partners home languages (English, French, Italian, Dutch and Polish), which the user will be able to select. This list may be extended to include other major European languages such as Spanish and German, although this will be decided during the design phase.

It will also be possible to enter information such as metadata descriptions in any language. However, the site will not support the automatic translation of all text (*i.e.* immediate translation 'on the fly'). This is not practicable since the automatic translation packages currently available are not very reliable. Any errors made in the translations of legally-binding information (such as metadata descriptions) may well lead to legal problems. All information entered by users will therefore remain exactly as entered during registration, in whatever language, and will not be translated. It will therefore be necessary to explicitly define the ASCII code, and the character sets that will be supported.

3.2.1 Home Page

This will provide new users with an introduction to the site and the service, explaining how the service operates, and providing links to all other areas of the site.

3.2.2 Registration Page

Both data providers and consumers new to the service will have to register with the broker through an on-line registration form. The form will consist of mandatory fields, such as organisation name, address and contact, which will have to be completed by the new user before registration is completed.

3.2.3 Metadata Catalog Page

A registered provider will be able to register new datasets with the catalog, or update existing metadata records on-line. Each provider will have their own password-protected area within the site, and they will therefore only be able to access and modify their own data. Details of those consumers who have purchased data from them will be available at all times through the GUI.

Each provider will only have knowledge of the consumers who have purchased their data.

Any visitor to the site will be able to browse the metadata catalog through the GUI, in order to get familiar with the service before registering, and see what sort of data is available. It is felt that this will encourage more users to register with the service. However only registered users will be able to search for and purchase any data. The consumers will have details of the providers of the data available through the Broker, but the confidentiality of all other consumers will be maintained.

3.2.4 Search/Extract/Request Page

The GUI will also provide the interface for registered consumers who wish to search for, extract and order data. The search criteria will be input on-line by the user. The user will be able to search by a number of different fields, such as time, location, and data type. These searchable fields will be decided during the design phase of the search engine. It will be possible to leave any of these fields blank, which will be interpreted as a 'wild card'. For instance, if a user only enters details for a location and leaves all other fields blank, then all datasets available for that location, regardless of type, format, nature, size or time will be presented.

The results of the search will be displayed through the GUI.

The user will be presented with a list of all the datasets that fit the search criteria, with details of the provider, metadata descriptions (including Q.A. procedures, data type, format, size, etc.), the price, and the delivery mechanisms that are available for that dataset.

If any part of a dataset lies within any part of the search location, as shown in figure 1, then this will be included in the results.

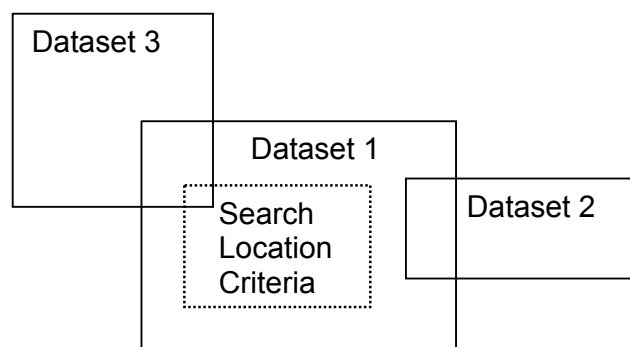


Figure 1. For a search area shown by the dotted line, datasets 1 and 3 would be presented in the search results. Dataset 2 would not be listed.

The search request and results will remain confidential at all times.

The metadata catalog will be supported by a graphical world map on which the user will be able to zoom. The location of selected datasets in the catalog

and the search area will be shown graphically on the map. The search location co-ordinates will be displayed as text, which the user will be able to enter manually.

It will also be possible to preview some datasets, if such a preview has been supplied by the provider (see section 3.4).

In addition, example 'icons' will be designed to represent each of the major data types displayed with the metadata record. The list of data types that will be represented in this way will be decided during the next phase of the project.

The consumer will be able to order the data through the GUI, or may choose to contact the provider directly. Links to the provider's site will be established.

3.2.5 Pricing Page

A three-dimensional pricing database will be accessible through the GUI to enable providers to define their specific pricing structure (see section 3.5.3). Consumers will also be able to view some of this information, such as the individual provider's pricing policy for data sub-sets, although they will not be able to edit it.

3.2.6 Purchasing Page

The consumer will be able to purchase the dataset they have selected automatically. The consumer will have to choose the preferred payment method from a list of available options (see section 3.5.4), and the payment and invoicing will be handled by the broker. The terms and conditions of purchase will be available on-line. These will either be the standard set, or the provider's customised set (see section 3.5.1).

There will be no constraints on the consumer once they have purchased the data with regard to how they use it, other than those set out in the terms and conditions. The provider will be able to provide advisory guidance for the use and interpretation of the data in the metadata description. For data that requires special agreements, consumers will have to agree these directly with the provider.

3.2.7 Broker Administration Page

Only the broker administrator will have access to this page, which will allow the broker to effectively maintain the service. The full design of this page will take place during the next phase.

3.3 Data Structures

3.3.1 Metadata Catalog

The metadata catalog will provide a list of all datasets that have been registered with the broker. Each dataset will therefore have its own record, providing all of the necessary information about the data.

Catalog Structure

The metadata catalog will provide as much descriptive information about the dataset that the provider wishes to give. Therefore, the metadata structure will be hierarchical or 'treed', presented on-line through a wizard-style series of input boxes. The structure will be divided into two main sections; one for the providers details, and one for the details of the dataset. There will be mandatory and optional fields to be completed in both sections. The mandatory fields will consist of essential basic information, such as the providers name and address, company registration number and contact details for the first section, and the time, date, location, parameters measured and units used for the dataset section. More complex details such as Q.A procedures adopted, accuracies, measuring instruments, reference ellipsoids, map projections used, *etc.* will be entered through higher levels of the 'tree', and will be optional. If any mandatory information is missing, the record will be rejected, and the provider warned of the missing information.

The provider will only have to complete the first section once, during initial registration. Each dataset registered thereafter will be associated with the provider's core details. The data section will be further subdivided into several categories, for the different types of data, with the appropriate fields for each type. There are likely to be up to about 5 main types (*i.e.* implicitly gridded, explicitly gridded, point data, text and images). The categories will be decided during the design phase.

Metadata Records

There will therefore be a record for each provider. There will also be a record for each individual dataset and data stream. Therefore every dataset will have to be provided with a metadata description during registration with the catalog. The format of this description will be set. Amongst the mandatory information, each record will require a reference time and location which will help in the search of datasets. The reference time will be either an explicit time, for datasets, or an implied time (*i.e.* current time) for data streams. All times will be Y2K compliant, and referenced to the beginning of the Christian era (*i.e.* 0000 A.D.). The location will be a geographical reference for the data. This will be in geographic co-ordinates (latitude and longitude) although the service may provide a number of other co-ordinate / projection systems if it is felt that this is necessary. This will require further research, and will be decided upon during the next phase.

The location will be given in one of two ways; either as a regular area defined by the top left and bottom right co-ordinates, or as a point location, with a region of interest in kilometres. This will be applicable even for non-georeferenced datasets, such as bibliographies, and it will be possible to define the whole world as the geographic location of the data. The limits will be set as -90° to 90° for latitude (with 0° as the equator) and -180° to 180° for longitude (with 0° as the Greenwich meridian) for the standard co-ordinate system. It will be possible to either manually enter the location definitions, or to draw them graphically on a map interface.

There will also be a finite list of irregular areas (such as continents, countries, zones, oceans and major seas) that will be selectable by the user as the reference location. These will be pre-defined by bounding co-ordinates that are likely to form rectangular boundaries around each area. The number of areas together with their bounding co-ordinates will be decided on during the next phase, but are likely to be taken from an existing list (e.g. those defined by GCMD or EDMED – see below).

Some of the fields will be selectable pick-lists (such as delivery mechanism), but many will be unlimited free text fields (such as general description of the data). This will allow a fully flexible level of complexity of information to be provided.

The metadata format must have the capacity for providers to adequately define the Q.A. associated with the data both qualitatively in the notes and quantitatively with the actual data, as well as any guidance notes on the use of data. The metadata descriptions will be legally binding, and may form part of the terms and conditions of purchase.

The onus will be on the provider to give advisory notes on the ‘fitness for purpose’ of the data, but the consumer will not be constrained in its use. The responsibility will then lay with the consumer to ensure the data is used appropriately.

It will be possible for the provider to prepare metadata record files offline, and then upload these to the metadata catalog for inclusion, provided that these files are in the appropriate, pre-determined format. This is in order to facilitate those providers who may have numerous datasets they wish to register or update at one time.

Metadata Format

A single existing metadata standard is unlikely to be appropriate for all the different types of data that may be registered. Therefore, a simple, existing metadata standard is likely to form the basis of the format, which will then be customised for the purpose of the project. The detailed design of the structure, and the mandatory and optional fields will be completed during the next design phase. There are a number of similar projects that are currently being developed under the EU IST Framework for which the development of a metadata catalog is fundamental. These include THETIS (and AVID),

COASTBASE, IWICOS and EDEN-IW. The metadata format that will be adopted for MERMAID is likely therefore to be developed in conjunction with these other projects, and the groundwork for this has already been made through contact via the first concertation meeting in Brussels.

The metadata format adopted is likely to be compatible with Nasa's Global Change Master Directory (GCMD). This is a comprehensive directory of descriptions of datasets of relevance to global change research. The GCMD database includes descriptions of datasets covering climate change, agriculture, the atmosphere, biosphere, hydrosphere & oceans, geology, geography, and human dimensions of global change.

Existing Metadata Standards

The existing standards under consideration for MERMAID are GELOS, FGDC, EDMED, MEDI and GML. Brief descriptions are provided below, and references to further information are provided in appendix 4.

The Global Environmental Information Locator Service (GELOS) is a globally distributed virtual library of Environment and Natural Resources Management (ERNM) data and resources. It has developed a European-wide metadata standard for the description, and cataloguing of information.

The Federal Geographic Data Committee (FGDC) co-ordinates the development of the National Spatial Data Infrastructure (NSDI), which encompasses policies, standards, and procedures for organisations to co-operatively produce and share geographic data. The FGDC approved the Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998) in June 1998. This standard defines the data elements that make up the required metadata record for each dataset.

EDMED was developed and coordinated by the British Oceanographic Data Centre (BODC) and was funded by the Marine Science and Technology Programme (MAST) of the European Commission. It is a comprehensive reference to the marine environmental data held within Europe. Datasets are catalogued in EDMED irrespective of their format, and records are held in simple ASCII text format.

The MEDI Pilot Project is a directory of information about marine related datasets, data catalogs and data inventories within the framework of the IOC's International Oceanographic Data and Information Exchange (IODE) system. The MEDI database structure has been based on the Global Change Master Directory (GCMD), and is fully compatible with it.

Geography Markup Language is an XML-based encoding standard for geographic information developed by the OpenGIS Consortium (OGS). It's current status is an RFC (Request For Comment) under review within the OpenGIS Consortium. The RFC is supported by a variety of vendors including Oracle Corporation, Galdos Systems Inc, MapInfo, CubeWerx and Compusult Ltd.

Updating Records

It will not be feasible to incorporate an automated procedure for the update of the catalog by the broker, and it will be the responsibility of the provider to update the metadata description, and the data itself, should these change.

The broker will not be responsible for validating data or metadata descriptions. If the vendor provides erroneous data or incorrect descriptions, then any liability will lie with the vendor.

Searching

The consumer will be able to search the metadata catalog by any combination of a set of specific fields. These will be decided on during the design of the search engine, but are likely to include the following:

- Time (and Period)
- Location
- Format
- Type
- Keyword

It should be noted that 'time' may have a varying influence and validity depending on the dataset. For instance, a weather forecast may be valid for 9 hours from a reference time, or the 'current' time of a map of surface circulation may extend back or forward in time for hours, days or even weeks. It is therefore likely that the user will be able to enter either an explicit time, or to imply a time, by entering 'current' time together with a period of interest (such as plus or minus 48 hours).

A flat map of the world will be presented through the GUI for easy selection of the location of interest. The user will be able to zoom-in on this map. In addition, the user will be able to manually enter the location co-ordinates, which will be identified on the map. The geographic co-ordinate system (latitude and longitude) will be the standard reference system, although it may be deemed feasible, following further research, to support other geographic projections.

3.3.2 Data Formats

There will be no limit to the type, size or format available through the broker. It will be the provider's decision on what data they make available, and in what format. It will also be the provider's decision as to what transmission mechanisms will be available for each dataset. This information will be provided upon registration of the dataset. This will give providers the flexibility to supply the data in the most popular, or suitable format for their consumers. The only restriction will be in the metadata format that the data descriptions will have to be provided in. The format of the data will be referenced in the metadata.

The broker will therefore be able to handle data of any format, which will keep the service as flexible as possible. However, in order to enable the extraction of subsets of data by the broker, proprietary formats will need to be set for the main data types of 'operational' data. It is these 'operational' datasets, which usually cover large geographic areas and time-periods. They are therefore usually large in terms of file size and likely to be those that consumers will want to extract small subsets from. It is envisaged that there will be four or five of such data types, and each of these types will need to have it's own standard format. The types to be supported, and the optimum formats to be used will need to be identified in the next phase. Providers will not be obliged to register their data in these formats unless they wish potential consumers to be able to extract subsets from them.

If specific applications are required in order to use a particular dataset, the development of these applications will be the responsibility of the providers. If they see the benefit of supplying such an application with their data, they will do so. However there will be nothing to stop a third party in purchasing the data, processing and adding-value to it, and then registering the processed data as a new dataset for sale to other consumers.

The broker will not be able to convert or re-format any data. Therefore datasets will only be available in the format as supplied by the providers. If a provider wishes the data to be available in more than one format, they will have to register each format as a separate dataset.

It should be noted that the ASCII standard that the service will adopt will follow that of the www consortium. Presently this is the 8-bit ISO8859 standard.

3.4 Data Warehouse

It will be possible to hold the datasets either at the broker's site, or the data provider's site. However, there will be a maximum size per dataset that the broker warehouse will be able to store. There will be a fixed rate per Mb for the storage of datasets up to this limit. For datasets larger than this, the storage at the warehouse will incur an additional fee, at an increasing rate per Mb. It is envisaged that the majority of data will actually be held by the providers. In such cases, the data access engine for the search and retrieval of data will be hosted by the provider. The broker will also host a data access engine for handling all data held at the warehouse.

The broker will not have the capability of processing any data. It will be possible for the broker to extract a sub-set of data from a larger dataset, based on a set of criteria, if the data is in one of the supported prescriptive formats. It will not however be possible to filter, analyse or further process this data before delivery. The broker will support the facility to display 'thumbnail' images (as *.GIF files) of the dataset if the provider wishes to supply these with the metadata record. It will not be compulsory to supply such an image, although it is envisaged that such a preview is likely to increase the marketability of the dataset.

3.4.1 Delivery Mechanisms

For each media there will be a set limit on the size of data that can be transferred, a typical transfer time and appropriate delivery mechanisms. For instance, the media that are likely to be supported are; CD-ROM, DVD, floppy disk, ZIP disk, JAZ disk, DAT tape, cartridge, FTP, e-mail, and paper copy. For e-mail, the file size limit might be set as 1 Mb, with a transfer time of 2 hours, and the delivery mechanism as the Internet.

The media that will be supported and the limits set will be agreed upon during the next phase of the project.

The provider will have to register the default delivery mechanisms that are available generally. Then for each dataset that is registered, the provider's default media will be associated with the dataset, and offered to potential consumers as selectable options. However, the provider will have the ability to override the defaults for a specific dataset. For instance, a provider may have registered floppy disks as a general media available, but may not wish to provide a 1Gb dataset on 700 floppy disks. Therefore for this dataset, the provider would remove floppy disks as an available mechanism and this would not be listed as a selectable option.

Consumers will therefore be provided with a list of all the transfer options for each dataset together with the typical transfer time. For delivery mechanisms that cannot be conducted on-line, such as the transfer of geophysical core samples, then it will be the provider's responsibility to deliver the requested dataset.

The FTP facility will be more user-friendly and secure than standard FTP facilities. Further details of the design of the automated FTP server will be provided during the next workpackage.

For mechanisms other than FTP, the automated checking of the receipt of datasets by consumers will not be possible. It will therefore be the responsibility of the consumers to check that they have received the correct dataset.

3.5 E-commerce

3.5.1 Terms and Conditions

There will be a single set of mutually acceptable standard Terms and Conditions for all purchases through the service. These will be agreed upon by the main providers and consumers (both internal and external) before completion of the project. Providers will have the option to supply their data under these standard terms and conditions, or to issue their own set of terms and conditions. It will be the provider's responsibility to highlight these differences. For consumers who are not happy with non-standard terms and conditions, they will have to contact the provider directly to negotiate the agreement. The standard Terms and Conditions will be subject to the jurisdiction of the country that hosts the broker web site, and the consumer will be warned by the broker of this fact, and that legal obligations may well vary from country to country.

3.5.2 Security

The broker site will be fully protected through the use of standard security mechanisms such as passwords, firewalls, secure socket layers, and so on. The details of the mechanisms to be adopted will be researched further during the next phase. However, the technological security of all aspects of the site is imperative, and it will be crucial to effectively market this issue in order to install confidence in the broker.

3.5.3 Pricing

In order to facilitate a differential pricing policy, a pricing model will need to be developed as a separate software model within the broker. An automated approach obviously imposes certain limitations, and a completely flexible model is not feasible.

However, a three-dimensional relational pricing database will be developed, which will handle a 'semi-flexible' pricing structure. The database will hold the discount that each consumer is entitled to from each provider, as well as the price ratio of data subsets (if applicable). By default, the discount will be 0%

for all consumers, and the price ratio 100% for all subsets. It will be the responsibility of each provider to edit these values accordingly. The database will be password-protected, and each provider will only have access to their own details. Consumers will only be shown the 'final' price of the dataset they are interested in, after all discounts have been accounted for. The details of the pricing structure of each provider will not be available to any third parties.

Upon initial registration with the broker, a new provider will be asked to select from a finite list, the generic categories of organisations that they may be prepared to offer discounts to. The provider will then be presented with a list of all registered consumers that belong to the categories that have been selected for a possible discount. The provider will have to assign the discount that they will give to each consumer. The discounts assigned will apply to all purchases from that provider. In addition, the provider will need to define a breakdown of subset ranges, based on file size, together with the cost of each range. For instance, a provider may define the subset breakdown in terms of percentage of total file size, as; up to 10%, up to 50%, and up to 100%, with the corresponding price ratios; 50%, 75%, 100%. Therefore, if a consumer who receives a 50% discount from a provider requests a subset that is 40% of the whole dataset (price ratio of 50%), then he will be charged 25% (50% of 50%) of the default price of the whole dataset. The breakdown range of subsets is totally flexible, in that any number of ranges, and any interval between each range would be acceptable. It will also be possible for providers to charge more for a subset if they wish. This gives optimum flexibility to the provider.

New consumers will be asked upon registration to select from the same list which category their organisation belongs to. Their details will be e-mailed to all the providers that have selected that category as being eligible for a possible discount. If a consumer feels they would be entitled to a discount from a particular provider but does not have one, then the consumer will need to apply directly to the provider for this discount.

It will be the responsibility of the provider to verify that the consumer's details are correct, and to update the pricing database if any changes are made.

Each dataset will be assigned a default price by the provider upon registration of that dataset. This is the price that will be charged to potential consumers, unless the consumer is entitled to a discount from that provider.

3.5.4 Payment Methods

There will need to be the following three payment methods available:

- credit card facility
- direct debit from consumers accounts
- one-to-one credit accounts between provider and consumer

The credit card method, which will be adopted after research from an existing secure on-line shop, will be open to any registered user.

The direct debit method will only be open to those users who select this as an option they are prepared to use, upon initial registration. Funds will be automatically debited from the consumer's bank account, and transferred to the provider's account. The users (both providers and consumers) who select this as an option will therefore have to provide their account details and any upper limits on transactions. Consumers will only be able to use this option if the provider who they wish to purchase the data from has selected this as an option, and any limits have not been reached.

The one-to-one account will be mutually agreed upon between the individual providers and consumers, and will therefore vary according to the relationship. The details of the account, such as credit limits, payment methods, and payment timescales will be decided between the provider and consumer. However, the broker will facilitate the establishment of new accounts, and will know the details of the account, in order to monitor the transactions. It will be the responsibility of the provider to keep the broker informed of payments made, and any changes to account details..

Regular statements will be sent to both providers and consumers by the broker.

3.5.5 Broker Administration Fee

There will be three payment methods to fund the service, which will all be payable by users of the service. These will give the maximum flexibility to the service, and ensure that the service is adequately paid for.

- 1) A subscription fee. This will be payable by providers on a monthly basis in order to remain registered with the service. Registered providers will be able to update their existing data, and add new datasets, and registered consumers will be able to browse the catalog.
- 2) A warehouse fee and commission. Providers who want the broker warehouse to store their data would pay 'rent' at a fixed rate per MB. The broker would also earn a commission (at a fixed percentage) on all data purchased by consumers from the warehouse.
- 3) A finders fee. The broker would also earn a fee (probably a commission) for all transactions that arise through the broker linking consumers directly with the provider.

The broker does not aim to be profit-making, and the fees levied will be to cover the overheads, support and maintenance only. The broker will be entirely owned by the project consortium.

4 EVALUATION PLAN

The project development will produce a working prototype of the data broker, based on this initial User Requirements Specification. This prototype will be fully tested at the culmination of the project, in June 2002, against an Evaluation Plan. This plan has been objectively drawn up in principal now, and mirrors the requirements stated in this document. The Evaluation Plan is provided in Appendix 5.

5 MOCK-UP APPLICATION

In order to demonstrate the principals of the data broker, and provide a basis of the architecture of the system, a mock-up application is being developed at an early stage of the project. This mock-up will provide a working graphical model that will aid the project consortium in the development of their ideas, and allow the easy demonstration of how the system will work to both the EU, and to external organisations. The mock-up will undoubtedly develop throughout the project, and will eventually culminate in the full, working prototype at the end of the project. A full description of the mock-up is provided in Appendix 6.



APPENDICES