

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Exploring the Sea: Heterogeneous Geo-Referenced Data Repository

Inês Davim Lopes Garganta da Silva

DISSERTATION



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Gabriel David

Second Supervisor: José Devezas

July 29, 2016

Exploring the Sea: Heterogeneous Geo-Referenced Data Repository

Inês Davim Lopes Garganta da Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor Carla Lopes

External Examiner: Doctor Alberto Simões

Supervisor: Doctor Gabriel David

July 29, 2016

Abstract

The marine environment is subject of an increasing attention and constitutes a dynamic and multidimensional environment that is very demanding for data collection and update, requiring large amounts of data. This data is unique since the campaigns where it is gathered are unrepeatable, due to the existence of a wide range of factors outside the researchers' control. Moreover, funding for these campaigns can be hard to come by. Nevertheless, these datasets are often underused as they are not available to all the involved stakeholders, or involve non-interoperable formats.

Currently, metadata and some of the data are registered in paper-based forms, which are later digitalized or transcribed to spreadsheets, with researchers placing emphasis in the publications rather than in the management of the collected data. Data provenance often relates to soil, water and biological samples, as well as sensors, ship routes, photos, videos, sounds and laboratorial analyses.

This problem is reflected in the large BIOMETORE project that involves several teams of marine researchers lead by Instituto Português do Mar e da Atmosfera (IPMA). The ultimate goal of BIOMETORE is the achievement and maintenance of the Good Environmental Status (GES) of the European Marine Waters. This project has eight campaigns, producing large amounts of marine data that should be organized in order to enable reusability by different stakeholders. On the other hand, the SeaBioData project, lead by INESC TEC, aims at developing a georeferenced database for BIOMETORE, that can integrate all available data and implement existing standards for data interoperability, as specified in directives such as INSPIRE. Building the database is essential to allow uniform data access by local researchers as well as the international community and, at the same time, reduce the required effort allocated to data management, promoting faster and more accurate scientific results.

In order to respect the INSPIRE directive, SeaBioData has adopted the data model from the OGC Sensor Observation Service. This data model has already been adopted by the international community, which ensures that the implementation relies on an interoperable approach. The project surveyed available technological options, as well as the datasets supplied by IPMA and has chosen the open source implementation from 52° North, since it supports the majority of the SOS model's concepts and provides a native REST API and Web Services. The 52° North data model does not support the storage of all of the data required by IPMA for internal usage. One of the main data modelling challenges was to extend the existing data model without altering the original tables, thus centralizing the data, while ensuring that the model is compliant with existing services. The extension had to follow the metadata structure defined by the Portuguese specification SNIMar, which implied the study and implementation of its metadata profile. The Darwin Core standard has been followed, in order to store more details of the taxonomic rank of the species. Furthermore, the 52° North data model has been extended in order to address the local needs of BIOMETORE, since the SOS model is focused on data concerning the observations, disregarding information about relevant entities, such as teams, campaigns, users, documents or responsible parties.

Resumo

O ambiente marinho é objeto de uma atenção crescente e constitui um ambiente dinâmico e multidimensional que é muito exigente para os sistemas de monitorização, em termos de recolha e atualização de dados. Estes dados são únicos uma vez que as campanhas onde foram recolhidos são irrepetíveis, devido à existência de diversos fatores que estão fora do controlo dos investigadores. Além disso, o financiamento para estas campanhas pode ser difícil de conseguir. No entanto, estes conjuntos de dados são muitas vezes subutilizados ao não estarem disponíveis para todas as partes envolvidas, ou por terem formatos não interoperáveis.

Atualmente, os metadados e alguns dados são registados em formulários de papel que são posteriormente digitalizados ou transcritos para folhas de cálculo, dando os investigadores mais ênfase às publicações do que à gestão dos dados recolhidos. Os dados provêm de amostras de solo, água e biológicas, sensores, rotas de navios, fotografias, vídeos, sons e análises laboratoriais.

Este problema reflete-se no projeto BIOMETORE, um projeto de grandes dimensões que envolve várias equipas de investigadores marinhos liderados pelo Instituto Português do Mar e da Atmosfera (IPMA). O objetivo final do BIOMETORE é a obtenção e manutenção do bom estado ambiental (Good Environmental Status, GES) das águas marinhas europeias. Este projeto contém oito campanhas que produzem grandes quantidades de dados marinhos; estes devem ser organizados de modo a permitir a sua reutilização pelas diferentes partes interessadas. Por outro lado, o projeto SeaBioData, liderado pelo INESC TEC, visa o desenvolvimento de uma base de dados georreferenciada para o BIOMETORE, que possa reunir todos os dados disponíveis e que implemente as normas existentes de interoperabilidade de dados, conforme especificado em diretivas como a INSPIRE. A construção da base de dados é essencial para permitir que investigadores locais e comunidade internacional tenham um acesso uniforme aos dados e, ao mesmo tempo, para reduzir o esforço na gestão de dados, promovendo resultados científicos mais rápidos e precisos.

De modo a respeitar a diretiva INSPIRE, o SeaBioData adotou o modelo de dados do OGC Sensor Observation Service. Este modelo foi adotado pela comunidade internacional, o que garante que a implementação constitui uma abordagem interoperável. Analisadas as opções tecnológicas disponíveis, bem como os dados fornecidos pelo IPMA, decidiu-se pela implementação open source do 52° North, uma vez que suporta a maioria dos conceitos do SOS e fornece uma API REST e Serviços Web nativos. O modelo de dados do 52° North não suporta o armazenamento de todos os dados exigidos pelo IPMA para uso interno. Um dos principais desafios na modelação de dados foi estender o modelo existente, sem alterar as tabelas originais, centralizando assim os dados e assegurando que o modelo é compatível com os serviços existentes. Foi também seguida a estrutura de metadados definida pela especificação portuguesa SNIMar, o que implicou o estudo e implementação do respetivo perfil de metadados. Seguiu-se ainda o standard Darwin Core, de modo a armazenar mais detalhes sobre a classificação taxonómica das espécies. Para além disso, adicionou-se uma extensão ao modelo de dados do 52° North de modo a responder às necessidades locais do BIOMETORE, uma vez que o modelo do SOS se foca nos dados relativos às observações, ignorando entidades como as equipas, campanhas, utilizadores, documentos ou responsáveis.

Acknowledgements

Este foi um percurso longo e árduo, que chega agora ao fim. Nesta reta final, fica a clara sensação de um balanço positivo, que me permitiu alargar horizontes, renovar e reforçar os meus conhecimentos.

Para que este trabalho pudesse ficar concluído foi fundamental o apoio e colaboração de várias pessoas que, de uma forma mais direta ou indireta, contribuíram para que tivesse a energia necessária para terminar esta etapa.

Ao Professor Doutor Gabriel David, pela mestria e empenho com que orientou a minha tese e ao Engenheiro José Devezas, pela dedicação contínua que depositou neste trabalho.

À equipa do SeaBioData agradeço a disponibilidade permanente e a ajuda.

Aos meus amigos de sempre que me ajudam a manter o equilíbrio na minha vida e me permitiram ter a energia que este processo me exigiu.

Aos meus colegas e amigos do InfoLab, agradeço a amizade e companheirismo, que foram fundamentais para a alegria permanente com que encarei este desafio.

À família do meu namorado, agradeço o carinho e hospitalidade com que sempre me recebem. Em particular ao Agostinho de Sousa Pinto por ser um exemplo próximo de uma referência na área de sistemas de informação.

À minha família pelo amor e esperança, que sempre depositam em mim. À minha irmã e à minha mãe agradeço em especial a dedicação e ao meu pai agradeço o exemplo de ser um académico de referência.

Aos meus avós, tios e primos agradeço por preencherem a minha vida.

Ao Hugo Gilvaia agradeço o amor e companheirismo. A sua permanente disponibilidade para me apoiar e ajudar neste e em todos os projetos da minha vida.

A todos agradeço mais uma vez, pois sem vocês não teria sido possível levar este projeto a bom porto.

Inês Davim Lopes Garganta da Silva

*“Learn from yesterday, live for today, hope for tomorrow.
The important thing is not to stop questioning.”*

Albert Einstein

Contents

1	Introduction	1
1.1	Context	1
1.2	Campaign Planning and Data Acquisition	2
1.2.1	Campaign Planning	2
1.2.2	Campaign Data	2
1.2.3	Sensor Data	3
1.2.4	Laboratory Data	4
1.3	Motivation and Goals	5
1.4	Summary	6
1.5	Report Structure	6
2	State of the Art	7
2.1	OGC Web Services	7
2.1.1	OGC Web Processing Service	7
2.1.2	OGC Catalogue Service for the Web	9
2.1.3	OGC Web Coverage Service	9
2.1.4	OGC Web Map Service	9
2.1.5	OGC Web Feature Service	10
2.1.6	OGC Sensor Observation Service	11
2.1.7	OGC Implementations	13
2.1.8	Observation & Measurements	14
2.2	INSPIRE Directive	14
2.2.1	Observations	15
2.3	Other Work	21
2.3.1	Geo-Seas	21
2.3.2	Baltic Operational Oceanographic System	21
2.3.3	Water Framework Directive	21
2.4	Marine Data Management Projects	22
2.5	Time Series Conceptual Models	22
2.6	Summary	22
3	Challenges of Marine Data Heterogeneity	25
3.1	Problem Description	25
3.2	An Information System to Integrate Marine Data	26
3.3	Summary	27

CONTENTS

4	Repository Architecture	29
4.1	Marine Research Data Workflow	29
4.2	Data Model	30
4.2.1	52° North	30
4.2.2	SNIMar and Darwin Core	35
4.2.3	Project Local Data	39
4.3	Summary	42
5	Discussion	45
5.1	Mapping the Data	45
5.1.1	Observations Data	45
5.1.2	Project Local Data	50
5.2	File Preservation	51
5.3	Summary and Conclusions	51
6	Conclusions and Future Work	53
6.1	Conclusions	53
6.2	Future Work	54
	References	55
A	Full Model	59

List of Figures

1.1	Conductivity-Temperature-Depth sensor.	3
1.2	Sieves.	4
1.3	Microscopy image of a species stained with rose bengal.	4
1.4	Liquid preservation.	4
1.5	Long-term maintenance in the Museum of Natural History of Berlin.	4
1.6	Abludomelita obtusata, female.	5
1.7	Phytoplankton.	5
2.1	Common sequence of WPS operations [Ope15].	8
2.2	CSW protocol sequence diagram [Ope07].	10
2.3	Political, coastline, and populated areas, Southeaster United States [Ope06].	11
2.4	WFS protocol diagram [Ope05].	12
2.5	Class diagram centred on the <i>OM_Observation</i> entity [INS11].	15
2.6	INSPIRE: single result in time for a single point [INS11].	18
2.7	INSPIRE: results with multiple values in time on a single point [INS11].	19
2.8	INSPIRE: results with multiple values in time on a curve [INS11].	19
2.9	INSPIRE: results with multiple values in time on a surface [INS11].	20
2.10	INSPIRE: feature-of-interest as specimen [INS11].	20
4.1	Representation of the different marine data collection stages and their interactions.	29
4.2	Table <i>observation</i> from the 52° North data model [Hol12b].	31
4.3	Table <i>featureofinterest</i> from the 52° North data model [Hol12b].	31
4.4	Table <i>observableproperty</i> from the 52° North data model [Hol12b].	32
4.5	Table <i>procedure</i> from the 52° North data model [Hol12b].	32
4.6	Table <i>offering</i> from the 52° North data model [Hol12b].	33
4.7	Table <i>series</i> from the 52° North data model [Hol12b].	33
4.8	52° North API console.	34
4.9	Metadata model based on SNIMar.	36
4.10	Taxonomic Rank model based on Darwin Core.	39
4.11	BIOMETORE local database model.	40
5.1	Benthos spreadsheet data example.	46
5.2	Mapped benthos data for the <i>Observation</i> table.	46
5.3	Mapped benthos data for the <i>Series</i> table.	47
5.4	Mapped benthos data for the <i>observableproperty</i> table.	47
5.5	Mapped benthos data for the results tables.	47
5.6	CTD data example.	48
5.7	Mapped CTD data for the <i>observation</i> table.	48
5.8	Mapped CTD data for the <i>Series</i> table.	48

LIST OF FIGURES

5.9	Mapped CTD data for the <i>observableproperty</i> table.	49
5.10	Mapped CTD result template for the <i>resulttemplate</i> table.	49
5.11	Mapped CTD data for the results tables.	50
5.12	Mapped data for the <i>campaign</i> table.	50
5.13	Mapped data for the <i>station</i> table.	50
A.1	Full model	60

Abbreviations

DB	Database
EMODNET	European Maritime Observation and Data Network
GES	Good Environmental Status
GMES	Global Monitoring for Environment and Security
GML	Geography Markup Language
INSPIRE	Infrastructure for Spatial Information in the European Community
MCS	Marine Core Service
OGC	Open Geospatial Consortium
O&M	Observations and Measurements
RDBMS	Relational Data Base Management System
SOS	Sensor Observation Service
UML	Unified Modeling Language
WCS	Web Coverage Service
CSW	Catalogue Services for the Web
WFD	Water Framework Directive
WFS	Web Feature Service
WMS	Web Map Service
WPS	Web Processing Service

Chapter 1

Introduction

Since we are living in a "data economy", information and data are the most important assets for research institutions and companies. The need for securely keeping the data and making it accessible for public-funded research, warranting that studies can be replicated or elaborated upon by other researchers, has been recently recommended by the European Commission. Although the emphasis is often on information transfer and electronic databases, biological and geological data always harbours much more information than the previously mentioned, constituting tangible testimonies of conditions, processes and events that can be constantly revisited and analysed in light of cutting-edge techniques and methods.

Environmental data contains a lot of spatial information. This information has been recognized as one of the most important elements that supports decision-making in many areas [LSW06]. It provides a geographical context which enables planning, management and resource allocation, allowing a better understanding, and thus better management of a given area [LSW06].

A standard recommendation is used to ensure interoperability and integrability of different datasets. An example for managing spatial data is the INSPIRE Directive [Com].

"In Europe a major recent development has been the entering in force of the INSPIRE Directive in May 2007, establishing an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment." [Com]

1.1 Context

This work has been developed in the context of a large project involving marine researchers and computer scientists. The project, called BIOMETORE¹, produces a big amount of data in the field of marine sciences for the Portuguese Seas. Its goal is the achievement and maintenance of the

¹<http://oom.arditi.pt/biometore/>

Good Environmental Status (GES) of the European marine waters. GES means that the marine resources are used in a sustainable level, ensuring their continuity for future generations.

Data formats of special interest to Portuguese marine waters are heterogeneous, being stored in excel sheets, detectors, instrument-specific data formats, photos, videos and paper ledgers, usually with metadata regarding the online data taking. The spread of data, data formats and corresponding metadata make it extremely challenging for the researchers to analyse and process the data collectively from different surveys, or correlate data from different sources or measurements. BIOMETORE requires a geo-referenced database able to gather all the different data in order to provide an integrated view of the environment.

1.2 Campaign Planning and Data Acquisition

Prior to launching the campaigns, they are pre-planned and organised. This phase will be detailed in this chapter, in order to provide a better insight of the BIOMETORE processes.

The teams from Instituto Português do Mar e da Atmosfera (IPMA) are responsible for the data acquisition process associated with the datasets used in our work. The collection and production of data has three phases, the campaign phase, where the data is collected *in situ*, the laboratory phase, where more data is produced under laboratory procedures *ex situ*, and the analysis phase, where results are derived from the data.

1.2.1 Campaign Planning

Campaigns are planned far in advance from departure. Researchers must consider several aspects such as available equipment, weather conditions and designated personnel. In addition, they have to decide which locations and stations they will visit, the type of samples and which forms they are going to fill.

In order to facilitate the planning of the campaigns, we must take all of this into consideration. As later explained in Chapter 4, our database was modelled based on these requirements, including for instance the choice of locations in the pre-campaign phase, as well as the precise coordinates of the several stations in each main location, which must be done simultaneously to sample collection, in order to be able to precisely specify the latitude and longitude of the samples.

1.2.2 Campaign Data

During the campaign, produced data mostly consists of metadata about the observed species or captured samples. It may be, for example, the site, latitude, longitude, datum, date, begin and end time, tide situation, tidal zone, depth and additional notes. This metadata is stored in paper ledgers, that are later archived in traditional dossiers or stored in properly configured mobile applications, for later integration in the main repository.

The data collection sites are defined beforehand, prior to the campaigns, and the data can be from biotic or abiotic samples, species distributions, fishing diaries, photographs, videos and

sensors. Some of the data is processed while still on board. The collected samples undergo some sampling methods like conservation (freezing, refrigerating, drying), sieving and fixing. The metadata is transcribed to spreadsheets and analysed to make operational decisions, such as the number of collections to be created from a certain data type.

1.2.3 Sensor Data

One of the most common types of data collected by sensors is called CTD data from Conductivity-Temperature-Depth. The term CTD is frequently used to denominate a package that includes the previous measures, plus data from auxiliary sensors that measures other parameters, such as dissolved oxygen, pH, fluorescence and turbidity. This data is used in applications such as: the calculation of deep-ocean currents based on the variations of density; the detection of areas with low oxygen, which can have a serious impact on the fish population; and the search of areas with strong density changes to examine the chemical and biological processes that go on there. An example of a CTD device is showed in Figure 1.1.



Figure 1.1: Conductivity-Temperature-Depth sensor.²

The data is extracted to hex files that include a header with information like the latitude, longitude, time, ship, cruise, and a body with the measures in hexadecimal. The IPMA researchers clean the data (remove bag flags for example) and then transform the file to the CNV format, that is, a converted engineering unit data file. To do this they use a program called SBE Data Processing³. In the CNV file, the body is converted from hexadecimal to engineering units.

These files have a predefined structure that can be extended with new header lines as defined in the file format. During the ETL⁴ phase, we built a parser based on an open source Python solution called Seabird⁵. We extended this package to read the name of the ship, the cruise number, the station and the depth. We also altered the parser of the latitude and longitude to accept cardinal points besides the already supported signed coordinates. In addition, we fixed a bug preventing the date format to be supported in the Microsoft Windows environment (i.e. we changed "%Y%m%s%H%M%S" to "%Y%m%d%H%M%S"). Thereby, the CNV data was ready to be parsed and loaded to the database.

²Photo 1.1 by [Wikipedia, CC BY 3.0](#)

³<http://www.seabird.com/software/sbe-data-processing>

⁴Extract, Transform and Load the data

⁵<http://seabird.readthedocs.io>

1.2.4 Laboratory Data

The samples collected in the campaigns are later analysed in the laboratory. The samples go through laboratorial processes like sieving (Figure 1.2), where wanted elements are separated from unwanted material or where the particle size distribution is characterized; coloring (Figure 1.3), that is used, for example, to prepare the sample for microscopic analysis, allowing the distinction between forms that are alive or dead at the time of collection; formol fixation, that is a critical step in the preparation of histological sections by which biological tissues are preserved from decay; and alcohol preservation (Figure 1.4), which is usually used to preserve most arthropods like insects, crustaceans and arachnids. The specimens analysed should be later deposited in a reference collection in an institution for long-term maintenance and access (Figure 1.5), thus ensuring that the animals are preserved in the best possible conditions.



Figure 1.2: Sieves.



Figure 1.3: Microscopy image of a species stained with rose bengal.⁶

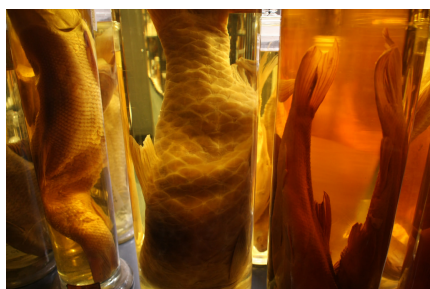


Figure 1.4: Liquid preservation.⁶



Figure 1.5: Long-term maintenance in the Museum of Natural History of Berlin.⁶

There is a huge amount of data produced in this phase. Examples include species identification (taxonomic rank, sex, life stage; Figure 1.6); phytoplankton concentration (the phytoplankton — Figure 1.7 — is the basis of the food chain of the oceans); biomass, that is the total quantity or weight of organisms in a given area or volume; granulometric composition, that is the determination of the size of the sample particles and the percentages of their occurrences; and chemical

⁶ Photo 1.3 by Danovaro R., Dell'Anno A., Pusceddu A., Gambi C., Heiner I. & Kristensen R. M., CC BY 2.0 / Photo 1.4 by Solja Virkkunen, CC BY-NC-SA 2.0 / Photo 1.5 by anna malina / CC BY-NC 2.0



Figure 1.6: *Abludomelita obtusata*, female.⁷

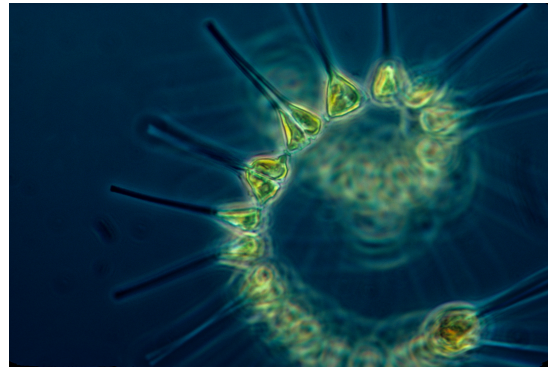


Figure 1.7: Phytoplankton.

composition, which refers to the arrangement, type, and ratio of atoms in molecules of chemical substances.

1.3 Motivation and Goals

The main goal of this work is to guarantee that the biological material and associated information, collected in the BIOMETORE project, as well as past and future data, will be fully accessible for present and future research.

The database design and the data integration will ensure uniform access to all researchers at a local and international level. This approach will allow researchers to focus on their research rather than managing their e-infrastructure needs, resulting in faster scientific results.

The goals are then to provide an adequate technological framework for the effective and efficient monitoring of particular ecosystems such as seamounts, and providing adequate and timely information, thus contributing for the monitoring of resources and the promotion and preservation of marine environment and biodiversity.

The system will be composed of two subsystems: on one hand, a repository that allows for campaign information recording, including video, physical and biological data; on the other hand, a system to assimilate and provide data from observations in various temporal scales (time series) according to an interoperable⁸ interface.

Information from two different domains will be integrated:

1. Information from the human pressures (fishing activities);
2. Information on the ecosystem monitoring.

The objectives of this work are:

1. Definition and design of a data management system to integrate present and future data;

⁷Photo 1.6 by [Hans Hillewaert](#) / CC BY-SA 4.0

⁸"Interoperability is a property of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, without any restricted access or implementation." [[dUdLL](#)]

2. Database architecture and design;
3. Propose/define the long term digital preservation of data and metadata.

1.4 Summary

Projects, like the BIOMETORE, collect large amounts of heterogeneous data. Surveying data acquisition methodologies and analysing existing datasets was one of the most important steps of our work, leading us to understand the structure and content of the various data types. A considerable amount of work was required in order to achieve a consistent data model. Providing a database capable of storing such a multitude of data formats, while providing a logical and integrated organization of the information, as well as supporting existing storage and operational needs that complied with the INSPIRE directive, was a considerable challenge.

1.5 Report Structure

The remainder of this work is organized as follows. In Chapter 2, we describe the state of the art and present previous works related to environmental information systems and frameworks. We also describe international norms as OGC and INSPIRE. In Chapter 3, we write about the challenges of the marine data heterogeneity and we make a propose for the solution of this problem. In chapter 4 we explain the data models and its components. In chapter 5 we discuss the workings of the model with examples from real data. And, finally, in Chapter 6 we present the conclusions and future work.

Chapter 2

State of the Art

In this chapter, we present the state of the art and related work, covering directives, standards and geographic projects. We also do a technological survey, where we review web services, frameworks and UML implementations of time series.

2.1 OGC Web Services

"The OGC (Open Geospatial Consortium) is an international non-profit organization committed to making quality open standards for the global geospatial community. These standards are made through a consensus process and are freely available for anyone to use to improve sharing of the world's geospatial data.

OGC standards are used in a wide variety of domains including Environment, Defence, Health, Agriculture, Meteorology, Sustainable Development and many more." [Con16a]

The OGC service models define a set of requirements for access and processing of spatial data, providing capabilities for sets of geographical features. The OGC baseline comprises more than 30 standards, as CSW, GML, GeoXACML, SensorML, WCS, WFS, WPS, WMS and GeoSPARQL [Wik16]. The current and relevant standards for interoperable marine environment modelling are presented below.

2.1.1 OGC Web Processing Service

"A Web Processing Service (WPS) is a web service¹ that enables the execution of computing processes and the retrieval of metadata describing their purpose and functionality. Typically, these processes combine raster, vector, and/or coverage data

¹Web services provide the ability to pass messages between computers over the Internet, therefore allowing geographically distributed computers to more easily share data and computing resources [JLGZ08].

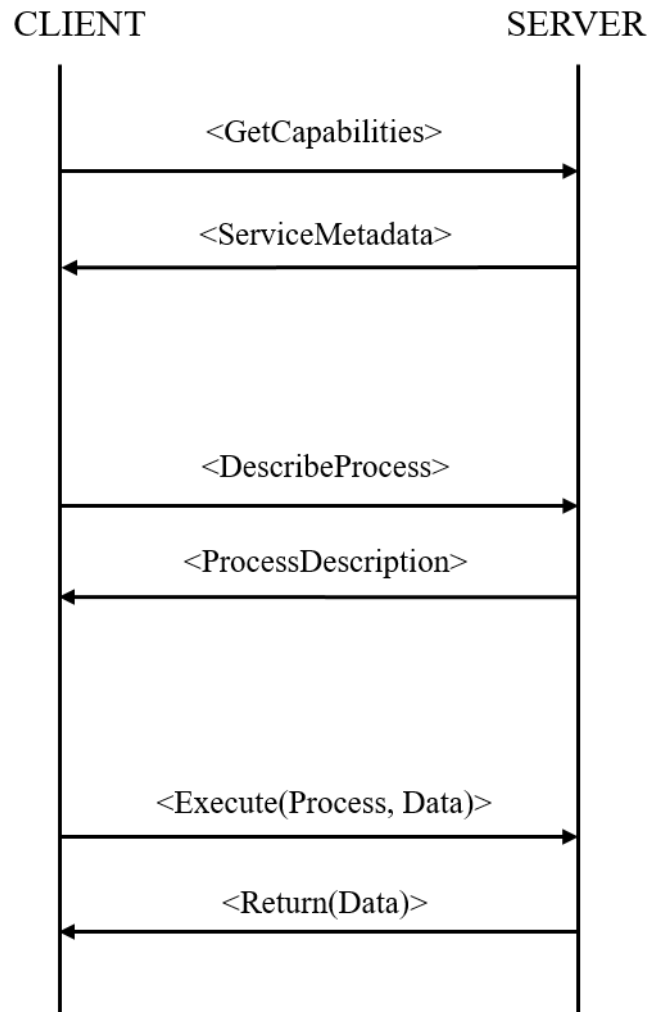


Figure 2.1: Common sequence of WPS operations [Ope15].

with well-defined algorithms to produce new raster, vector, and/or coverage information." [Ope15]

The main WPS functionalities are the `GetCapabilities()` that allows a client to request metadata about the capabilities and processes of the server. It is complemented by the methods `DescribeProcess()`, that enables a client to request metadata related to certain processes offered by the server, and `Execute()`, that allows a client to execute a process with an identifier and desired data inputs and output formats [Ope15].

In Figure 2.1, the OGC [Ope15] illustrates the common sequence of WPS operations. A client should first make a `GetCapabilities()` request to the server to get an updated listing of available processes. Then, it may request a `DescribeProcess()` to find out more information about particular processes offered, including the supported data formats. To run a process with a given input data, a client will make an `Execute()` request [Ope15].

2.1.2 OGC Catalogue Service for the Web

"Catalogue services support the ability to publish and search collections of descriptive information (metadata) for data, services, and related information objects. Metadata in catalogues represent resource characteristics that can be queried and presented for evaluation and further processing by both humans and software." [Ope07]

The main operations of the Catalogue Service for the Web (CSW) are: `GetCapabilities()` that allows a client to retrieve metadata about the Catalogue Service instance; `DescribeRecord()` which enables a client to get the type definitions used by the metadata of one or more registered resource types; `GetDomain()` which allows a client to fetch the domain (allowed values) of a metadata property or request parameter; `GetRecordById()` retrieves the default representation of catalogue records using their identifier; `Transaction()` which defines an interface for creating, modifying and deleting catalogue records; and `Harvest()` that pulls data into the catalogue [Ope07].

In Figure 2.2, the OGC [Ope07] displays the request/response message pairs for all the operations defined for the catalogue service. It shows three classes of operations. The Service Class which has the `GetCapabilities()` operations; the Discovery Class that has the operations used to determine the information model of the catalogue; and the Management Class that has operations used to create or change the records in the catalogues [Ope07].

2.1.3 OGC Web Coverage Service

"The OGC Web Coverage Service (WCS) supports electronic retrieval of geospatial data as "coverages" – that is, digital geospatial information representing space/time-varying phenomena." [Ope12]

The main operations that the WCS interface specifies are: `GetCapabilities()` that allows a client to request metadata about selected coverages offered by the server and `GetCoverage()` that enables a client to request a coverage included in a chosen range of properties at a selected set of spatio-temporal locations, expedited in coverage encoding formats [Ope12]. The coverage formats are *MultiPointCoverage*, *MultiCurveCoverage*, *MultiSurfaceCoverage*, *MultiSolidCoverage*, *GridCoverage*, *RectifiedGridCoverage* and *ReferenceableGridCoverage* [Ope10a].

2.1.4 OGC Web Map Service

"A Web Map Service (WMS) produces maps of spatially referenced data dynamically from geographic information. This International Standard defines a "map" to be a portrayal of geographic information as a digital image file suitable for display on a computer screen. A map is not the data itself. WMS-produced maps are generally rendered in a pictorial format such as PNG, GIF or JPEG, or occasionally as vector-based graphical elements in Scalable Vector Graphics (SVG) or Web Computer Graphics Metafile (WebCGM) formats." [Ope06]

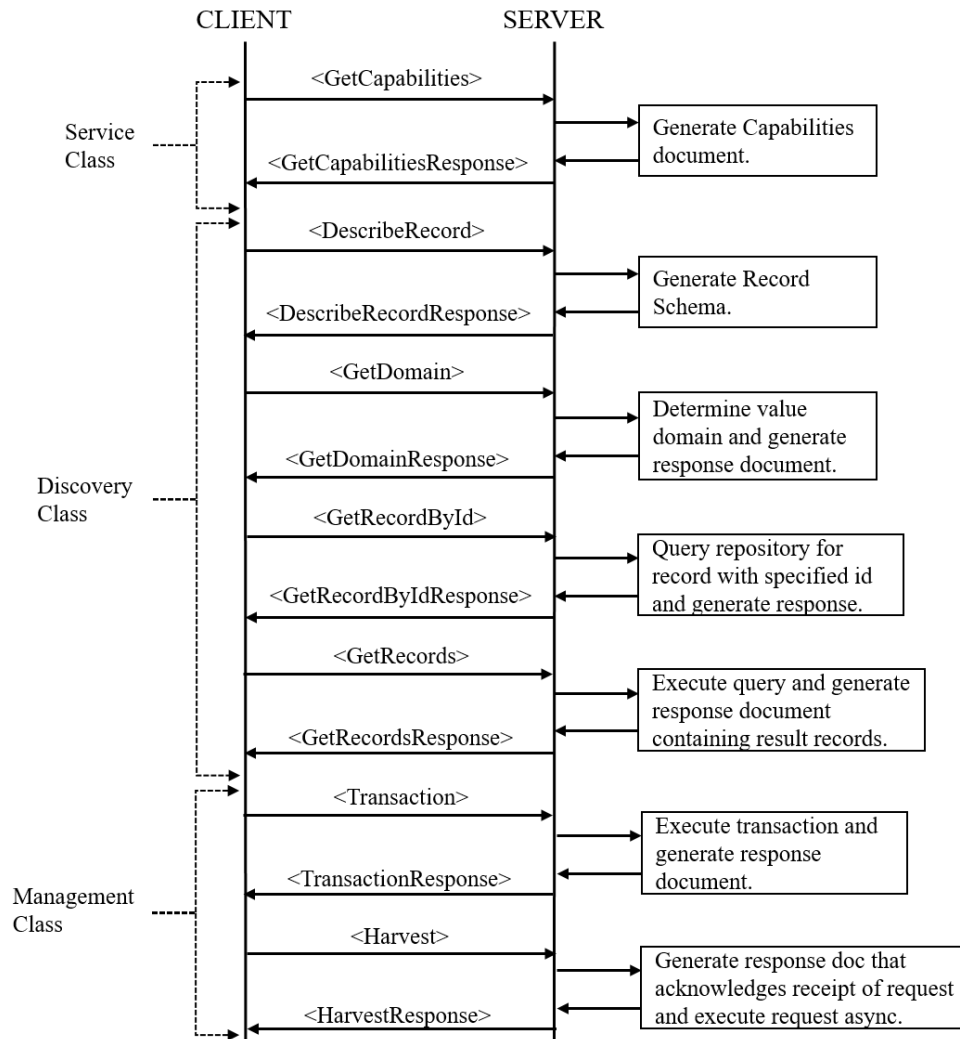


Figure 2.2: CSW protocol sequence diagram [Ope07].

The main operations of the WMS are: `GetCapabilities()` to obtain service metadata and `GetMap()` which returns a map of the spatial information layer requested [Ope06]. The Figure 2.3, from OGC [Ope06] shows an example of an output for the given URL:

```

http://b-maps.com/map.cgi?VERSION=1.3.0&REQUEST=GetMap&
CRS=CRS:84&BBOX=-97.105,24.913,-78.794,36.358&
WIDTH=560&HEIGHT=350&LAYERS=BUILTUPA_1M,COASTL_1M,POLBNDL_1M&
STYLES=0xFF8080,0X101040,BLACK&FORMAT=image/png&BGOLOR=0xFFFFFF&
TRANSPARENT=TRUE&EXCEPTIONS=INIMAGE

```

2.1.5 OGC Web Feature Service

"The Web Feature Service (WFS) represents a change in the way geographic information is created, modified and exchanged on the Internet. Rather than sharing geo-

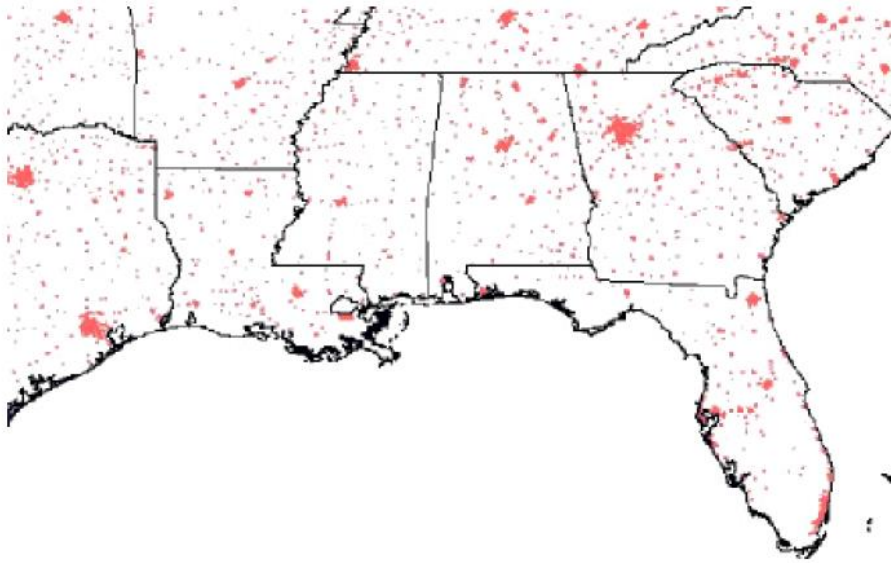


Figure 2.3: Political, coastline, and populated areas, Southeaster United States [Ope06].

graphic information at the file level using File Transfer Protocol (FTP), for example, the WFS offers direct fine-grained access to geographic information at the feature and feature property level. Web feature services allow clients to only retrieve or modify the data they are seeking, rather than retrieving a file that contains the data they are seeking and possibly much more. That data can then be used for a wide variety of purposes, including purposes other than their producers' intended ones." [Ope14]

The main operations in order to support transaction and query processing are: `GetCapabilities()` that indicates which feature types and what operations the web feature can service and support respectively; `DescribeFeatureType()` that describes the structure of any feature type of the web feature; and `Transaction()` that contains the operations to modify features, as create, update and delete [Ope10b].

In Figure 2.4, the OGC [Ope05] shows a simplified protocol diagram illustrating a typical transaction protocol, and the messages that might be passed between a client application and a web feature service. A typical flow would be, for example, a client sends a `DescribeFeatureType()` request and the server responds with a data layout description document.

2.1.6 OGC Sensor Observation Service

"The Sensor Observation Service (SOS) standard is applicable to use cases in which sensor data needs to be managed in an interoperable way. This standard defines a Web service interface which allows querying observations, sensor metadata, as well as representations of observed features. Further, this standard defines means to register

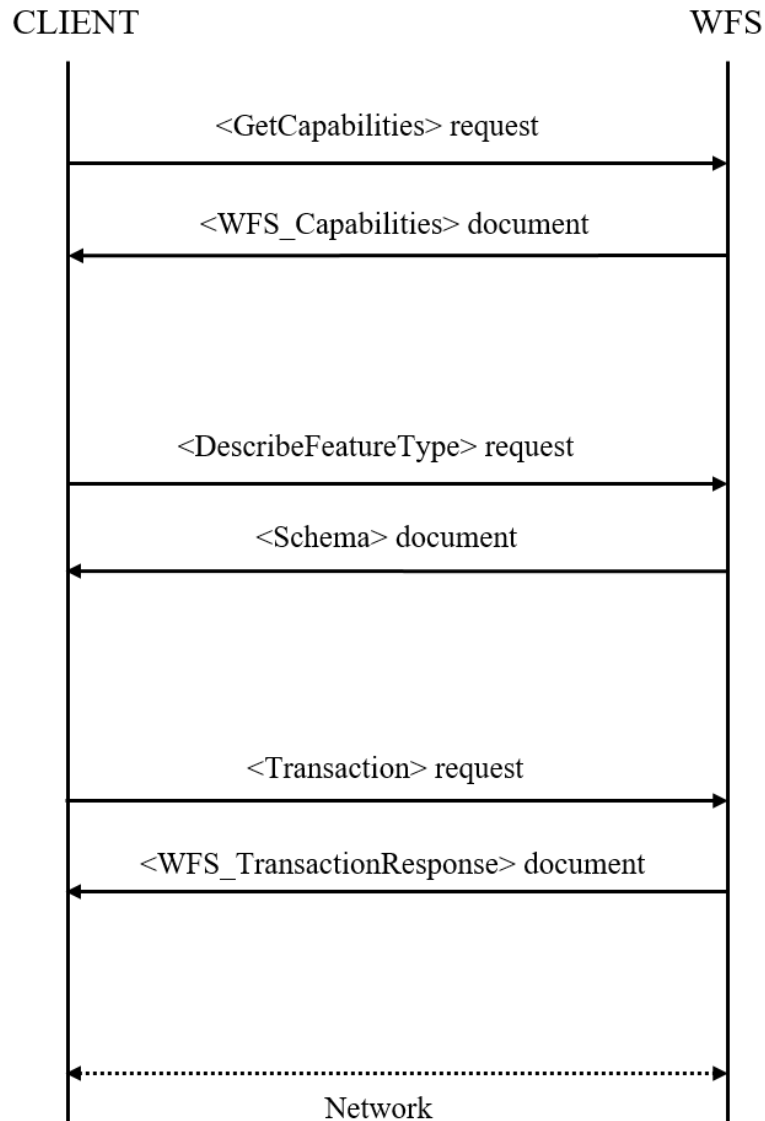


Figure 2.4: WFS protocol diagram [Ope05].

new sensors and to remove existing ones. Also, it defines operations to insert new sensor observations." [Con16b]

The SOS operations are separated in three profiles. The Core Profile has three operations provided by every SOS implementation. The Transactional Profile, optional, is used to register sensors and insert observations into the SOS. The Enhanced Profile has additional operations [OGC11].

Core Profile

`GetCapabilities()` Returns a service description with information about the service interface and the sensor data;

`DescribeSensor()` Returns a description of one specific sensor, sensor system or data producing procedure;

`GetObservation()` Provides access to sensor observations and measurement-data.

Transactional Profile

`RegisterSensor()` Allows the register of new sensors with the SOS via the service interface;

`InsertObservation()` Allows the client to insert new observations for a sensor system.

Enhanced Profile

`GetResult()` Allows a client to repeatedly obtain data from the same set of sensors;

`GetObservationById()` Returns an observation based on an identifier;

`GetFeatureOfInterest()` Returns one or more descriptions of features of observations;

`GetFeatureOfInterestTime()` Returns the time periods for which the SOS will return data;

`DescribeFeatureType()` Returns the XML schema for the specified feature;

`DescribeResultModel()` Returns the XML schema for the result of a certain observation type.

2.1.7 OGC Implementations

OGC is a set of standards, meaning that it only defines services interfaces and not implementations. Some open source² implementations of the services are described next.

52° North is an open international network of partners from research, industry and public administration. Its communities develop new concepts and technologies in the field of Geoinformatics, like managing real-time sensor data [fGOSSG16]. Their SOS implementation provides an interface that makes the sensor data available through an interoperable web based interface [fGOSSG14];

deegree is an open source software for spatial data infrastructures. It offers data access, visualization, discovery and security, implementing open standards from OGC like WMS, WFS and WPS and INSPIRE [dee16];

MapServer is a geographic data rendering engine. It allows for the creation of geographic image maps that can direct users to the content, providing spatial context when needed [Fou16]. The MapServer is not a full-featured geographic information system, nor it is intended to be, instead it excels in presenting spatial data on the web. It supports OGC standards like WMS, non-transactional WFS, WCS, SOS and O&M [Wik13];

²Open source software is computer software with its source code made available.

istSOS is an OGC SOS server implementation that allows the managing and dispatching of observations from sensors. It also provides a graphical user interface and a RESTful web API for administration procedures [MA15].

2.1.8 Observation & Measurements

The Observations & Measurements (O&M) is an international standard that defines a conceptual schema for the provision of measurements data. The model supports the exchange of information describing observation acts and their results, both within and between distinct technical and scientific communities [Ope13]. This standard arises from the OGC Sensor Web Enablement (SWE), that is concerned with establishing interfaces and protocols to enable a "Sensor Web", where applications and services are able to access sensors and the observations generated by them. Some of the components of the SWE are the O&M, the SOS and the SensorML³ [Ope13].

In the model the Observation is the central table, and has associated procedures, features, results and observed property. Its attributes are the phenomenon time, result time, valid time, result quality and parameters to describe any event of relevance to interpret the observation.

2.2 INSPIRE Directive

"INSPIRE (Infrastructure for Spatial Information in Europe) is based on the infrastructures for spatial information established and operated by the 28 Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications, with key components specified through technical implementing rules. This makes INSPIRE a unique example of a legislative "regional" approach." [Com]

The central idea of the INSPIRE Directive is that the INSPIRE web services exist to allow an easy and harmonized access of geospatial data both at national and thematic borders of the EU. Thus, the datasets should be adjusted, when possible, to the data specifications of each INSPIRE theme. In addition, financial savings can be achieved if the data distribution is appropriate between INSPIRE related topics such as Atmospheric Conditions, Meteorological Geographical features and the Oceanographic Geographical features [INS11].

The O&M standard is useful for data discovery using the primary entities of the model, the feature-of-interest, observed property and procedure. It can also assist in data fusion through different disciplines. The O&M is a very generic standard, thus allowing for a variety of design patterns depending on the domain and on the use cases to be supported [INS11].

³"The primary focus of the Sensor Model Language (SensorML) is to provide a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations." [OGC16]

2.2.1 Observations

"An Observation is an action whose result is an estimate of the value of some property of the feature-of-interest, at a specific point in time, obtained using a specified procedure." after Cox 2008 [INS11]

An observation is associated with a temporal attribute, representing either the instant or period when a number, term or other symbol was assigned to a phenomenon. It involves the use of procedures like sensors, instruments, algorithms or process chains. The result of an observation is an estimate value of a property of a feature. The use of a common model allows observation data using different procedures to be unambiguously combined [Ope13].

The O&M standard defines the *OM_Observation* type. Figure 2.5, from INSPIRE [INS11], shows the basic *OM_Observation* type. As we can see, *OM_Observation* has five attributes *phenomenonTime*, *resultTime*, *validTime*, *resultQuality* and *parameter*, and establishes many-to-one relationships to the process, feature and property type tables.

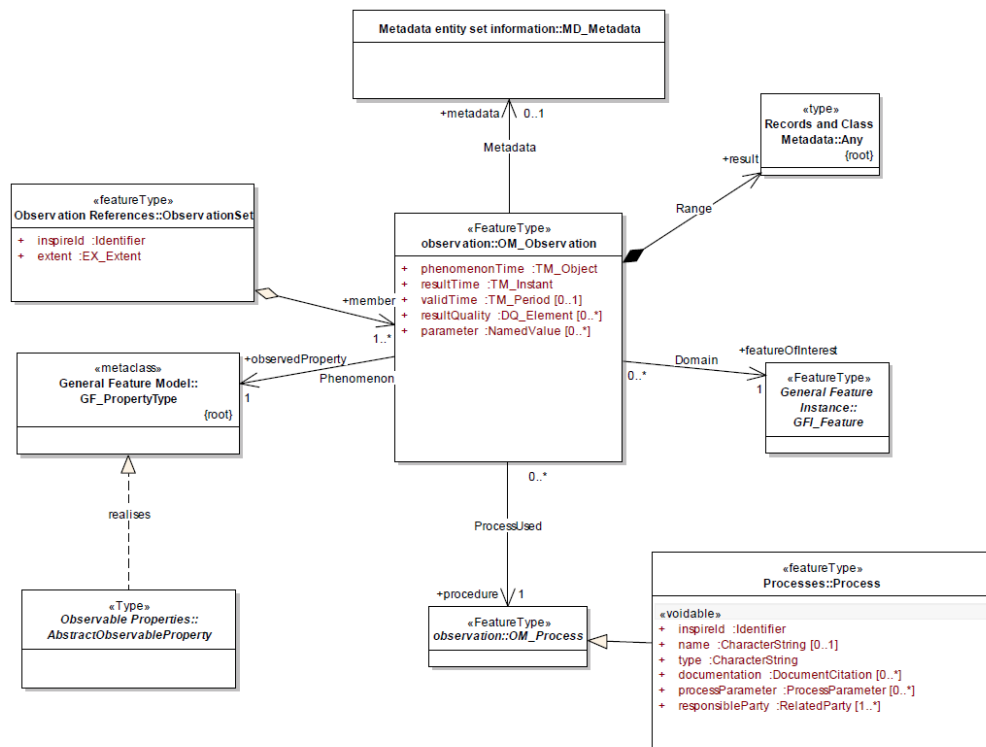


Figure 2.5: Class diagram centred on the *OM_Observation* entity [INS11].

Associations

Feature-of-interest

The association Domain links *OM_Observation* with the *GFI_Feature* that is the subject of the observation. The feature-of-interest (FoI) is the representation of the real world object the property is being estimated on.

Some of the types that are used as *featureOfInterest* are: species occurrence point, water column, trajectory, grid, water sample and rock sample.

In INSPIRE, the feature-types may serve as the feature-of-interest (for direct observations) or the sampled-feature (where a sample was taken from). The feature-types will often be withdrawn from the feature-concept-dictionary [Cen] and related feature-type-catalogue [ECJRC14]. For the feature-of-interest or sampled-feature the feature types should be derived from the *SF_SpatialSamplingFeature*.

Property

The association Phenomenon links *OM_Observation* with *GF_PropertyType*. This describes the property of the feature-of-interest that is being estimated in the observation.

Some of the types used as *observedProperty* are: species identification, height, O₃ hourly mean, salinity, water temperature and nitrate concentration.

The *observedProperty* should be assigned with persistent identifiers, such as Uniform Resource Identifiers (URI), allowing it to be referenced in observation instances. In INSPIRE, the *AbstractObservableProperty* class is defined as a realisation of the metaclass *GF_PropertyType*.

Procedure

The association ProcessUsed links *OM_Observation* with *OM_Process*. This describes the process or methodology used in the estimation of the result of an observation.

Some of the terms used to refer to the procedure are: method, sensor, instrument and analytical process.

The procedures may be expressed in SensorML or any other appropriate formulation. It should also be assigned with URIs. The feature type Process is derived from *OM_Process* that provides the minimum requirements, within INSPIRE, for the description of procedural information.

Result

The association Range links *OM_Observation* with the estimate of the property on the feature-of-interest generated by the procedure.

Some of the terms used to refer result are: observation value, measurement value, observation, value and analysis.

The Result can have many different encodings that depend on the nature of the stored data, which can be from a single value to a multidimensional result set. In INSPIRE, they omit the definition of specific result types. Instead they provide recommendations about the several result types on the various OGC codifications, as well as other formats that are more suited to specific purposes.

Attributes

Parameter

The *parameter* attribute describes any event-specific value of relevance to interpreting the observation. It will typically complement the description of the observation procedure. The data type *NamedValue* allows the provision of key/value pairs.

PhenomenonTime

The *phenomenonTime* attribute describes the time when the result applies to the property of the feature-of-interest. This could be the time when a specimen was collected or when a procedure was performed, but it may also be in the future in case of forecasts, or in the past for archaeological or geological observations. The *TM_Object* type allows time instants, time periods or temporal topologies.

ResultQuality

The *resultQuality* attribute must be provided when the result consists of a single value, or a set of values with the same quality. For complex results, where the quality varies across result values, the quality should be provided for each value in the result.

ResultTime

The *resultTime* attribute describes the time when the result became available. For some observations this is similar to the *phenomenonTime*. However, there are also important cases where they may differ, as a specimen analysed in the laboratory, where the *phenomenonTime* is when the specimen was captured, while the *resultTime* is when the laboratorial analysis was finished.

ValidTime

The *validTime* attribute describes the time period when the result is intended to be used. This is commonly used in forecasting applications.

2.2.1.1 Observations & Feature-Of-Interest

In order to provide a better comprehension of the O&M design patterns used in INSPIRE, we present a series of examples from the directive [INS11], structured by the dimensionality of the feature-of-interest of the observation. The differentiation point lies in the result.

The simplest design pattern provides a single result on a single point. This result can be a measurement, a categorization or other type of observation on the feature-of-interest. Figure 2.6 shows an example on how a measurement performed on an individual can be structured. In the example, the *featureOfInterest* is the "species occurrence point", the *observedProperty* is "Height", and it was observed using the "Magic Triangulation" procedure. The result is the "tree height" in meters.

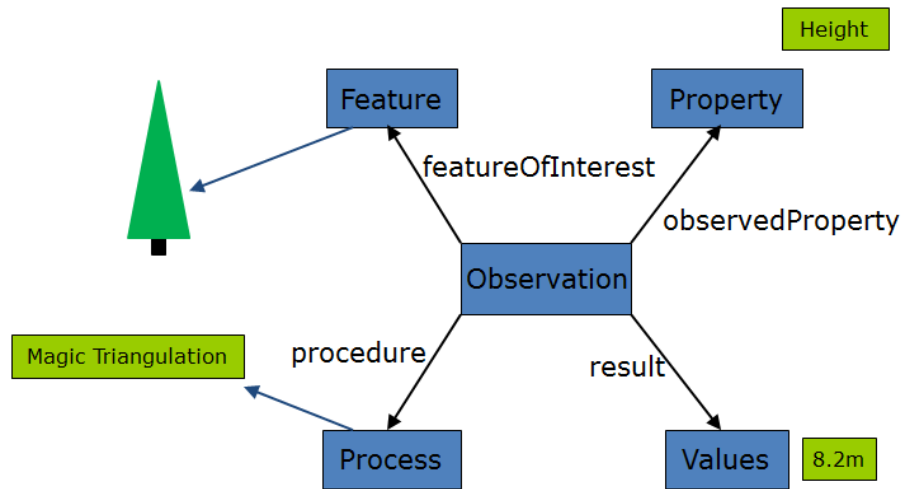


Figure 2.6: INSPIRE: single result in time for a single point [INS11].

In others cases, the observation can be a repeated measurements taken on a single point in space (time series). Figure 2.7 shows an example of how multiple measurements over time can be structured. This example is related to the one above on the height of a tree. In this case, the height of the tree is measured over a yearly period. Similarly, the *featureOfInterest* is "species occurrence point", the *observedProperty* is "Height" and, once again, it was observed using the "Magic Triangulation" procedure. The produced result is the "tree height" in meters, with a timestamp (year) provided for each height measurement.

Another type of design pattern provides single or multiple results, in time, on a curve. Figure 2.8 shows an example where the *featureOfInterest* represents the trajectory of a moving ship making sea surface temperature measurements. The trajectory is the *samplingFeature* with the spatial sampling feature (e.g. a *SF_SamplingCurve*) being the ocean. All measurement locations should be positioned in the trajectory described by the *featureOfInterest*. The produced result is the time, location and water temperature triple. The location can be relative (5m from the start) or absolute (geographic coordinates).

Finally, the design patterns provide single or multiple result values in time on a surface. Figure 2.9 shows the *featureOfInterest* represented as a grid over the ocean. In this example, the ocean temperature is modelled for each grid cell over time. All measurement locations should be

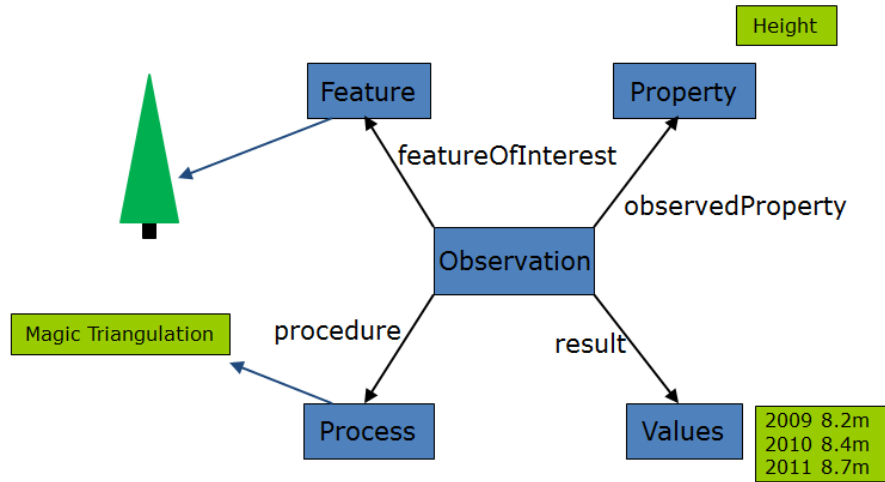


Figure 2.7: INSPIRE: results with multiple values in time on a single point [INS11].

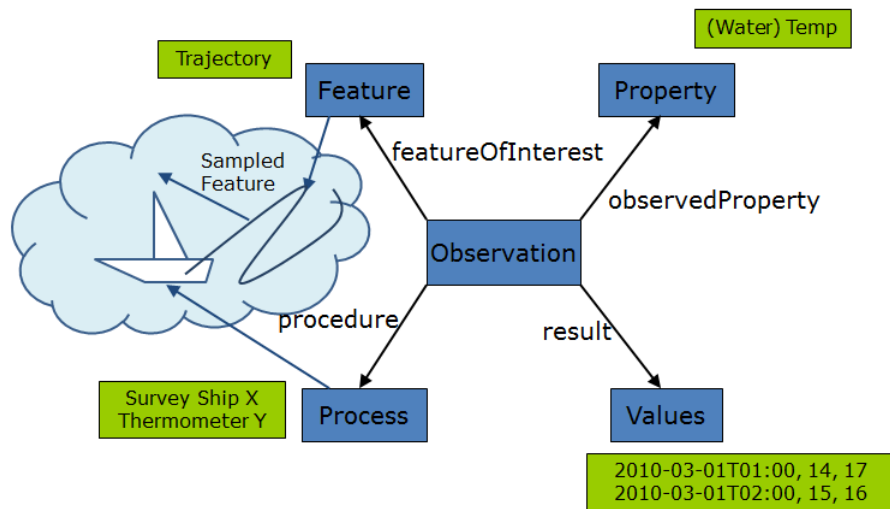


Figure 2.8: INSPIRE: results with multiple values in time on a curve [INS11].

positioned within the boundaries of the *featureOfInterest*. The result contains the grid cell and time in the domain, and correspondent range with the water temperature.

The feature-of-interest can also be a specimen (instead of a location). In this design pattern, the *featureOfInterest* is a sample or a specimen collected from the sampled feature and analysed in an external laboratory. The samples can either be analysed once, or several times in regular intervals. In Figure 2.10, we show an example of the latter, where the biochemical oxygen demand (BOD)

was measured in waste treatment plants. The measured was made by taking one sample and studying the BOD evolution over time.

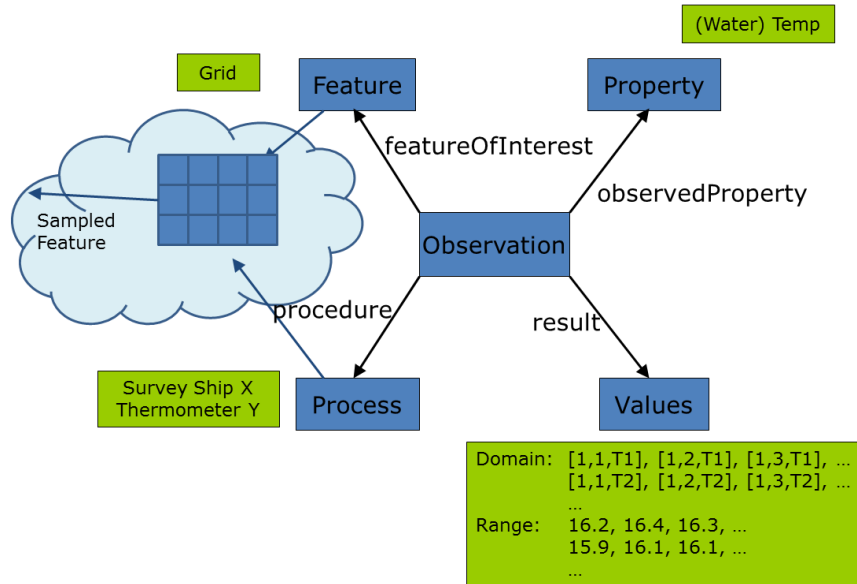


Figure 2.9: INSPIRE: results with multiple values in time on a surface [INS11].

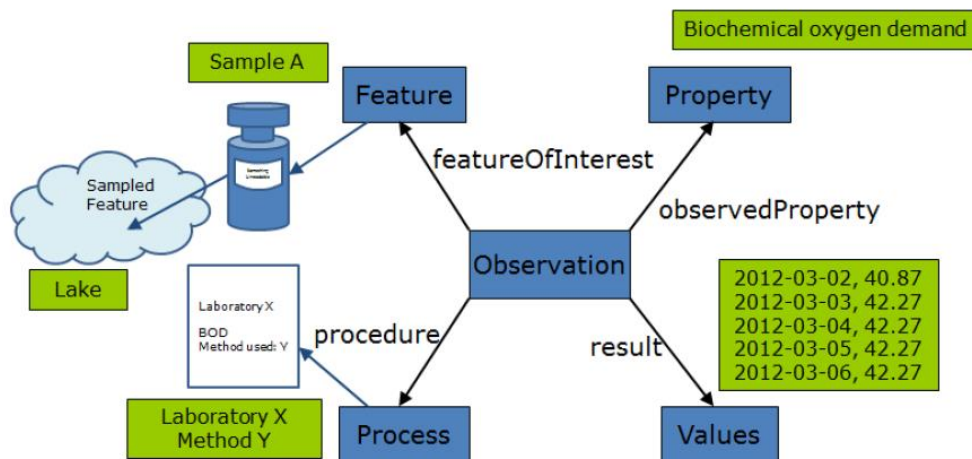


Figure 2.10: INSPIRE: feature-of-interest as specimen [INS11].

2.3 Other Work

Modelling the data before analysing it is becoming increasingly important to ensure that it is properly understood, stored and later processed [JALR14]. In Europe, an important share of marine geological and geophysical observations is collected and analysed by research institutions and national geological surveys. These often host substantial volumes of data collected by the industry, government departments, academia and environmental organizations [PDS12].

Several studies and projects [PDS12, Soe08, FFM, Soe08, Soe08, JZT, JALR14] have been developed to obtain data models and frameworks that can contain/manage data from geographic studies, that involves different sources and types of data.

2.3.1 Geo-Seas

The Geo-Seas [GS16] EU framework project is implementing an e-infrastructure of 26 marine geological and geophysical data centres. The objective of this project is to achieve interoperability with already existing data-sharing initiatives that are well established in other fields of marine science. At the same time it also aims to be compliant and boost the European INSPIRE directive [PDS12]. Geo-Seas has developed a metadata model which aims to address domain specific issues and enrich ISO19115⁴ metadata with scientifically-relevant content while at the same time providing interoperability with initiatives in other scientific disciplines [PDS12].

2.3.2 Baltic Operational Oceanographic System

The Baltic Operational Oceanographic System — BOOS [BOO] was initiated when the BOOS Memorandum of Understanding was signed in 2001. The members of BOOS are institutions from Sweden, Finland, Russia, Estonia, Latvia, Lithuania, Poland, Germany and Denmark, who are committed in national operational oceanographic studies. The objectives of BOOS are to provide an integrated service for the safe and efficient support of off-shore activities, environmental management, safety and sustainable use of marine resources [Soe08].

2.3.3 Water Framework Directive

The Water Framework Directive (WFD) [Com16a] aims to provide a common framework for the protection and improvement of surface waters (rivers, lakes, transitional waters and coastal waters) and groundwaters in Europe. The environmental data is managed through a Relational DataBase Management System (RDBMS) and a GIS Server [FFM]. Our solution is also based on a relational database.

⁴ISO 19115 provides metadata information about the identification, extent, quality, spatial and temporal aspects, content, spatial reference, portrayal, distribution, and other properties of digital geographic data and services [ISO14].

2.4 Marine Data Management Projects

Three major programs were initiated from the EU commission, including the INSPIRE directive, plans for the GMES Marine Core Service (MCS) and a new European recommendation to build an European Maritime Observation and Data Network (EMODNET) proposed by the Integrated Maritime Policy for the European Union [Soe08].

In the context of the 6th [Com15] and 7th [Com16b] framework of the European Union, there were several R&D projects funded with marine data management as main subject [Soe08]:

SeaDataNet [IFR] aims to develop a system that provides access to marine datasets and data products from 36 countries in Europe [Soe08].

ECOOP (European COastal-shelf sea Operational observing and forecasting system) is focused on the implementation and development of forecasting and observing systems that contain long-term systematic observations of marine parameters of the European Regional and coastal Seas [Soe08].

MyOcean [Ser] is the implementation project of the GMES Marine Core Service, aiming at deploying the first concerted and integrated pan-European capacity for ocean monitoring and forecasting.

MERSEA (Marine Environment and Security for the Open Sea) is an essential component of the Ocean and Marine services element of GMES [Soe08].

2.5 Time Series Conceptual Models

One of the most problematic types of geographic data is the time series, which is a set of repeated measurements taken on a single point in space [INS11]. Some conceptual models have been created and analysed to model this type of data. One of the studies suggests a multidimensional model of a data warehouse with data mining integration. They provide a UML profile for the modelling of the time series analysis [JZT]. Another work [JALR14] proposes a UML extension able to represent any set of structurally complex hierarchically ordered data. They also apply data mining techniques to the resulting data structure for knowledge discovery.

Time series are also part of the data types that our work contains. After studying the best way to model this type of data, we concluded that the use of existing and proven data models would be the best option, such as OGC SOS implementations like istSOS or the one from 52° North.

2.6 Summary

Throughout this chapter, we have described some directives, standards, services, frameworks, software, programs and proposals for data models. When it comes to modelling geographic data we see that there are many rules and policies to be met. While, in terms of implementation, this

may be more complex, it can bring several advantages, including the interoperability factor, which is very important in this type of data.

The main challenges we faced were the integration and modelling of the covered heterogeneous data. While there is work on how to model this type of data, some points were still open to improvement. The carried studies only covered particular cases as fish-captures in Alicante [JZT] or in stabilometry field [JALR14]. In this work, we present a data model created to encompass a variety of situations as the collection of fishing data, the *in situ* collection of biotic and abiotic data, laboratorial analysis data, photographs, videos and data collected from sensors (time series).

State of the Art

Chapter 3

Challenges of Marine Data Heterogeneity

In this chapter we develop the challenges of the marine data caused by the heterogeneity of the ocean research and, consequently of the types of data. Afterwards, we present a solution to this problem, an extension to the INSPIRE based data model.

3.1 Problem Description

"Humankind is extremely reliant on the oceans, as a source of food and raw materials, as a climate regulator, for transportation, for disposal of waste products, and for recreation. According to the UN Atlas of the Oceans, approximately 44 per cent of the worlds population live within 150 km of the coast and this is likely to increase. The marine environment provides immense opportunity in terms of economic potential, with activities such as shipping, oil and gas mining, fishing and aquaculture." [LSW06]

The marine environment is dynamic and multidimensional, constituting a difficult area for data collection and updating. Data is often collected on a project-based approach (e.g., BIOMETORE) and is rarely shared between different organizations. In other situations, data is collected under regular programs and, in this case, it is usually more widely distributed. However, in both cases data must be stored and distributed in an integrated way, according to standards. There is a lot of data collected on the marine and coastal environment, but it is often not available to all interested parties. While data availability is certainly an issue, it can be hindered by the lack of interoperability, without which access can be difficult and restricted [LSW06].

Regardless of the lack of data and the temporal and spatial heterogeneity of the data that are limiting the understanding of natural processes, the large amount of data makes the researchers

spend a lot of time in basic activities such as data gathering and transformations. It is also clear that storing data in spreadsheets or text documents is no longer a practical solution to manage a large number of data sets [FFM].

According to Fernandez et al. [FFM], implementing an environmental information system brings benefits like:

1. Centralized information management with support for concurrent users.
2. Centralized scientific knowledge.
3. Removal of the intermediate data gathering and maintenance responsibilities from scientist or water professionals.
4. Centralized homogeneity and data validation.

3.2 An Information System to Integrate Marine Data

"Collecting, organizing and manipulating environmental data under an environmental information system has enabled the analysis of data overcoming heterogeneity under an integrated framework. The application of Environmental European Directives at regional scale requires the ability to manage a large amount of data from operational monitoring networks, real time sensors and derived information obtained from analysis." [FFM]

As stated on Chapter 2.5, managing time series is a challenge. This type of data is represented at a low-level of abstraction and its management is expensive. According to Zubcoff et al. [JZT], most analysts face two major problems: (i) the cleaning of the large number of potentially analysable data; and (ii) the correct definition of data mining algorithms to be used. Also according to Zubcoff et al., in a time-series analysis time is the main variable used to describe the data that is being analysed. Nevertheless, a period of time can be recorded using different scales. For example, in a point-of-sales system, sales are saved using the more granular level of the timestamp. However, for the purpose of analysis, investigators can aggregate the data by day, week, month, etc., according to their needs. Time series analysis is a powerful technique to discover patterns and trends in temporal data. In spite of the gap between software solutions, there is a common technological trend, the use of relational databases to store and manage environmental information, and the use of web technologies to share and facilitate the access to information [FFM].

So, for the solution, we propose the development of a relational data model that will allow the integration of data collected and analysed in the scope of the campaigns (specimens, photographs, videos) and the data collected by the sensors (time series). This model must be developed to allow not only data storage but also its extraction for analysis and use in many types of studies. The model must also have characteristics that allows it to evolve and add new parameters, as well as a perspective for the preservation of data and INSPIRE compliance.

3.3 Summary

There is a huge amount of data collected by sensors and researchers about the marine waters and their ecosystems. However, a lot of this data is being lost because of the poor organization of the data, making it hard to be re-analysed and shared with other institutions or even departments. With this in mind, developing an information system to manage this type of data brings many advantages, even more if this system is compliant with European directives as INSPIRE. The first step to develop a good and functional information system is to build up a good model for the data. This model must ensure that all the proposed requirements are properly implemented.

Chapter 4

Repository Architecture

In this chapter, we present a brief description of the workflow of marine research. Then we present the developed data model, which has three main components: the 52° North implementation, the metadata model and the local data required for BIOMETORE-like campaigns.

4.1 Marine Research Data Workflow

The first stage of any campaign is its preparation, where the teams, forms and stations are decided, and the vessel is prepared. Next, the campaign itself is carried out, where data and metadata are collected and locally stored in a mobile device or computer (most of the times the devices will be offline). In the end of the campaign (if not sooner), the collected data is sent through the web services layer and it is deposited in the database. We can see this workflow in Figure 4.1.

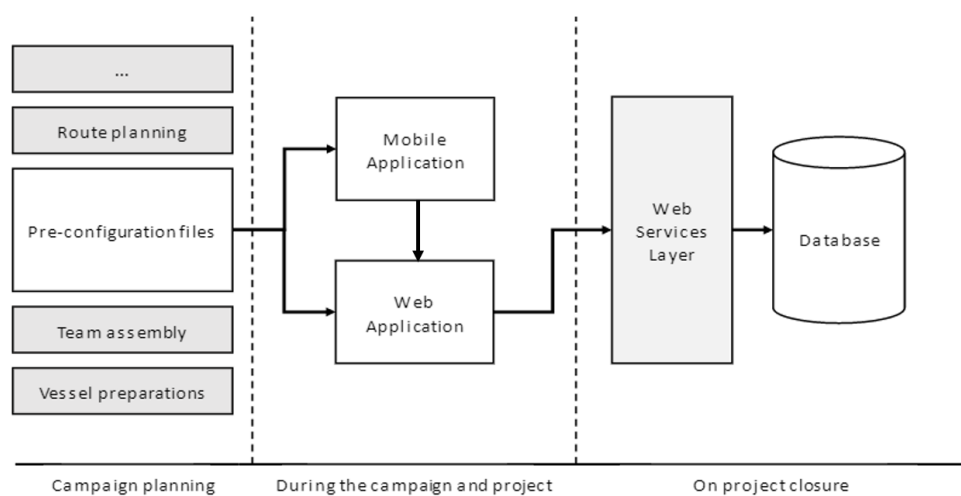


Figure 4.1: Representation of the different marine data collection stages and their interactions.

4.2 Data Model

All of the studies and work that have been described so far resulted in the data model shown in Figure A.1, in Appendix A. This model can be divided in three sections, the 52° North model, the metadata models of SNIMar and Darwin Core, and the local data from BIOMETORE. The 52° North part of the model is responsible for the storage of the data collected in the campaigns as well as the data produced in laboratory. The SNIMar metadata enables data relation in future analyses, describing the type of stored data, and the process used to collect it. The Darwin Core is used to store additional data about the identification of the species through a normalized taxonomic rank storage. The BIOMETORE local data is where everything related to the preparation of the campaigns and organisation of the IPMA data is stored. All three section are essential to the long term storage and analysis of the data.

4.2.1 52° North

After studying existing open source implementations, we chose the one from 52° North. We made this decision because 52° North implements the OGC SOS specification that relies on the OGC Observation & Measurements standard. Although the model is an implementation of the SOS, that only predicts sensor data, we decided to use the same model to also store other semantically comparable data collected in the campaign and produced in laboratory, mapping it as observations, and ensuring data integration simply by using the same structure.

Another reason for choosing 52° North is that it also implements the OGC SOS 2.0 specification that has four extensions: Core, Enhanced, Transactional and Result Handling, which in total equals to 13 operations [Hol16]. Also, 52° North is well known in the community, which facilitates the search for information and overall support.

The OGC O&M data model was already described in Chapter 2.2.1. Given the 52° North model is based on it, in this section we will only present a brief description of the main tables, based on the 52° North SOS data models [Hol12a].

Figure 4.2 shows the *observation* table. It aggregates the data for an observation, like time, procedure and feature-of-interest, and the observation value (or result) in separate tables. The result table from O&M is replaced by the value tables. The supported types of value are text, count, numeric, boolean, blob, category, geometry and data array.

In Figure 4.3 we show the *featureofinterest* table, representing the real world object to which the observation belongs. The geom column represents the geometry of the feature-of-interest; in particular it is represented using the PostGIS geometry type.

Figure 4.4 shows the *observableproperty* table. It references the phenomenon which has been observed. It stores its identifier, name and description. As seen in Chapter 2.2.1.1, the observed property can be, for example, a species identification, height, water temperature or biochemical oxygen demand. The table *compositephenomenon* allows us to create a parent-child connection between different observable properties.

Repository Architecture

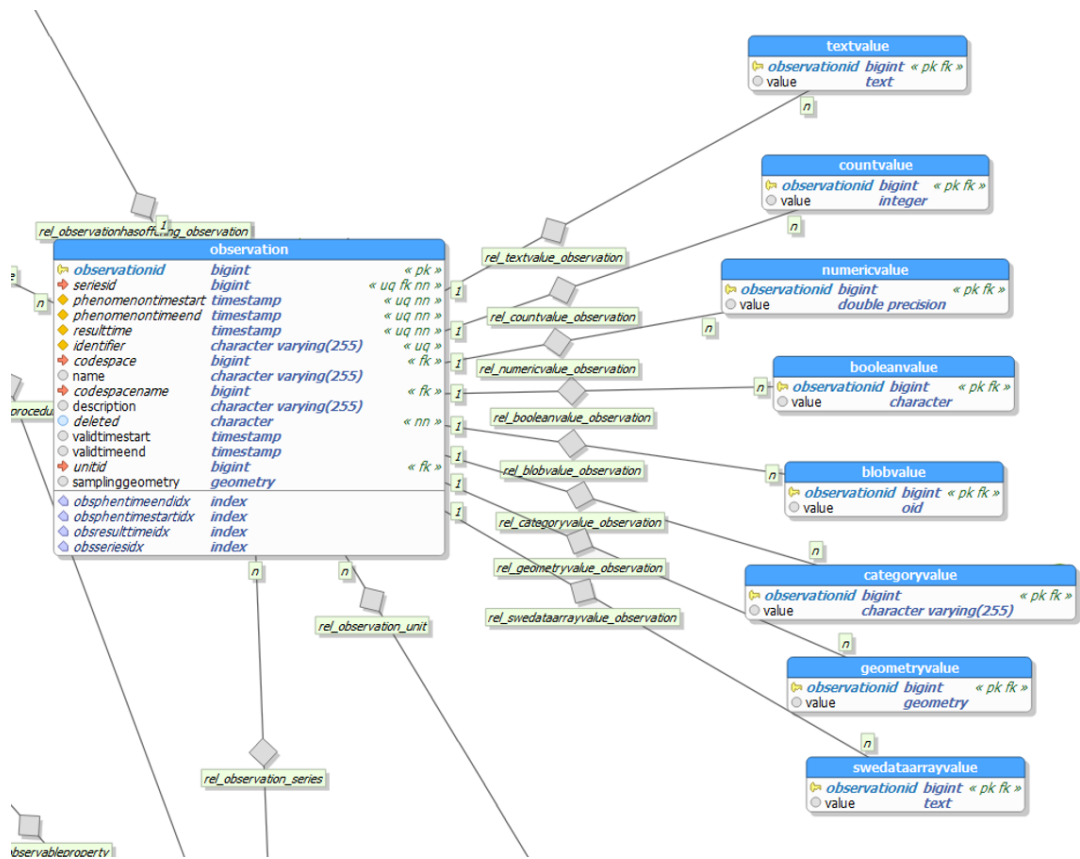


Figure 4.2: Table *observation* from the 52° North data model [Hol12b].

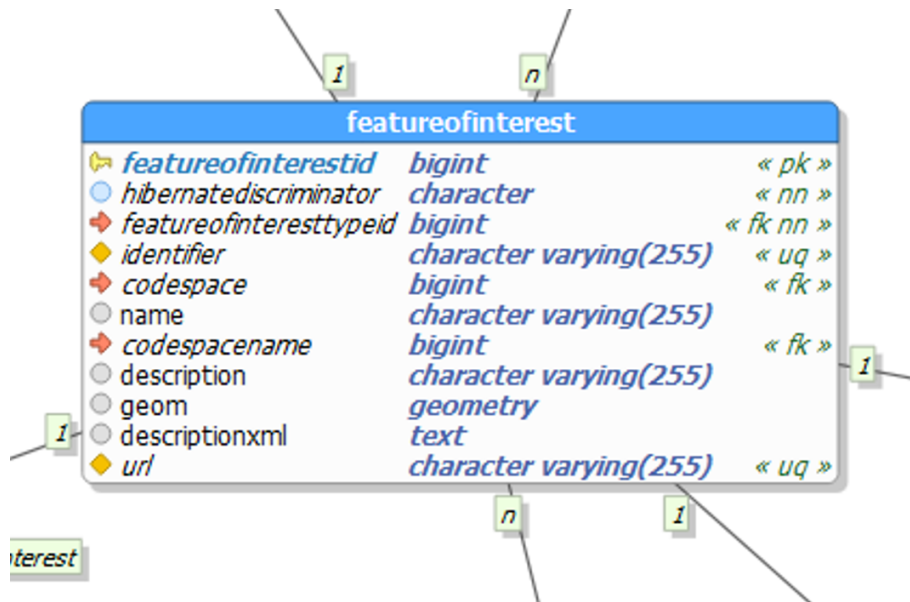


Figure 4.3: Table *featureofinterest* from the 52° North data model [Hol12b].

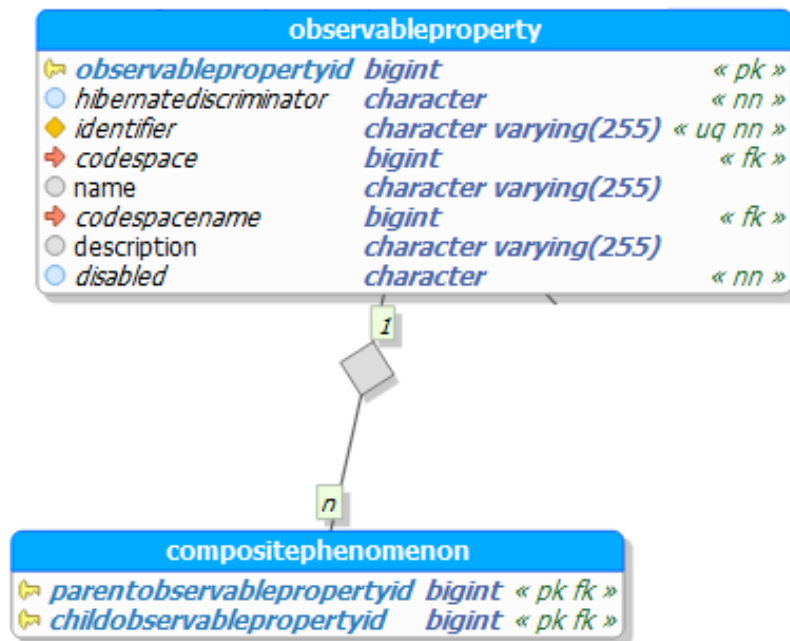


Figure 4.4: Table *observableproperty* from the 52° North data model [Hol12b].

Figure 4.5 shows the *procedure* table. It stores data about the procedure used to collect/analyse the data, like its identifier, name, description and a possible description file. The procedure can be a sensor or a method like the "Magic Triangulation".

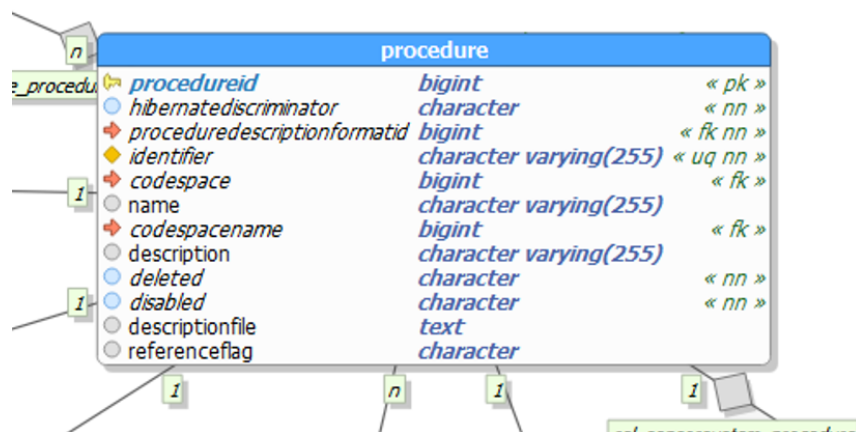
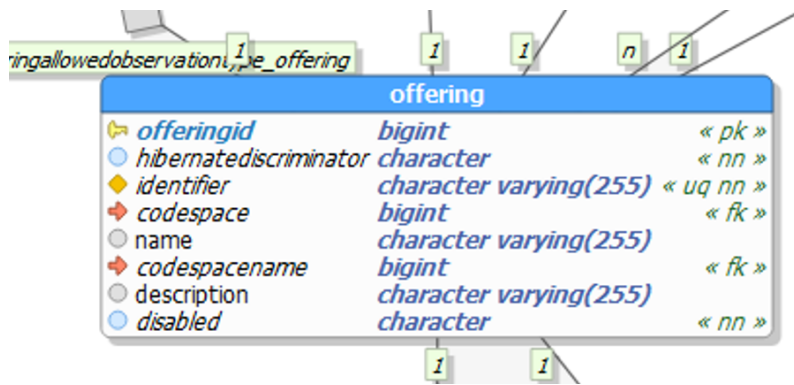
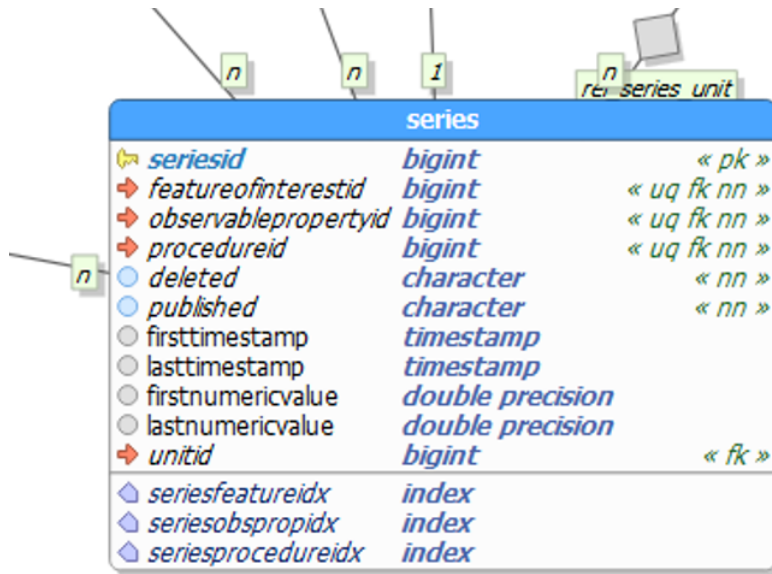


Figure 4.5: Table *procedure* from the 52° North data model [Hol12b].

The table from Figure 4.6 is an extension to the OCG O&M data model available in 52° North. The *offering* allows the grouping of observations according to certain thematic criteria, for example grouping per country or region. It is connected to the tables *featureofinterest* and *observation*.

Figure 4.6: Table *offering* from the 52° North data model [Hol12b].

Finally, the table from Figure 4.7 is also an extension. The *series* table is where we store the information from the time series. Its attributes include first time stamp, last time stamp, first numeric value and last numeric value. This table also has another function: it combines the tables *featureofinterest*, *observableproperty* and *procedure*, enabling information cross-referencing, not only with time series, but also with other types of collected data.

Figure 4.7: Table *series* from the 52° North data model [Hol12b].

The operations of the four extensions of the OGC SOS 2.0 specification, available through the 52° North API (Figure 4.8) are described by Carsten Hollmann [Hol16]. This specification has small differences from the one from Chapter 2.1.6.

Home
Client
Documentation
Admin

52°North SOS Test Client

Choose a request from the examples or write your own to test the SOS.

Examples

NOTE: Requests use example values and are not dynamically generated from values in this SOS. Construct valid requests by changing request values to match values in the Capabilities response.

NOTE: For security reasons, the transactional SOS operations are disabled by default and the *Transactional Security* is activated by default with allowed IPs *127.0.0.1*. The transactional operations can be activated in the [Operations settings](#) and the *Transactional Security* can be deactivated in the [Transactional Security tab of the settings](#).

SOS
2.0.0
JSON
GetCapabilities

Load a example request ...

Service URL

http://sensorix.inesporto.pt:8080/52n-sos-webapp/service

Request

POST
application/json
application/json

```

1 {
2   "request": "GetCapabilities",
3   "service": "SOS",
4   "sections": [
5     "ServiceIdentification",
6     "ServiceProvider",
7     "OperationsMetadata",
8     "FilterCapabilities",
9     "Contents"
10  ]
11 }

```

Send

Response

```

200 OK
Date: Thu, 02 Jun 2016 13:12:59 GMT
Server: Apache-Coyote/1.1

```

Figure 4.8: 52° North API console.

1. Core

`GetCapabilities()` Request a self-description of the service;

`GetObservation()` Request observation data encoded in O&M;

`DescribeSensor()` Request information about a certain sensor, encoded in SensorML.

2. Enhanced

`GetFeatureOfInterest()` Request an encoded representation of the feature that is the target of the observation;

`GetObservationById()` Request data of a specific observation.

3. Transactional

`InsertSensor()` Publish new sensors;

34

Repository Architecture

`UpdateSensorDescription()` Alter the values that describe a sensor;

`DeleteSensor()` Remove the record of a given sensor;

`InsertObservation()` Publish observations.

4. Result Handling

`InsertResultTemplate()` Add a result template to the server that describes the structure of the values to be inserted;

`InsertResult()` For uploading raw values according to the structure defined in the *InsertResultTemplate*;

`GetResultTemplate()` Obtain the result structure and encoding;

`GetResult()` Obtain the raw data for specific parameters.

Besides these operations, they implement two more:

`GetDataAvailability()` Enables the filtering of existing time series by their identifying metadata fields (as procedure, feature-of-interest and observed property) and the offering they may belong to;

`DeleteObservation()` Removes the specified observation.

4.2.2 SNIMar and Darwin Core

In order to maximize data interoperability, we used two more standards in addition to INSPIRE and OGC: SNIMar for the metadata about the services and datasets description, and Darwin Core for the metadata about the taxonomic ranks of the species.

SNIMar is a project within the European Economic Area Grants (EEA Grants) funded through the Financial Mechanism of the European Economic Area (FM EEA). It is based on the National Metadata Profile of Geographic Information (Perfil Nacional de Metadados de Informação Geográfica, Perfil MIG), which in turn is based on the ISO 19115, ISO 19119 and on the requirements of the INSPIRE directive [Sis15].

SNIMar [Sis15] declares that metadata about geographic information is no more than a textual description, in a standardized form, of geographic information. Its documentation is essential for the identification and technical assessment of spatial datasets, as well as aspects related to the access and usage of services for geographic information.

In Figure 4.9, we show the metadata model based on the SNIMar document [Sis15]. It has two main tables, the *md_dataidentification* and the *sv_serviceidentification*. As the names suggest, the first stores information about the data (datasets or dataset series) and the second about the services. We did not model all the fields required by SNIMar, given that they will be made available in XML format, through the `GetCapabilities()` operation of the 52° North API. We simply added some of the most relevant fields that might be queried in the future. Next, we describe each table:

Repository Architecture

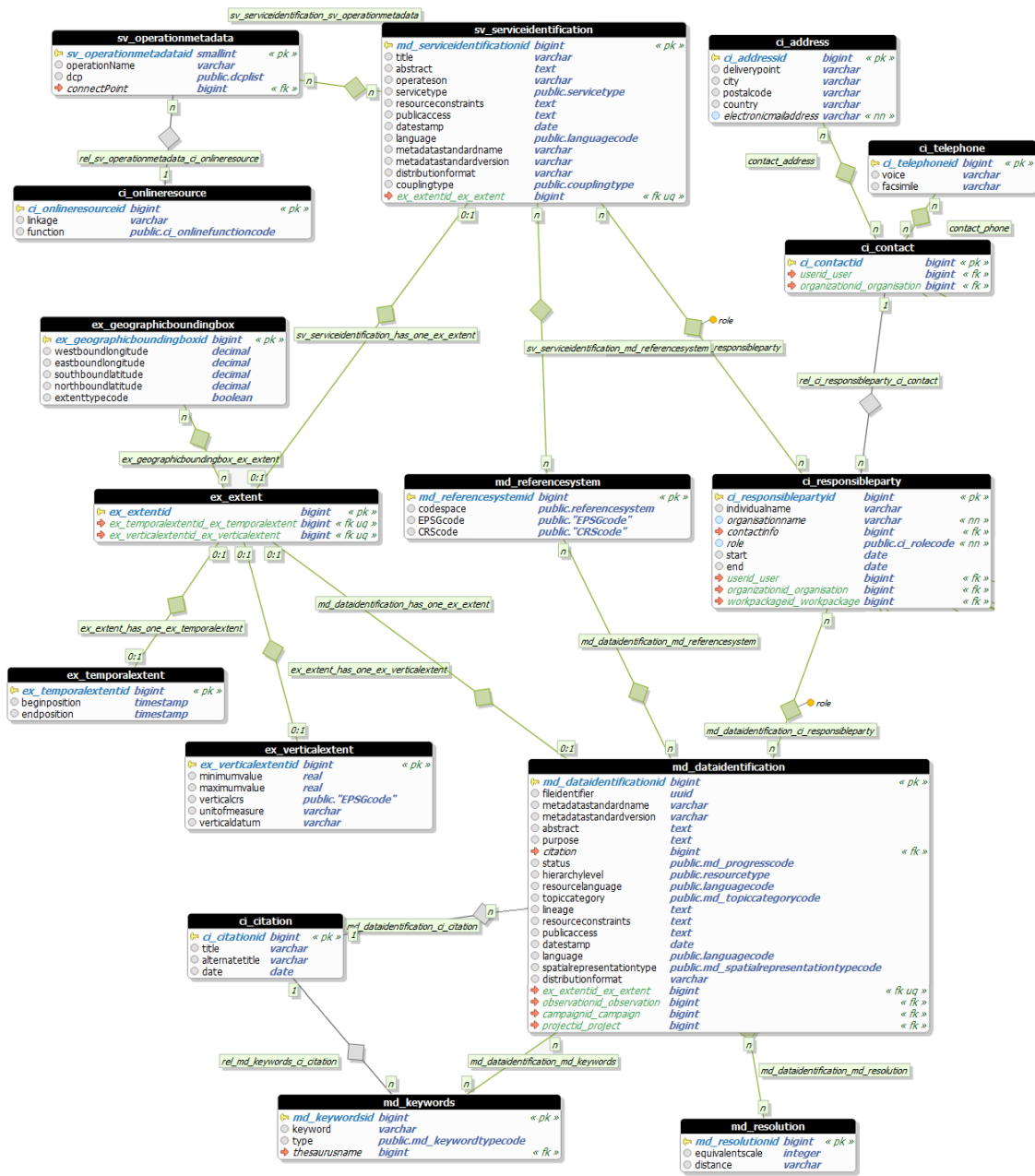


Figure 4.9: Metadata model based on SNIMar.

md_dataidentification

The table *md_dataidentification* describes the datasets or the dataset series. It has a citation of the resource, spatial resolutions (*md_resolution*), information about the reference systems (*md_referencesystem*), information about the spatial extent (*ex_geographicboundingbox*), descriptive keywords (*md_keywords*) and responsible parties (*ci_responsibleparty*) with a specific role as point of contact. In its attributes it stores information like an abstract, purpose, status (as completed, obsolete, ongoing and required), topic category (as biota,

boundaries and elevation), lineage with information about the acquisition processes or the databases used on the construction of the data, datestamp (date of creation or last update), spatial representation type (as vector, grid, video or text table) and resource constraints that are restrictions and legal pre-requisites for accessing and using the resource.

sv_serviceidentification

The table *sv_serviceidentification* describes the services. Like the *md_dataidentification*, it has information about the reference systems (*md_referencesystem*), the spatial extent (*ex_geographicboundingbox*) and responsible parties (*ci_responsibleparty*). Furthermore, it has information about the operations available in the service (*sv_operationmetadata*). In its attributes it stores information like the service type (as discovery, view or download), the associated resource under *operateson* and the coupling type of the services with the datasets (as loose, tight or mixed).

md_keywords

The *md_keywords* stores keywords that characterizes the resource, its type and reference.

ex_extent

The *ex_extent* describes the extension of the resource by means of geographic extent (*ex_geographicboundingbox*), temporal extent (*ex_temporalextent*) and vertical extent (*ex_verticalextent*).

ex_geographicboundingbox

The *ex_geographicboundingbox* describes the geographic position of the resource in a square area. The *extenttypecode* is a boolean, where true means that exists data within the given area and false that doesn't exist data within the given area.

ex_temporalextent

The *ex_temporalextent* is the time period for which the spatial data is valid.

ex_verticalextent

The *ex_verticalextent* is the altimetry extension of the spatial dataset.

md_referencesystem

The *md_referencesystem* stores information about the reference system used in the resource. It can be an EPSG Geodetic Parameter Set, or CRS (Coordinate Reference System).

md_resolution

The *md_resolution* stores a detailed level of a spatial dataset expressed on a scale factor or distance on the ground.

ci_citation

The *ci_citation* stores a citation to reference the resource. It has a title, alternate title and date.

sv_operationmetadata

The *sv_operationmetadata* stores information on all operations provided by the service. It has the operation name, the Distributed Computing Platforms (DCP) within a code list and the connect point. This table will be used as the basis for the catalogue of the services.

ci_onlineresource

The *ci_onlineresource* stores information on online resources. It has the link to the resource and its function (as download, information or search).

ci_responsibleparty

The *ci_responsibleparty* stores information about the person or organisation responsible for the resource or metadata.

Darwin Core is a set of standards and includes a glossary of terms intended to facilitate the sharing of information about biological diversity. It is primarily based on taxa and its occurrence in nature is documented by observations, samples, specimens and related information [Sta15].

In Figure 4.10, we used the Darwin Core standard as a basis for the data model of the taxonomic rank of the species:

kingdom (*Animalia*).

phylum (*Arthropoda*).

class (*Malacostraca*).

order (*Amphipoda*).

family (*Melitidae*).

genus (*Abludomelita*).

taxonomicclassification

The table *taxonomicclassification* stores the remaining necessary information about the taxonomic rank. This includes the scientific name (*Abludomelita obtusata*), the author and date of the name (*Montagu, 1813*), the taxonomic status (as accepted, valid, synonym or misapplied), the parent name usage (the taxon name, of the direct and closer higher-rank parent taxon) and taxon remarks, as comments or notes about the taxon name. It also has the GUID for the taxonomic portal EU-NOMEN¹.

¹<http://www.eu-nomen.eu/portal/taxon.php?GUID=urn:lsid:marinespecies.org:taxname:102788>

Repository Architecture

The taxonomic rank information is linked to the 52° North model through the result category value.

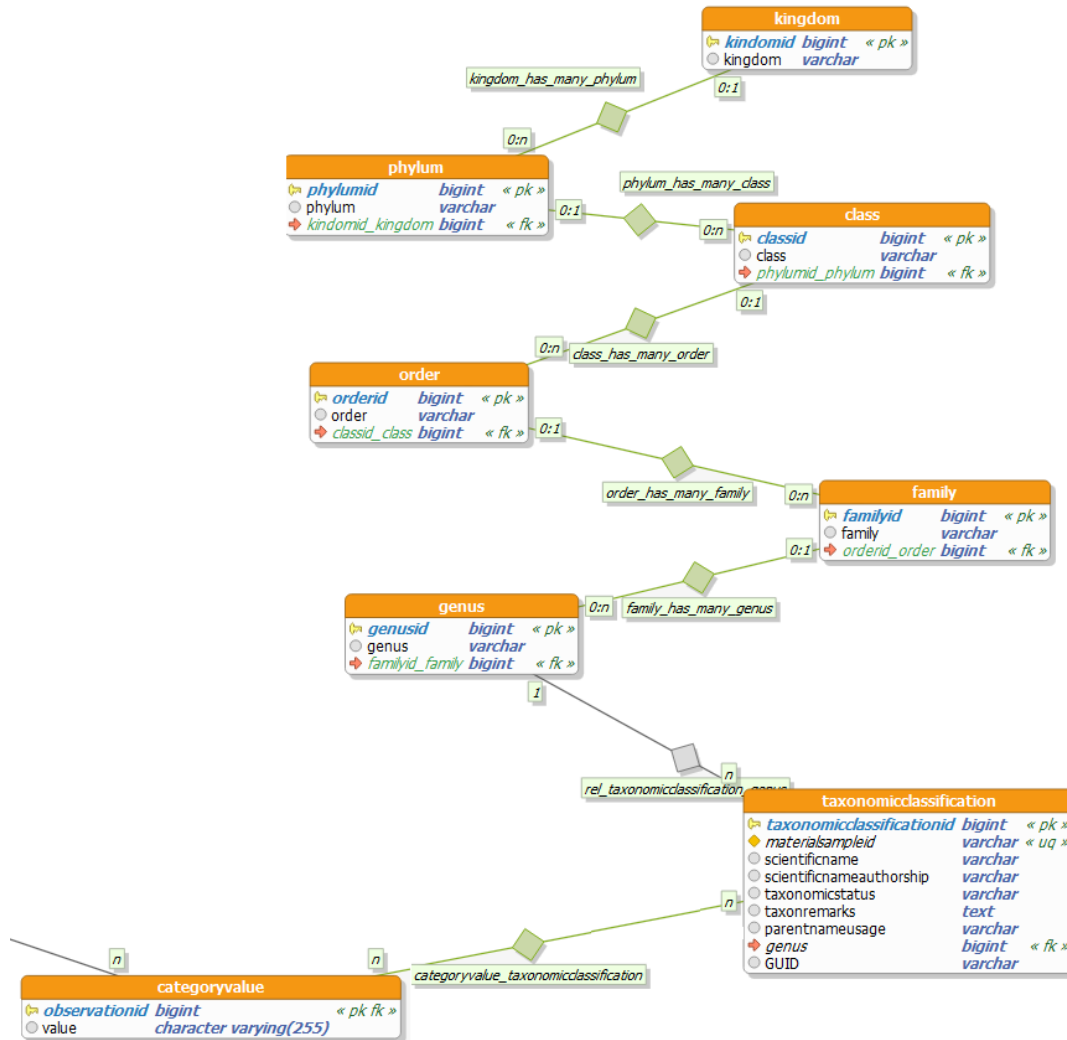


Figure 4.10: Taxonomic Rank model based on Darwin Core.

4.2.3 Project Local Data

So far, the presented data models focus on the observations, their data, procedures, results, feature-of-interest, observable property and metadata. Given that certain metadata is missing for the BIOMETORE, we have developed an extension of the original model from 52° North. This enables the storage of data about the users, contacts, responsible parties, organization, campaigns, vessels, stations and forms.

Repository Architecture

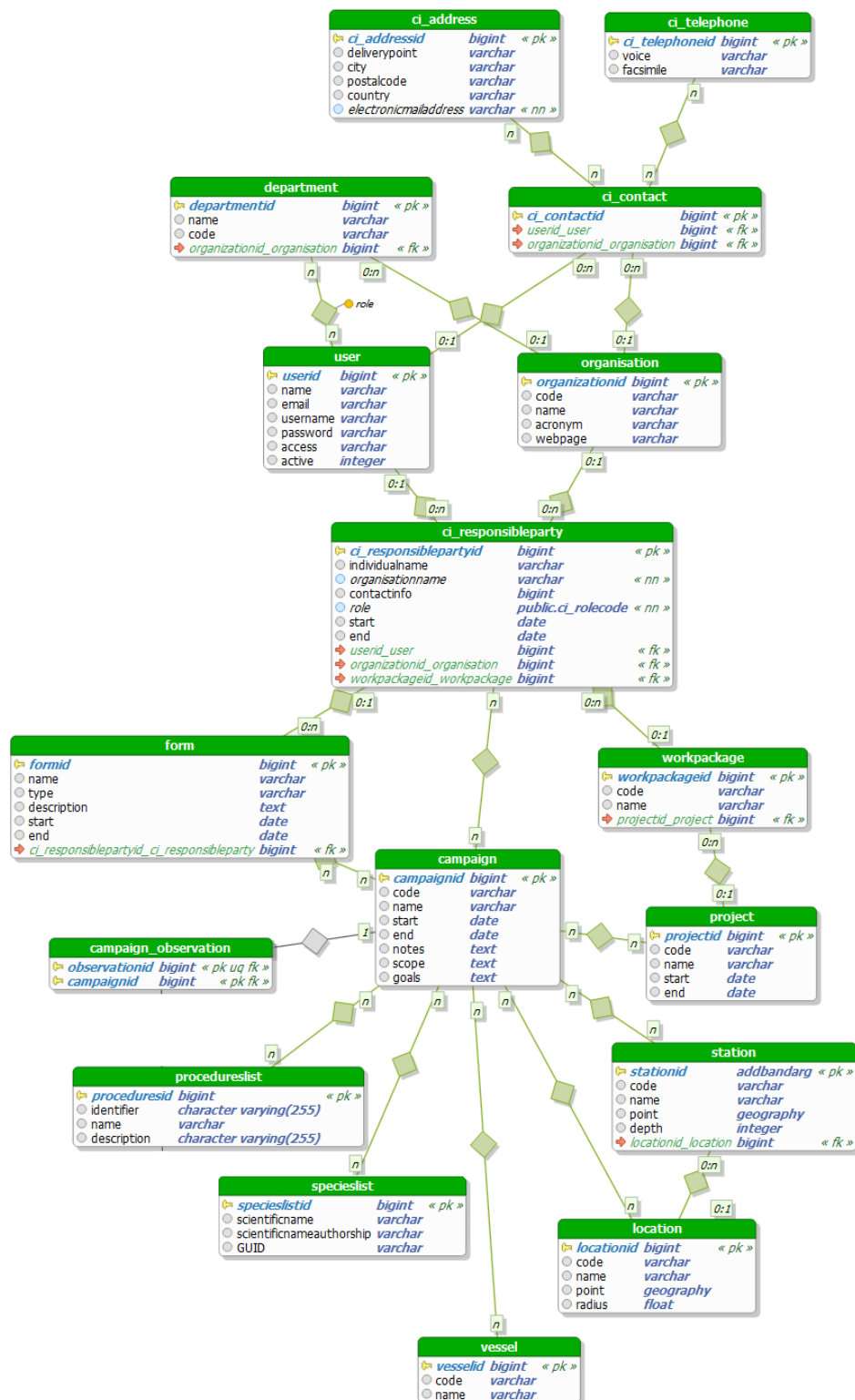


Figure 4.11: BIOMETORE local database model.

This model has two main tables: the *ci_responsibleparty* and the *campaign*. The *ci_responsibleparty* represents the responsible entity (person or organisation) for the campaign, observation, form and procedure. The *campaign* belongs to a project, has stations, vessels, forms and locations.

Next, we explain each table:

ci_responsibleparty

As previously stated, this table can either represent a user or an organization, which is identified by the presence or absence of the *individualname* attribute — if it is filled, then it is a user. Regardless of being a user or organisation, the responsible party always has an organization name (if it is a user then the organization name represents his/her affiliation). The role of the responsible party in the campaign is also stored (for example, *pointOfContact*), as well as the start and end dates of its involvement in the campaign. It is also from this table that is possible to retrieve the teams, by joining the campaign, start and end dates of the responsible party and the work package.

user

The *user* represents a person and user of the system. It has contacts and belongs to departments of an organisation.

organisation

The *organisation* has departments and contacts. Its attributes are a code, name, acronym and webpage.

department

The *department* belongs to an organisation and has several users with specific roles. Its attributes are a name and a code.

workpackage

The *workpackage* represents the work package where the responsible party belongs to. This is used to manage data access and to ensure that a user can only access data from his work package.

ci_contact

The *ci_contact* contains multiple addresses and telephones.

ci_address

The *ci_address* stores the delivery point, the city, postal code, country and e-mail address.

ci_telephone

The *ci_telephone* stores the voice and fax number.

project

The table *project* represents the highest level of granularity of the model, given almost every entity is under a particular project. It has a code, name and start and end dates.

campaign

After the project, *campaign* is next in granularity level. It is within a campaign that observations are stored. The campaign has projects, stations, locations, vessels, forms and responsible parties. In its attributes it has a code, name, goals, scope and a start and end dates.

form

The table *form* stores information related to the forms used in the campaigns. They are also associated with the observation where they were used and the procedures performed in their filling.

vessel

The *vessels* represents the ships used in the campaigns. Not all campaigns use vessels.

station

The *station* represents the point where a sample took place. It has a code, name, geographic point and depth. The stations can be pre-planned but their point and depth must be updated with real, more accurate, values after the sample has been collected.

location

The *location* is a list of actual locations, that won't probably change. Like the "Great Meteor" that is a sea-mount with a specific geographic point and area.

specieslist

The *specieslist* is used in the filling of the forms during the campaign. Thereby, the filling work of the researchers is minimized. It has the scientific name of the species, the author that had identified it and the GUID for the EU-NOMEM portal.

procedurelist

The *procedurelist* is used in the pre-planning of the campaign, when making new forms and choosing the procedures. It is also used during the campaign in a similar fashion to the *specieslist*. It has the identifier of the procedure, its name and description.

4.3 Summary

We developed a relational data model with three main parts: (1) the model from 52° North, where the data from the observations is going to be stored; (2) the metadata model of SNIMar to describe

Repository Architecture

the datasets and services, and the Darwin Core standards to normalize the information about taxonomic ranks; (3) a data model for the project local data, including data about users, campaigns, vessels, stations and forms. Combining these three sets we built a robust model able to stock all the data required by IPMA.

Chapter 5

Discussion

In this chapter, we analyse the mapping of the data to the model, with the goal of illustrating practical applications of our data model, while discussing decisions, strengths and difficulties that we have identified.

5.1 Mapping the Data

In order to demonstrate the good fitting between our heterogeneous data and our relational data model, we use the test data provided by IPMA, in various distinct formats. Next, we discuss how the data is mapped and provide examples of the most challenging parts. A few columns were hidden since are only complementary to the model and are not required to understand it.

5.1.1 Observations Data

In this section, we discuss the data mapping to the 52° North model. As stated in previous chapters, it is in this part of the model that the observations data (campaign and laboratory) is stored. To do the mapping we used files provided by IPMA such as benthos and CTD files. The benthos are spreadsheets with data collected both on campaigns and laboratory. The remaining spreadsheets contain analogous data. We also used a CTD file to do the mapping. All other CTD files have the same structure since they only diverge on the observable properties.

5.1.1.1 Test Case 1

In Figure 5.1, we show an example of spreadsheet data (Figure 6.1.1) with information about the campaign, including station, latitude, longitude, procedure and notes. In this particular case, we extract and load to the 52° North model the latitude and longitude, the date and time and the procedure used to collect the data. Figure 6.1.2 shows a table with the result of a taxonomic observation made in laboratory on the sample collected on D1 station. Figure 6.1.3 shows the

Discussion

Estação	Amostra	Data	Hora	LAT N	LONG W	Equipamento colheita	Notas
D1	DA1	19/09/2015	13:05	37°55,881'	8°50,279'	Smith-McIntyre 0,1m2	Sines 150042; Rumo a D1- Sines 141927.HSX. 2 Fotos junto ao porto

Figure 6.1.1: Data collected *in situ*

Estação	Phylum	Classe	Ordem	Família	Gênero	Espécie	Autor
D1	Mollusca	Bivalvia	Veneroida	Mactridae	Mactra	Mactra stultorum	Linnaeus, 1767
D1	Mollusca	Bivalvia	Veneroida	Montacutidae	Kurtiella	Kurtiella bidentata	Gofas & Salas, 2008
D1	Arthropoda	Malacostraca	Amphipoda	Ampeliscaidae	Ampelisca	Ampelisca spooneri	Kroyer, 1842

Figure 6.1.2: Taxonomic ranking (made in laboratory)

Estação	Data	Pcad vazio(g)	Pcad+Sed seco a 100°C(g)	Pcad+Sed seco a 450°C(g)
D1	19/09/2015	11,9056	12,1150	12,1112

Figure 6.1.3: First measures to calculate MOT (made in laboratory)

Figure 5.1: Benthos spreadsheet data example.

measurement of Pcad, which is the crucible¹ weight, in several phases in order to calculate the MOT (Total Organic Matter).

In Figure 5.2, we show the table *observation* filled with the previous spreadsheet data. It includes the phenomenon times, name, description, result time, among others. Please note the connection to the *series* table through the *seriesid*.

observation									
observationid	seriesid	phenomenon timestart	phenomenon timeend	resulttime	identifier	codespace	description	unitid	sampling geometry
1	1	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1	http://www.opengis.net/d ef:nil/OGC/0/unknown	Preliminary observation	1	38.1613888888889 -8.91083333333334
2	2	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_1_1	http://www.opengis.net/d ef:nil/OGC/0/unknown	Taxonomic Mactra	1	38.1613888888889 -8.91083333333334
3	3	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_1_2	http://www.opengis.net/d ef:nil/OGC/0/unknown	Taxonomic Kurtiella	1	38.1613888888889 -8.91083333333334
4	4	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_1_3	http://www.opengis.net/d ef:nil/OGC/0/unknown	Taxonomic Ampelisca	1	38.1613888888889 -8.91083333333334
5	5	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_2_1	http://www.opengis.net/d ef:nil/OGC/0/unknown	MOT Pcad Vazio	2	38.1613888888889 -8.91083333333334
6	6	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_2_2	http://www.opengis.net/d ef:nil/OGC/0/unknown	MOT Pcad + Sed seco 100°	2	38.1613888888889 -8.91083333333334
7	7	2015-09-19 T13:05	2015-09-19 T13:05	2015-09-19 T13:05	http://inescporto.pt/ observation/DA1_2_3	http://www.opengis.net/d ef:nil/OGC/0/unknown	MOT Pcad + Sed seco 450°	2	38.1613888888889 -8.91083333333334

Figure 5.2: Mapped benthos data for the *Observation* table.

Analysing the table, we can see that all observations belong to different series (Figure 5.3), but, as we will describe in the next test case, several observations can belong to the same serie. This is determined upon insertion via the selection of the proper API request. The data can be inserted, as independent observations, using the `InsertObservation()` function or it can be

¹"A crucible is a heat-resistant container in which materials can be heated to very high temperatures." [Wik15]

Discussion

inserted using the `InsertResult()`, based on a result template for the registration of several observations in one request. For the first test case, we used the `InsertObservation()`, due to the different observable properties (Figure 5.4).

Series										
seriesid	featureof interestid	observable propertyid	procedureid	deleted	published	firsttime satmp	lasttime stamp	fsrtnumeric value	lastnumeric value	unitid
1	1	1	1	0	0	2015-09-19 T13:05	2015-09-19 T13:05			1
2	2	2	2	0	0	2015-09-19 T13:05	2015-09-19 T13:05			1
3	2	2	2	0	0	2015-09-19 T13:05	2015-09-19 T13:05			1
4	2	2	2	0	0	2015-09-19 T13:05	2015-09-19 T13:05			1
5	3	3	3	0	0	2015-09-19 T13:05	2015-09-19 T13:05	11,9056	11,9056	2
6	3	4	3	0	0	2015-09-19 T13:05	2015-09-19 T13:05	12,1150	12,1150	2
7	3	5	3	0	0	2015-09-19 T13:05	2015-09-19 T13:05	12,1112	12,1112	2

Figure 5.3: Mapped benthos data for the *Series* table.

observableproperty					
observable propertyid	identifier	codespace	name	codespacename	description
1	http://inescporto.pt/ observableProperty/1	http://www.opengis.net/ def:nil/OGC/0/unknown	Benthos_analysis	http://www.opengis.net/def/ nil/OGC/0/unknown	Benthos Analysis
2	http://inescporto.pt/ observableProperty/1	http://www.opengis.net/ def:nil/OGC/0/unknown	species_id	http://www.opengis.net/def/ nil/OGC/0/unknown	Species identification
3	http://inescporto.pt/ observableProperty/2_1	http://www.opengis.net/ def:nil/OGC/0/unknown	pcad_vazio	http://www.opengis.net/def/ nil/OGC/0/unknown	Pcad vazio
4	http://inescporto.pt/ observableProperty/2_2	http://www.opengis.net/ def:nil/OGC/0/unknown	pcad_sed_seco_100	http://www.opengis.net/def/ nil/OGC/0/unknown	Pcad+Sed seco a 100°C
5	http://inescporto.pt/ observableProperty/2_3	http://www.opengis.net/ def:nil/OGC/0/unknown	pcad_sed_seco_450	http://www.opengis.net/def/ nil/OGC/0/unknown	Pcad+Sed seco a 450°C

Figure 5.4: Mapped benthos data for the *observableproperty* table.

In Figure 5.5, we show the results of the observations. For the blue rows, the results are category values and, for the orange rows, the results are numeric values.

categoryvalue		numericvalue	
observationid	value	observationid	value
2	Macrta stultorum	5	11,9056
3	Kurtiella bidentata	6	12,1150
4	Ampelisca spooneri	7	12,1112

Figure 5.5: Mapped benthos data for the results tables.

Discussion

5.1.1.2 Test Case 2

In Figure 5.6 we show an example of a CTD file, its header and readings. In this case, we extract to the 52^o North the procedure (name of the sensor), timestamp, latitude, longitude, name of the variables read and its results.

Sea-Bird SBE 9 Data File:			
1	* FileName = E:\02020513\I007.dat		
2	* Software Version Seasave Win32 V 5.39c		
3	* Temperature SN = 4849		
4	* Conductivity SN = 2902		
5	* Number of Bytes Per Scan = 30		
6	* Number of Voltage Words = 4		
7	* Number of Scans Averaged by the Deck Unit = 1		
8	* System UpLoad Time = May 19 2013 19:38:46		
9	** Ship: Noruega		
10	** Cruise: 02020513		
11	** Station: I007		
12	** Lat: 40 00.0 N		
13	** Long: 09 37.0 W		
14	** Date (dd-mm-YYYY): 19-05-2013		
15	** Time (hh:mm): 19:40		
16	# nscan = 9		
17	# nvalues = 34690		
18	# units = specified		
19	# name 0 = prDM: Pressure, Digiquartz [db]		
20	# name 1 = t090C: Temperature [ITS-90, deg C]		
21	# name 2 = c0S/m: Conductivity [S/m]		
22			
214	# datcnv_skipover = 0		
215	# file_type = ascii		
216	*END*		
217	1.191	13.0194	-0.000009
218	1.067	13.0196	-0.000015
219	1.191	13.0194	-0.000012
220	1.133	13.0193	-0.000009
221	1.191	13.0193	-0.000009
222	1.067	13.0194	-0.000015
223	1.191	13.0193	-0.000015
224	1.133	13.0193	-0.000015
225	1.133	13.0191	-0.000015
226	1.191	13.0193	-0.000015
227	1.133	13.0191	-0.000015
228	1.067	13.0191	-0.000007
229	1.258	13.0191	-0.000001
230	1.067	13.0190	0.000002
231	1.191	13.0190	-0.000007
232	1.133	13.0188	-0.000018
233	1.133	13.0188	-0.000018
234	1.191	13.0186	-0.000015
235	1.133	13.0188	-0.000015

Figure 5.6: CTD data example.

Figure 5.7, shows the same *observation* table, but now storing the data from the CTD file.

observation									
observationid	seriesid	phenomenon timestart	phenomenon timeend	resulttime	identifier	codespace	description	unitid	sampling geometry
8	8	2013-05-19 T19:38:46	2013-05-20 T03:40:26	2013-05-20 T03:40:26	http://inescporto.pt/ observation/I0071_1	http://www.opengis.net/ def:nil/OGC/0/unknown	CTD prDM	4	40 -9.616666666666667
9	8	2013-05-19 T19:38:46	2013-05-20 T03:40:26	2013-05-20 T03:40:26	http://inescporto.pt/ observation/I0071_2	http://www.opengis.net/ def:nil/OGC/0/unknown	CTD t090C	3	40 -9.616666666666667
10	8	2013-05-19 T19:38:46	2013-05-20 T03:40:26	2013-05-20 T03:40:26	http://inescporto.pt/ /observation/I0071_3	http://www.opengis.net/ def:nil/OGC/0/unknown	CTD c0S/m	5	40 -9.616666666666667

Figure 5.7: Mapped CTD data for the *observation* table.

In the second test case, given the regularity of the data, we were able to create a result template (Figure 5.10) to facilitate data loading. So we added a general observed property ("CTD_Output", Figure 5.9) and a single entry in the *series* table (Figure 5.8). Next, we connected the observable properties, within the CTD file, to the observed properties table using the composite phenomenon to establish parent-child relationships.

series										
seriesid	featureof interestid	observable propertyid	procedureid	deleted	published	firsttime satmp	lasttime stamp	fsrstnumeric value	lastnumeric value	unitid
8	4	6	3	0	0	2013-05-19 T19:38:46	2013-05-20 T03:40:26	1.191	-0.000015	1

Figure 5.8: Mapped CTD data for the *Series* table.

Discussion

observableproperty					
observablepropertyid	identifier	codespace	name	codespacename	description
6	http://inescporto.pt/observableProperty/3	http://www.opengis.net/def/nil/OGC/0/unknown	CTD output	http://www.opengis.net/def/nil/OGC/0/unknown	CTD output
7	http://inescporto.pt/observableProperty/3_1	http://www.opengis.net/def/nil/OGC/0/unknown	prDM	http://www.opengis.net/def/nil/OGC/0/unknown	Pressure, Digiquartz
8	http://inescporto.pt/observableProperty/3_2	http://www.opengis.net/def/nil/OGC/0/unknown	t090C	http://www.opengis.net/def/nil/OGC/0/unknown	Temperature
9	http://inescporto.pt/observableProperty/3_3	http://www.opengis.net/def/nil/OGC/0/unknown	c0S/m	http://www.opengis.net/def/nil/OGC/0/unknown	Conductivity

Figure 5.9: Mapped CTD data for the *observableproperty* table.

resulttemplate				
resulttemplateid	offeringid	observablepropertyid	resultstructure	resultencoding
1	3	6	<pre> "resultStructure": { "fields": [{ "type": "time", "name": "phenomenonTime", "definition": "ww.opengis.net/def/property/OGC/0/PhenomenonTime", "uom": "http://www.opengis.net/def/uom/ISO-8601/0/Gregorian" }, { "name": "CTD_output", "definition": "http://inescporto.pt/observableProperty/3", "fields": [{ "type": "quantity", "name": "prDM", "definition": "http://inescporto.pt/observableProperty/3_1", "uom": "db" }, { "type": "quantity", "name": "t090C", "definition": "http://inescporto.pt/observableProperty/3_2", "uom": "deg_C" }, { "type": "quantity", "name": "c0S/m", "definition": "http://inescporto.pt/observableProperty/3_3", "uom": "S/m" }] }] } </pre>	<pre> "resultEncoding": { "tokenSeparator": ";", "blockSeparator": "##" } </pre>

Figure 5.10: Mapped CTD result template for the *resulttemplate* table.

In Figure 5.11, we show the results for the readings of the sensors, per observation.

Thus we have: for the first test case, with two observations each and three observable properties, six entries in the *observation* table plus six entries in the *series* table; for the second test case with one observation and three observable properties, three entries in the *observation* table and one entry in the *series* table.

Discussion

numericvalue	
observationid	value
8	1.191
8	1.067
8	1.191
8	1.133
8	1.067
9	13.0194
9	13.0196
9	13.0194
9	13.0193
9	13.0193
10	-0,000009
10	-0,000015
10	-0,000012
10	-0,000009
10	-0,000009

Figure 5.11: Mapped CTD data for the results tables.

5.1.2 Project Local Data

Test cases in the previous section didn't cover any information besides the latitude and longitude, to contextualize the observations. We were missing information like the personnel involved, the campaign and project where the observations belonged and even the stations/locations where the data was collected. In order to fill in this information, we collected data from the campaign reports. For example, in Figure 5.12, we can see the campaigns where the data from the benthos and CTD was gathered and in Figure 5.13 the stations where it was collected.

campaign						
campaignid	code	name	start	end	scope	goals
1	02020513	BIOMETORE1	03/05/2013	25/05/2013	BIOMETORE Project (EEA Grants - PT02 Programme)	Characterization of pelagic and benthic ecosystems of the seamounts. Collect hydrographic data of study areas.
2	2030915	ARQBIOMETOR E-43-V15	05/09/2015	31/10/2015	BIOMETORE Project (EEA Grants - PT02 Programme)	Describe species diversity of demersal and deep-water fishes and several invertebrates up to 2000 m depth.

Figure 5.12: Mapped data for the *campaign* table.

station				
stationid	code	name	point	depth
1	I007	I007	40, -9.616666666666667	223
2	D1	D1	38.16138888888889, -8.908333333333333	13,4

Figure 5.13: Mapped data for the *station* table.

5.2 File Preservation

In order to ensure data preservation, it is also necessary to store several types of files, such as photos, videos, sounds, digitalized documents and spreadsheets.

The photos, videos and sounds can be stored in the 52° North model as a feature of interest or observable property, depending on the context. For example, the photo of a fish, in the moment it was taken, is associated with an observation and can be stored in the observable property of a result. However, it can also be the feature of interest used to make the expert judgement of the taxonomic rank.

The rest of the documentation associated with the projects, campaigns or observations can be stored in the *md_dataidentification* under the *fileidentifier* attribute using a URI (Uniform Resource Identifier) to locate it. These documents can be digitalized forms used in the campaign, campaign reports, spreadsheets and other legal documents.

While this method of storing files through an external identifier is functional, it is not ideal. The task of file storage will be explored in the future, where we will research better alternatives, to improve efficiency and organisation.

5.3 Summary and Conclusions

After mapping the data into the model, we were able to validate our proposal's robustness and completeness. The 52° North part of the model enabled us to access the data in several different contexts. The metadata part, despite not containing all the fields requested by the SNIMar, allowed for a better comprehension of the data and its context, like the campaign and the project. The service identification metadata, in addition to describing the service, allowed us to build a catalogue of the service operations. The taxonomic classification metadata provided a better description of the species, besides their scientific name. Regarding the BIOMETORE extended model, it has proved to contain all the necessary fields, allowing for a better contextualization of the observations, as well as the management of users, teams, contacts, organisations, data access and data required for the preparations of the campaigns.

This part of the model is the most flexible in regard to changes, having been modelled in order to be simple and extendible if necessary in the future. For example, if, in the campaign, it is necessary to store more information than the actual model allows, it will only be necessary to add one or more attributes to the *campaign* table. On the other hand, if the current model cannot handle the required changes, it is possible to extend it with additional tables. Although the model is easily edited, it is necessary to consider any attribute changes, such as renaming or adding new attributes. For example, the Object-Relational Mapping (ORM) has to be updated in order to ensure consistency regarding cursors and reception results variables.

Regarding the first part of the model, it is arduous to alter and we do not advise it, given, in the future, it could cause issues with new versions of the 52° North model. In case it does not meet some of the requirements, we suggest the creation of a separate table as an extension of the

Discussion

52° North table, adding new fields to the extended model, in order to connect it to the table where it belongs. For the metadata part, we advice caution in adding new fields, since they must be compliant with the European and Portuguese norms, such as SNIMar.

Chapter 6

Conclusions and Future Work

We have iterated over our data model multiple times, integrating heterogeneous data while focusing in interoperability by implementing existing directives and standards. In this chapter, we present an overview of our work that lead us to a stable and usable data model for the SeaBioData project.

6.1 Conclusions

The work we presented here, focused on the development of a data model to store and manage data from marine research projects. These projects produce a big amount of data that needs to be organized in order to enable its preservation and facilitate its analysis. The huge amount of data being collected by sensors and researchers about marine waters and their ecosystems is being lost because of the actual poor organisation of the data. This makes it hard to re-analyse the data, hindering or even making it impossible to share data with other institutions or even departments.

Driven by data loss and poor organisation, and lacking the possibility for data re-usability or external sharing, European institutions and experts developed a set of standards and directives. Throughout this work we have seen some of them, including INSPIRE and OGC web services that enabled the development of a model compliant with European norms. As an interoperable model it facilitates data exchange between departments, institutions and even countries.

We aimed to develop an information system to manage data at an European level, in the sense that it could easily be reused across all members states. The first step to develop an effective and functional information system was to build up a well designed model for the data. To do that, we started by surveying data acquisition methodologies and analysing existing datasets. This survey allowed us for a better understanding of the structure and content of the various data types, which includes biotic and abiotic data, laboratorial results, data of fishing activities, photos, videos, sound files and time series obtained from sensors.

Conclusions and Future Work

After analysing the state of the art, as well as the data and metadata that would be represented within components of the model and the requirements and needs of the researchers, we came to a solution. We built a relational data model that can stock all the different types of data required by IPMA. This model has three main parts. The first part is where we stored the data associated with observations, and collected within campaigns, and from sensors and laboratorial analysis. For this part of the model, we used an open source version of 52° North, that implements the OGC SOS specification. The second part corresponds to the metadata model used to describe the available services, that provides additional information about the observations and campaigns, such as a detailed description on how the data was collected, an abstract of the contents of the datasets and its purpose, and even supports the storage of documents. We also added tables to store additional information about the taxonomic rank of the species identified within an observation, including its kingdom, phylum, class, order, family, genus, scientific name and author. The third part of the model is where we stored all the extra data required by the project administration in the information system, including users, teams, contacts, organisations, departments and workpackages, important to define the access levels of the users. It also stored additional information required for the organisation of future campaigns, with lists of vessels, procedures, forms, locations and stations. Finally, the data model we proposed also supports campaign and project entities, as well as the connection between campaigns and observations.

We validated our data model by mapping real data provided by IPMA to the resulting model and concluded that the model is robust and supports all the data required by this organisation. The proposed data structure can easily be extended in the future, in order to support additional administration data, by altering the third part of the model. While the first part of the model is the most challenging to alter in the future, our validation process indicates that, given its implementation of the OGC SOS, it will easily support any unexpected changes to data storage requirements. We chose to leave this implementation unchanged, in order to enable future version updates of the 52° North platform.

Summing up, the data model has reached a high maturity level and is ready to support the information system that is being developed within SeaBioData project.

6.2 Future Work

In the future, we will improve the file storage scheme, as well as build the information system that accesses the database we prepared. As the models stands it is possible to store all the required files, but this process can be improved by exploring alternatives, such as a file system directory hierarchy or a file repository. In particular, we will focus on the development of a file system structure that allows for a better organisation and maintenance of the files. Regarding the information system, it has been developed alongside the relational data model, within the SeaBioData project, but it is still lacking in several API operations and front-end developing. Given the practical nature and real-world characteristics of this project, the development of an information system supported on a robust data model is critical to the quality of day-to-day operations.

References

- [BOO] BOOS. The BOOS vision. <http://www.boos.org/>. Accessed June 2016.
- [Cen] European Commission Joint Research Centre. INSPIRE feature concept dictionary. <http://inspire.ec.europa.eu/featureconcept>. Accessed June 2016.
- [Com] European Commission. INSPIRE DIRECTIVE. <http://inspire.ec.europa.eu/>. Accessed June 2016.
- [Com15] European Commission. FP6, 2015. https://ec.europa.eu/research/fp6/index_en.cfm. Accessed June 2016.
- [Com16a] European Commission. The eu water framework directive - integrated river basin management for europe, 2016. http://ec.europa.eu/environment/water/water-framework/index_en.html. Accessed June 2016.
- [Com16b] European Commission. FP7, 2016. http://ec.europa.eu/research/fp7/index_en.cfm. Accessed June 2016.
- [Con16a] Open Geospatial Consortium. OGC, 2016. <http://www.opengeospatial.org/>. Accessed June 2016.
- [Con16b] Open Geospatial Consortium. Sensor Observation Service, 2016. <http://www.opengeospatial.org/standards/sos>. Accessed June 2016.
- [dee16] deegree.org. Deegree, 2016. <http://www.deegree.org/>. Accessed June 2016.
- [dUdLL] Association Francophone des Utilisateurs de Logiciels Libres. Definition of Interoperability. <http://interoperability-definition.info/en/>. Accessed June 2016.
- [ECJRC14] Ispra European Commission Joint Research Center. Feature Catalogue 'INSPIRE application schemas', 2014. <http://inspire.ec.europa.eu/data-model/approved/r4618/fc/>. Accessed June 2016.
- [FFM] A. Puente J. A. Juanes F. Fernandez, V. Velarde and R. Medina. Answering Environmental European Directives through information systems. In *OCEANS 2011 IEEE - SPAIN*, pages 1–7.
- [fGOSSG14] 52° North Initiative for Geospatial Open Source Software GmbH. Sensor Observation Service, 2014. <http://52north.org/communities/sensorweb/sos/>. Accessed June 2016.
- [fGOSSG16] 52° North Initiative for Geospatial Open Source Software GmbH. 52° North, 2016. <http://52north.org/about/52north>. Accessed June 2016.

REFERENCES

- [Fou16] Open Source Geospatial Foundation. MapServer, 2016. <http://mapserver.org/>. Accessed June 2016.
- [GS16] Geo-Seas. Geo-Seas, 2016. <http://www.geo-seas.eu/>. Accessed June 2016.
- [Hol12a] Carsten Hollmann. 52° north sos data models, 2012. <https://wiki.52north.org/bin/view/SensorWeb/SosDataModeling>. Accessed June 2016.
- [Hol12b] Carsten Hollmann. The 52° north sos standard data model, 2012. https://wiki.52north.org/SensorWeb/SosDataModeling#The_52North_SOS_standard_data_model. Accessed June 2016.
- [Hol16] Carsten Hollmann. Sos 4.x documentation, 2016. <https://wiki.52north.org/bin/view/SensorWeb/SensorObservationServiceIVDocumentation>. Accessed June 2016.
- [IFR] IFREMER. SeaDataNet. <http://www.seadatanet.org/>. Accessed June 2016.
- [INS11] INSPIRE Working Group on Observations & Measurements. D2.9 Guidelines for the use of Observations & Measurements and Sensor Web Enablement - related standards in INSPIRE Annex II and III data specification development, 2011.
- [ISO14] ISO. ISO 19115-1:2014, 2014. http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798. Accessed June 2016.
- [JALR14] Maria A. Martinez Juan Pazos Juan A. Lara, David Lizcano and Teresa Riera. A UML profile for the conceptual modelling of structurally complex data: Easing human effort in the KDD process. *Information and Software Technology*, 56(3):335–351, 2014.
- [JLGZ08] Timothy L. Whiteaker David R. Maidment Jonathan L. Goodall, Jeffery S. Horsburgh and Ilya Zaslavsky. A first approach to web services for the National Water Information System. *Environmental Modelling & Software*, 23(4):404–411, 2008.
- [JZT] J. Pardillo J. Zubcoff and J. Trujillo. A UML profile for the conceptual modelling of data-mining with time-series in data warehouses. *Inf. Softw. Technol. (Netherlands)*, 51(6):977 – 92.
- [LSW06] Abbas Rajabifard Lisa Strain and Ian Williamson. Marine administration and spatial data infrastructure. *Marine Policy*, 30(4):431–441, 2006.
- [MA15] Cannata Massimiliano and Milan Antonovic. istSOS-project documentation, 2015. <http://istsos.org/en/latest/doc/>. Accessed June 2016.
- [OGC11] Inc. Open Geospatial Consortium. Sensor Observation Service (SOS) Introduction, 2011. http://www.ogcnetwork.net/SOS_Intro. Accessed June 2016.
- [OGC16] OGC. Sensor Model Language (SensorML), 2016. <http://www.opengeospatial.org/standards/sensorml>. Accessed June 2016.
- [Ope05] Open Geospatial Consortium Inc. *Web Feature Service Implementation Specification*, 2005.

REFERENCES

- [Ope06] Open Geospatial Consortium Inc. *Open Geospatial Consortium Inc . OpenGIS® Web Map Server Implementation Specification*, 2006.
- [Ope07] Open Geospatial Consortium Inc. *OpenGIS® Catalogue Services Specification*, 2007.
- [Ope10a] Open Geospatial Consortium. *OGC® GML Application Schema - Coverages*, 2010.
- [Ope10b] Open Geospatial Consortium Inc. *OpenGIS Web Feature Service 2.0 Interface Standard*. 2010.
- [Ope12] Open Geospatial Consortium. *OGC® WCS 2.0 Interface Standard- Core: Corrigendum*. 2012.
- [Ope13] Open Geospatial Consortium. *Geographic information — Observations and measurements*, 2013.
- [Ope14] Open Geospatial Consortium Inc. *OpenGIS Web Feature Service 2.0 Interface Standard — With Corrigendum*. 2014.
- [Ope15] Open Geospatial Consortium. *OGC WPS 2.0 Interface Standard*, 2015.
- [PDS12] R. Lowry Paolo Diviacco and D. Schaap. Marine seismic metadata for an integrated European scale data infrastructure: The FP7 Geo-Seas project. *Bollettino di Geofisica Teorica ed Applicata*, 53(2):243–252, 2012.
- [Ser] Copernicus Marine Environment Monitoring Service. cmems.met.no. <http://myocean.met.no/>. Accessed June 2016.
- [Sis15] Sistema Nacional de Informação do Mar. *Perfil SNIMar v.0.9 – Perfil de Metadados*, 2015.
- [Soe08] Kai C Soetje. To be on the right path from BOOS to an integrated pan-European marine data management system. In *US/EU-Baltic International Symposium: Ocean Observations, Ecosystem-Based Management and Forecasting - Provisional Symposium Proceedings, BALTIC*, page The Institute of Tallinn University of Technology, 2008.
- [Sta15] Biodiversity Information Standards. Darwin core, 2015. <http://rs.tdwg.org/dwc/>. Accessed June 2016.
- [Wik13] Wikipedia. MapServer, 2013. <https://pt.wikipedia.org/wiki/MapServer>. Accessed June 2016.
- [Wik15] Wikipedia. Crucible, 2015. [https://en.wikipedia.org/wiki/Crucible_\(disambiguation\)](https://en.wikipedia.org/wiki/Crucible_(disambiguation)). Accessed June 2016.
- [Wik16] Wikipedia. Open Geospatial Consortium, 2016. https://en.wikipedia.org/wiki/Open_Geospatial_Consortium. Accessed June 2016.

REFERENCES

Appendix A

Full Model

Full Model

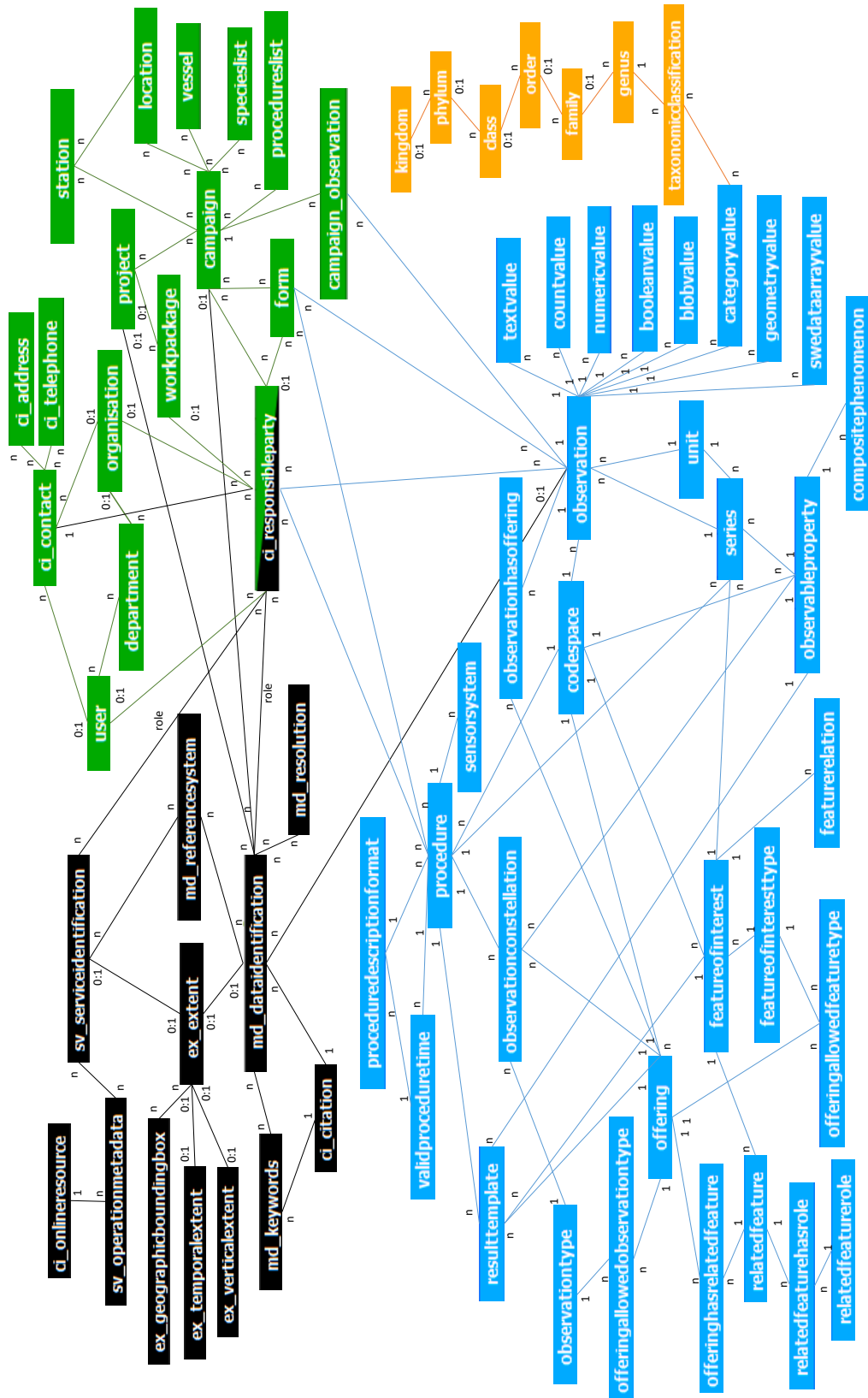


Figure A.1: Full model