Integrating grid and trajectory data via a web service: case study of iceberg movement

Nguyen Duc Ha
February, 2010
Integrating grid and trajectory data via a web service: case study of iceberg movement

by

Nguyen Duc Ha

Thesis submitted to the International of Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: Geoinformatics

Thesis Assessment Board

Thesis advisor
First supervisor: Dr. U.D. Turdukulov
Second supervisor: V. Venus MSc
Advisor: Ms Dr. C.A. Blok

Thesis examiners
Chairman: Prof.Dr. M.J. Kraak
External examiner: Dr. G. Andrienko
Disclaimer

This document describes work undertaken as part of a programme of study at the International Institute for Geo-information Science and Earth Observation. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the faculty.
Abstract

The understanding about behaviours of moving objects is required in many areas of science and technology. However, many unanswered questions related to moving objects still remain. The main reason is that moving objects do not exist alone. In the dynamic environment, they are influenced by many factors. These factors hide significant knowledge for understanding and predicting object movement behaviours. Therefore, to study behaviours of moving objects, the relations between moving objects and influencing factors need to be identified. The development of tracking devices, satellite technology, grid exchange mechanism and web services bring unprecedented opportunity for researchers and scientists to discover these relations. Nevertheless, the difference in data formats (trajectories are recorded in two/three dimensional vector while environmental factors are often in multidimensional grid) and spatio-temporal resolution become obstacles for data integration. In addition, the large quantity of trajectory and grid data is difficult for conventional integration process.

To ease the integration process and facilitate users from different research fields and distinct data integration needs, in this research, we aim to build up a web service that provides the capability to integrate trajectory data with grid data from remote servers. Concerning the advantageous of data structure that includes time as a dimension, data availability and service supporting, we decide to concentrate on multidimensional grid data of the Earth Science community. The result of this research is a web service with an HTML graphical user interface that allows users just upload their own trajectory data file, indicate grid data source links and receive the integrated data set. The integration mechanism is built up by using IDV and Jython script. It has capability to access grid data from OPeNDAP servers through OPeNDAP protocol then integrate grid and trajectory data. Some functions for interpolation and evaluating uncertainty of interpolated value are also supported. To support user through web environment, the integration mechanism is embedded on an Apache server containing Mod_python module. The server uses Python script as a middleware to get and respond requests from client side. The implemented prototype was fulfilled research objective. Iceberg trajectory data and grid data of sea surface temperature, wind direction are used as a case study. The integrated data set could be used for visualization applications (such as Google Earth) or trajectory models. This may help users to understand the relation between moving objects and influencing factors.

Key words: Data integration, Trajectory, Grid data, Iceberg, Unidata, OPeNDAP, IDV
Acknowledgements

I would like to take this opportunity to express my appreciation to all people who assisted me to accomplish this research.

First at all, I am grateful and appreciate Mr. Huurnerman, the course Director of Geoinformatics for giving me a chance to study Msc course at ITC.

I would like to forward my deep gratitude to Dr Ulanbek Turdulkulok and Dr Connie Blok for their support in advice, material wise and their valuable comments to make sure that the final output should be in a good quality. They spend most of their time to read my report and made comments that helped me to have a readable thesis in a consistent manner. I would like to thank them so much and I appreciated for what they did. In other hand, I would like to thank Mr Valentijn Venus for his advice and giving some IDV script that were very helpful for my research.

I would like to thank Prof. Dr. Menno-Jan Kraak for his advice and encouragement as my chairman. He helped me in different aspects and I appreciate for the support and kindness he had shown to me during my research.

I would like to thank Mr Don Murray and Mr Jeff McWhirter – IDV developers for their advice and guidance about IDV. I also would like to thank Ms Nguyen Thi Thu Ha and Mr Mobushir Riaz Khan for their advice while doing the research.

I deeply appreciate my friends: Mr Nguyen Hoang Long and Mrs Nguyen Thanh Nga, who always support, encourage me during my study.

Furthermore, I appreciate all my classmates of Geoinformatics for the support and good time we had during our stay and studies. Also, I thank to other Vietnamese friends who are/were here to make the foreign land like home.

Last but not least, my special thanks to all members of my family for their constant and un-conditional support throughout my whole studies. Words cannot express my emotion that I feel for them from the bottom of my heart.
...Dedicated to my parents...
## Table of contents

### Contents

1. **Introduction** ......................................................................................................................... 1  
   1.1. Background .......................................................................................................................... 1  
   1.2. Problem statement and justification ...................................................................................... 2  
   1.3. Related work ....................................................................................................................... 2  
   1.4. Research identification ........................................................................................................ 3  
      1.4.1. Research objectives ........................................................................................................ 3  
      1.4.2. Research questions ......................................................................................................... 4  
      1.4.3. Innovation aimed at ....................................................................................................... 4  
   1.5. Project set-up ....................................................................................................................... 5  
      1.5.1. Method .......................................................................................................................... 5  
   1.6. Organization of the thesis .................................................................................................... 6  
2. **The need for integration of grid and trajectory data: use and user aspects** ................................... 8  
   2.1. Introduction ....................................................................................................................... 8  
   2.2. Data integration issues and requirement to integrate trajectory and grid data ................... 8  
      2.2.1. Data integration issues .................................................................................................. 8  
      2.2.2. Requirement to integrate trajectory and grid data ......................................................... 9  
   2.3. Users of weather and iceberg data ..................................................................................... 10  
   2.4. User requirements of weather and iceberg data .................................................................. 12  
   2.5. Data characteristics .......................................................................................................... 13  
   2.6. Uncertainty ....................................................................................................................... 14  
      2.6.1. Uncertainty concepts .................................................................................................... 14  
      2.6.2. Sources of uncertainty ................................................................................................. 15  
      2.6.3. Uncertainty sources of the case study .......................................................................... 15  
   2.7. Conclusion ......................................................................................................................... 16  
3. **Data integration techniques** .................................................................................................. 17  
   3.1. Introduction ....................................................................................................................... 17  
   3.2. The choice of a data system ............................................................................................... 17  
   3.3. OPeNDAP (Open source Project for a Network Data Access Protocol) ......................... 20  
   3.4. OPeNDAP clients .............................................................................................................. 22  
      3.4.1. OPeNDAP clients ........................................................................................................ 22  
      3.4.2. Integrated data viewer (IDV) ..................................................................................... 23  
   3.5. Conclusion ......................................................................................................................... 24  
4. **Designing and implementing the prototype for data integration system** .................................. 25  
   4.1. Introduction ....................................................................................................................... 25  
   4.2. Desktop-based integration mechanism ............................................................................... 25  
      4.2.1. Data integration ........................................................................................................... 25  
      4.2.2. Data interpolation and uncertainty ............................................................................... 28  
   4.3. Building up data integration web service ....................................................................... 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1.</td>
<td>Graphical User Interface</td>
<td>34</td>
</tr>
<tr>
<td>4.3.2.</td>
<td>Server side Python script</td>
<td>35</td>
</tr>
<tr>
<td>4.4.</td>
<td>Results</td>
<td>36</td>
</tr>
<tr>
<td>4.5.</td>
<td>Evaluation</td>
<td>40</td>
</tr>
<tr>
<td>4.6.</td>
<td>Conclusion</td>
<td>41</td>
</tr>
<tr>
<td>5.</td>
<td>Conclusions and recommendations</td>
<td>42</td>
</tr>
<tr>
<td>5.1.</td>
<td>Introduction</td>
<td>42</td>
</tr>
<tr>
<td>5.2.</td>
<td>Conclusions</td>
<td>42</td>
</tr>
<tr>
<td>5.3.</td>
<td>Recommendations for further research</td>
<td>44</td>
</tr>
<tr>
<td>6.</td>
<td>References</td>
<td>46</td>
</tr>
</tbody>
</table>
List of figures

Figure 1-1: The flow chart of methodology ................................................................. 5
Figure 2-1: The basic components of the triad framework (Peuquet, 1994) .................. 9
Figure 2-2: Quadrant A-D (Source: URL3) ................................................................. 11
Figure 3-1: Multidimensional grid data (time is a dimension) ................................. 18
Figure 3-2: Geospatial data/two dimensional grid data ........................................... 18
Figure 3-3: Unidata and OGC architectures ................................................................. 19
Figure 4-1: Inverse distance weighting. In green, the assumed iceberg size. In dashed line, grid cells that iceberg overlay. In blue, the centre of measured grid cells .................................. 29
Figure 4-2: The conceptual model of data integration system via web service .......... 32
Figure 4-3: The system model in steps ................................................................. 1
Figure 4-4: Antarctic iceberg data provided by NIC ............................................. 36
Figure 4-5: Iceberg trajectory dataset ...................................................................... 37
Figure 4-6: HTML form for putting data source link, spatial resolution, interpolation method and time accuracy ................................................................. 38
Figure 4-7: HTML form for selecting data choices .............................................. 39
Figure 4-8: Integrated dataset (opened in Microsoft Excel) ...................................... 39
Figure 4-9: Visualizing integrated dataset on Google Earth .................................... 40
List of abbreviations

CSV: Comma Separated Value
ES: Earth Science
GIScience: Geo-Information Science
GIS: Geographic Information System
IDV: Integrated Data Viewer
NIC: National Ice Center
NDVI: Normalised Difference Vegetation Index
OPeNDAP: Open-source Project for a Network Data Access Protocol
OGC: Open Geospatial Consortium
SST: Sea Surface Temperature
THREDDS: Thematic Real-time Environmental Distributed Data Services
1. Introduction

1.1. Background

Movement is a main component of many activities. In many areas of science and technology such as meteorology, biology, sociology, transportation engineering, it is important to be aware of movement behaviours (Dodge et al., 2008). Many researches have been studying behaviours of moving objects such as animal movements, transportation movements. With regard to the rapid development of inexpensive positioning technologies, trajectory data and other data related to dynamics of moving objects are obtained in quickly increasing quantities (Andrienko et al., 2007). However, moving objects are influenced by various factors that impact and constraint their movements (Dodge et al., 2008). For example, the movement of bird is affected by temperature and vegetation distribution (Underwood and Chapman, 1985) or the movement of airplane is affected by weather conditions (Ahlstrom, 2005). These influencing factors hide significant knowledge for understanding and predicting the movements. In order to get an understanding about movement behaviour of dynamic objects, it is required to take a closer look at the movement itself. In other words, it is necessary to know not only geographical space and time (Butenuth et al., 2007) but also what and how external factors affect and limit the movement (URL.1). When the causes and consequences of object movement behaviours are determined, prediction models can be established. To evaluate and verify the quality of these models, the predicted data of movement behaviours and influencing factors need to be compared with actual data.

In recent years, a lot of attention has been paid to environmental monitoring with the use of satellite data (Shelestov, 2008). Using satellite is an innovative technology that has made it possible to monitor rapidly changing phenomena on the surface of the earth (Sohrabinia and Khorshiddoust, 2007). To study the external impacts that affect moving objects, remote sensing imageries from satellite sensors play an important role. These additional data are available and can be accessed from many satellite data servers as MODIS (Moderate Resolution Imaging Spectroradiometer) (URL.19), NOAA (National Oceanic and Atmospheric Administration) (URL.18) servers etc. Together with the development of satellites and grid data from these satellites, grid web services have been developed. Grid web services as THREDDS (Thematic Realtime Environmental Distributed Data Services) and Web Coverage Services (OGC WCS) have been become powerful tools in collecting and disseminating grid data and especially geo-data by using OPeNDAP (Open-source Project for a Network Data Access Protocol), OGC WCS, HTTP (Hypertext Transfer Protocol) or other remote data access protocols.

Since icebergs play a significant role in navigation, communication networks and climate change studies, monitoring and tracking icebergs have been considered and performed by many researchers. Large Antarctic icebergs are identified and tracked by National Ice Centre using a variety of satellite sensors from 1976 till present. The resulting data are reported approximately every 1-5 days (Ballantyne and G.Long, 2002) and provided through the World Wide Web. “Icebergs might be a huge threat if they go in the shipping routes, causing risks of collision with ship navigation. They may
also have impact on fixed and floating oil installations, and risks of scour on sub-sea installations, such as pipelines and wellheads. The consequences of iceberg impact can be severe in terms of the structural integrity of installations, and the safety of personnel and environment” (Veitch and Daley, 2000). Climatologists also use iceberg data to study climate change because of the potential sensitivity of icebergs. The process of calving of large iceberg can help to indicate the rapid change of climate, and have a potential influence on the global sea level increase (Kenneally and Hughes, 2006). Movement or calving of icebergs are influenced by temperatures, wind, tides, radiation, seasonal variation, and many other parameters (Benn et al., 2007; Jansen et al., 2007). Thus, the movement and disintegration of icebergs cannot be seen as independent events. With lack of required functionality, it is difficult to integrate, analyze and understand the current movement patterns and hard to verify evolution models of iceberg.

1.2. Problem statement and justification

In order to investigate the relationship between the movement behaviours of objects and the influencing factors, or to evaluate the quality of prediction models, a dataset including movement behaviour and influencing factor data is needed. Data of moving objects are used by different users in many regions on the world; therefore, they should be interoperable. These data should be formed in a common format; consequently they can be exchanged, read and written by different programs in a common set of exchange format. Moving objects are frequently tracked and reported in two/three dimensional vector format. On the other hand, external data as weather, vegetation data are often in multidimensional grid format with different spatiotemporal resolutions. Grid format can be considered as a set of two, three or multidimensional array of parameter values.

For performing study in movement behaviours or the relationship between moving objects and surrounding environment, decision makers frequently have to find out methods to exchange the grid data into a well-suited Geographic Information System (GIS) format before these data can be used in GIS tools. Therefore, the problems addressed here are: how to extract relevant data of external factors in grid format via web services; and how to integrate them with vector data (trajectories) with different spatiotemporal resolution so that they become an interoperable dataset? Furthermore, since geographic information is used for decision making process, uncertainty or error when performing the data integration can drive users to undesirable consequences (Devillers and Jeansoulin, 2006). Hence, uncertainty or error quantification when performing the data integration is also required.

1.3. Related work

Due to the importance of understanding the relationship between moving objects and influencing factors, and also the need of evaluating the quality of prediction models, there are many researchers tried to cope with these issues.

The trend of integrating data and taking advantage of web services for data system interoperability are considered in many organizations such as OGC (Open Geospatial Consortium) and Unidata organizations (URL.2). However, there are abundant researches that have used the combination of trajectory data and influencing data from satellite data, a few works have been done for creating a mechanism for integrating trajectory and grid data.
Froude (2008) introduced TRACK web application that allows storm identification and tracking software to be performed from a web browser. The user can create a list of jobs (such as different data periods) in their web browser, and then these requirements are sent to the server. The remote datasets of storm are accessed using the OpenDAP protocol from other data servers. Once the job have finished running, the output is put on the server for users to download or plot.

Ipopo (2009) tried to manage the problem of integrating datasets with different temporal resolutions. She used iceberg data and sea surface temperature (SST) data as an example so that users could explore the relationship between iceberg calving and SST. To fulfil that purpose, SST images were downloaded from the web manually, georeferenced and converted from raster format into vector data. She managed to integrate and visualize iceberg trajectories and corresponding SST, however, the process was manual and is hardly suitable for large number of trajectories.

VisualEsse is a project in Codeplex (an Open source project community) (URL 5). It developed a thick client-server system that allows users deliver GPS (Global Positioning System) tracks over Earth surface (using Microsoft Virtual Earth) and discover values of meteorological parameters (such as temperature, wind) along tracks. These meteorological parameters are stored on database server. The integrated values can be saved to Microsoft Excel worksheet.

The issue of integrating grid and trajectory data were raised; however integrated dataset should be easy for user to access. Therefore, this study tends to integrate grid and trajectory data which will be automatically downloaded in a common format; and integrated dataset will be disseminated via web environment.

1.4. Research identification

1.4.1. Research objectives

The main research objective of this research is to help users to explore the relationship between the iceberg trajectories and metocean (meteorological and oceanographic) variables (such as sea surface temperature, wind direction, wind speed, air temperature) by making data integration mechanism available as a web service.

Based on the main research objective, there are specific objectives addressed for this research:
1. To identify and understand the user requirements for studying integrated iceberg and metocean data.
2. To identify the data extracting method that can obtain metocean data from data services.
3. To understand the nature of data sets to be integrated.
4. To determine and implement the integration technique for iceberg and metocean data.
5. To determine the uncertainty and error quantification when performing the data integration.
6. To evaluate the proposed prototype.
1.4.2. Research questions

Based on the objectives the following research questions have been formulated:

1. To identify and understand the user requirements for studying integrated iceberg and metocean data.
   • Who are the users and what are the requirements for integrating different spatial temporal datasets?
   • What kind of metocean data need to be derived?

2. To identify the data extracting method that can obtain climate data from data services.
   • Where can metocean data be extracted from?
   • What is the remote data access protocol?
   • What method should be applied to extract climate data?

3. To understand the nature of data sets to be integrated.
   • How are the data formatted?
   • What are the spatial and temporal resolutions of the datasets?

4. To determine and implement the integration technique for iceberg and metocean data.
   • What kind of spatial statistical climate data measure is needed (max, min or average)?
   • What protocol should be used for forming dataset?
   • Which integration methods should be applied for integrating metocean data in relation to iceberg data?
   • Which interpolation methods should be applied for interpolating metocean data in relation to iceberg data?

5. To determine the uncertainty and error quantification when performing the data integration.
   • Which method should be applied for determining the uncertainty and error quantification of integrated data?

6. To evaluate the proposed prototype
   • Which method should be used to evaluate the prototype?

1.4.3. Innovation aimed at

This research aims to provide a new infrastructure that allows integrating trajectory data and grid data in different spatial-temporal resolutions and formats and makes integration mechanism available in a web environment via web service.
1.5. Project set-up

1.5.1. Method

The research methodology was based on the addressed research objectives. The methodology will be used to answer the research questions as explained in the flow chart (Figure 1-1).

Figure 1-1: The flow chart of methodology
1. Identify users and their requirement: literature study on the factors influencing the iceberg movement and calving, review of users of icebergs data and their requirements

2. Identify data sources and method for extracting data: as mentioned before, due to the development of satellites, metocean data from satellite data server will be used. Depending on the data availability of influencing factors in research area and period, relevant grid data sources will be determined. From then, data access protocol and methods to access these data remotely need to be identified.

3. Identify data characteristic: Data characteristics such as grid format, projections, spatial and temporal resolutions need to be understood.

4. Integrating grid and trajectory data:
   - Identify what kind of spatial statistical metocean data value is needed: In case of large moving objects, their size may overlap many grid cells. Hence, spatial statistical metocean measure of interest will be determined (max, min or average value or all of them).
   - Identify integration approach: find out the method for integrating grid and trajectory data
   - Identify interpolation approach: in case of value of metocean at tended location for integration is missing, find out the method for interpolating metocean data.
   - Implementation of integrating grid and trajectory data.

5. Quantification of uncertainty and error: The data quality components and data quality requirements will be determined. From these, the quality of the dataset will be evaluated.

6. Performance evaluation: Finally, the prototype will be evaluated in order to determine its capabilities and limitations.

1.6. Organization of the thesis

The structure of the rest of my thesis is organized as follows:

Chapter 2 provides the description of the need for integration of grid and trajectory data. For case study of iceberg, users of iceberg and metocean data, and their requirement of integrated data set in order to explore behaviours of icebergs are also introduced. The chapter describe some basic characteristics of grid and trajectory data. Then, the overview of data integration and data uncertainty issues is represented.

Chapter 3 introduces some technique choices what are envisaged to use for building up a data integration system. In the chapter, data models, data access protocol and clients software are introduced and compared and the most suitable ones are chosen.

Chapter 4 explains the design and implementation processes for the prototype of data integration system via a web service. The system consists of three main parts: server side, client side and middle ware. Iceberg trajectory and metocean data sources are used as input for the web service. The
integrated data set is visualized in Google Earth to illustrate result of integration process. The evaluation for the prototype is also given in chapter.

Chapter 5 presents the conclusions, recommendations and main finding of the research.
2. The need for integration of grid and trajectory data: use and user aspects

2.1. Introduction

Understanding, explanation, and prediction of real world phenomena are always needed for various purposes. Moving objects and their surrounding environmental are usually monitored by different organizations using different techniques. These differences cause difficulties in integrating both datasets. This study focuses on the method of data integration in order to facilitate users who attempt to understand movement behaviours, the influencing factors of the movement, and especially the relationship between them.

The purpose of this chapter is highlighting the use and users of trajectory and grid data, look at frequently asked questions concerning moving objects and the methods to explore movements in relation to their environment. The main focus is on case study: iceberg trajectories and metocean data. The chapter gives an overview of uncertainty issues concerning data integration. The methods, techniques and tools for actual data integration will be further discussed in the following chapters.

2.2. Data integration issues and requirement to integrate trajectory and grid data

2.2.1. Data integration issues

“Data integration is the process that involves merging data from multiple databases of multiple sources and offering the user with a unified view of these data” (Lenzerini, 2002). Data integration can be described as an overall process that resolves data heterogeneity and datum selection (Kumar et al., 2006).

In a GIS environment, integration of spatial temporal datasets from different sources (which may have different survey times, different degrees of geospatial resolution or different projections) is a very important issue for data analysis and exploration. The integration of heterogeneous geospatial data offers possibilities to manually and automatically derive new information, which is not available when using only a single data source (Butenuth et al., 2007). Furthermore, it allows for a consistent representation and the updating propagation from one data set to the other. Integration assists in pattern identification and facilitates end uses in decision making.

Therefore, the issue of designing data integration systems is important in current real world applications. To manage the huge amount of data from different sources, data integration systems need to offer a uniform interface and ability to exchange data across multiple systems (Dong et al., 2007).
2.2.2. Requirement to integrate trajectory and grid data

Our world is not static and movement is a vital part of almost all processes and activities on the world. In a general sense, movement is “the change in the physical position of an entity with respect to some reference system (geographical space) within which one can access positions” (Rinzivillo et al., 2008). In the geographically dynamic context, movement of an object which maintains the same identity is a change in location over time (Andrienko et al., 2008).

The path made by a moving object can be called as a trajectory. In other words, “a trajectory is a sequence of locations, each associated with a timestamp, describing the movement of a point” (Gudmundsson et al., 2009).

The need and interest in studying and understanding movement behaviours are not new. Many researches of movement have been performed in the fields of human, transportation and animal analysis.

Movement implies change and to detect change in spatial temporal data, Peuquet (1994) distinguishes three components of geodata and three basic questions in a conceptual framework of temporal dynamics in geographic information systems. These components relate to location (where), objects (what) and time (when).

![Figure 2-1: The basic components of the triad framework (Peuquet, 1994)](image)

Geographic phenomena can be approached from three different angles enabling three basic kinds of questions (Peuquet, 1994):

- **When + Where → What**: (What) illustrates the object or a set of objects in which it is present at a certain location or set location (Where) at a certain time or set of times (When).
- **When + What → Where**: (Where) depicts the location or set of locations used by a certain objects or set of objects (What) at a certain time or set of times (When).
- **Where + What → When**: (When) illustrates the times or set of times that certain objects or set of objects (What) used a certain location or set of locations (Where).

The stated questions aim at understanding the behaviours and characteristics of moving objects. It is clear that with spatial-temporal data and especially trajectory data, location and time cannot be seen
as attributes only, but as dimensions in themselves as well. Combinations of these components allow users to pose more complex questions. Those above questions could help scientists in discovering changes of moving objects. For example, in case of airplane traffic management, the answers to questions about the position of a particular flight at present or a future time can be provided. And, in case of bird movement, the position of a bird species in a particular period can be given.

However, many unanswered questions still remain or the available answers are not satisfactory. The reason is that moving objects do not exist alone. More or less, other factors influence the behaviours of these moving objects. To study the behaviours of moving objects, in their dynamic environment, the effects of other factors have to be taken in account. In case of bird movement, birds are cause of many airplane accidents. Many researches have been done to reduce bird strikes. For this case, knowing the position of bird species in time is not enough. Scientists have to learn what kind of vegetation surround airports where birds attract. Those vegetation types will then be replaced or limited. Thus, in most of cases, “What” in the triad should include not only moving entity attributes but also its influencing factors.

In the last few decades, the limitation of data, methods and techniques prevented insight into the relationships. Data sets of trajectories and their environment over a long period and large area were not available. Recently, this situation has changed. The development of sensors and automatic acquisition tools leads to the rapid increase of available digital data sets, data models and a diversity of data types. This is especially true for geospatial data which are obtained by many institutions, as mapping or environmental agencies and companies (Butenuth et al., 2007).

Due to the need of understanding behaviours of moving objects and their relationship of them with influencing factors, integrated data sets including these factors are needed. With the diversity of available data sources nowadays, data integration becomes a good and applicable solution to bring a clear link to answer user’s questions.

### 2.3. Users of weather and iceberg data

The changes in global climate cause glaciers and ice shells to shrink and calve apart to thousands of irregularly formed icebergs. Iceberg movement is unrestricted movement in the polar region (especially Antarctic region). However, the movement of icebergs cannot be seen as independent event, it depends on winds, tides, ocean current, seasonal variations, and perhaps more. This movement is related to environmental changes and can be hazards for human activities.

In studying iceberg behaviour and data about influencing factors, users try to understand the relationship between these data. They try to discover the cause of iceberg movement and events such as the appearance and disappearance, the fracturing and calving, and the change in amounts of iceberg. For the purpose of this study, users of icebergs and other data can be classified into three main groups as follow:

- Institutions and groups involves in data collections and monitoring
- Navigators, navigation planners and engineers
- Scientists like oceanographers and climatologists
For monitoring the change of iceberg behaviours, there are various institutions with different groups involved in iceberg data collection. NIC (National Ice Center) and BYU (Brigham Young University) are two mainly involved institutions. NIC has responsibility to identify, coordinate, name and track Antarctic icebergs which satisfy two basic requirements. The first requirement is that the iceberg must measure at least 10 nm (nautical mile) along the long axis. The second requirement is the iceberg be south of 60S latitude (URL.3). In addition, NIC will not track iceberg what disappear longer than 30 days.

The Antarctic quadrants are divided counter-clockwise in following mode:

![Antarctic Quadrants](URL3)

**Figure 2-2: Quadrant A-D (Source: URL3)**

A = 0-90W (Bellinghausen/Weddel Sea)
B = 90W-180 (Amundsen/Eastern Ross Sea)
C = 180-90E (Western Ross Sea/Wilkesland)
D = 90E-0 (Amery/Eastern Weddel Sea)

When an iceberg is first sighted and matches the above requirements, its point of origin is documented by NIC. The derivation of iceberg names is from the Antarctic quadrant where icebergs were originally sighted. For instance, A-20 is sequentially the 20th iceberg which is found by NIC in Antarctica between 0-90W (Quadrant A).

Remote sensing techniques are mostly used in data collection. The obtained and processed data are disseminated via the web and used by many researchers and scientists for different purposes. Navigators, navigation planners and engineers use iceberg data to study the location and the change in locations of icebergs. The reason is that iceberg movement may threaten ship navigation, industry operation or other human activities in the ocean. For instance, between 1882 and 1890, 14 vessels were lost and 40 seriously damaged due to ice. A large number of whaling and fishing vessels were lost or damaged by ice at the same time (URL.4). The loss of 1500 passengers of the “unsinkable” Titanic in 1912 made the world aware of potential disasters from icebergs. These incidents and awareness led to the establishment of systems and organizations such as the International Ice Patrol.
INTEGRATING GRID AND TRAJECTORY DATA VIA A WEB SERVICE: CASE STUDY OF ICEBERG MOVEMENT

(IIP) to track icebergs against ship collisions. Consequently, iceberg information is helpful to promote safe navigation for maritime community and help engineers to plan construction activities.

Apart from navigators and engineers, Antarctic icebergs are also the focus of scientific attention because they hold the answer to questions about the Earth’s past, present and future climate. Savage et al. (Savage et al., 2000) recognize that the calving process of icebergs is caused by two main reasons: “wave erosion of material” and “stresses induced by buoyant forces and or grounding”. Others have indicated more potential causes and relations, for example: climate warming is also noted as a reason for the breaking up of icebergs. It has influence on icebergs because it makes temperature vary with depth in the oceans in which it influences the strength of ice shelves and leads to the formation of icebergs (Jones, 2007). Oceanographers investigate the effect of icebergs since the formed large icebergs have a great impact on altering the dynamics of marine ecosystems (Arrigo et al., 2002). Giant icebergs are very effective at blocking light and decreasing the growth of phytoplankton (Arrigo and van Dijken, 2003).

2.4. User requirements of weather and iceberg data

As mentioned before, users have many purposes to study metocean and iceberg data. Navigators want to know where an iceberg is drifting to and some researchers wonder why it is there and not at another position. Others want to know how much fresh water from iceberg is dumped into the ocean, or how the iceberg and its melting affect ocean environment (URL.5). Users discover that changes on icebergs can be used as an indicator of climate change.

Some of questions that navigators, navigation planners and engineers usually have to deal with are:

- To which places and in which directions do icebergs move?
- How long does an iceberg stay in a particular location?
- What are the expected speeds of movement in different areas?
- When, where and how does iceberg calving occur?
- Where does an iceberg drift after calving?

Oceanographers and climatologists want to get answers to questions like:

- What is the sea surface temperature before calving?
- When and where does the calving process occur frequently?
- What is the effect of weather factors on iceberg movement?
- Which circumstance may lead to iceberg calving?

To address these different questions, users need to be supported by enough data sets of iceberg and influencing factors. In an attempt to recreate or predict iceberg trajectories and calving, it is paramount to understand accurately the data needed to use in the model. According to Death, Siegert et al (2006), wind stress, the ocean currents and the sea-surface temperature as three most dominant forcing fields for computing an iceberg’s dynamic and thermodynamics. These weather data are used as input parameter in several models to simulate and predict the movement, calving and melting of iceberg such as El-Tahan et al (1983), Banke and Smith (1984), Bigg et al (1997) etc. For instance, Kubat et al. (2005) introduced an operational model which is developed at CIS (Canadian Ice Service)
to forecast iceberg drift and calving over the Grand Banks region. The prediction of calved piece size
distribution and deterioration are also given. For the model inputs, water and air drag, water current,
wind wave, wind drag force and waterline length of the iceberg are used. Reliable field observations
from two tracks of two icebergs were used to examine the model prediction. The comparisons of
model prediction and observed data show that used input parameters (especially water current) have a
strong influence on prediction of iceberg behaviours. Also, the availability and reliability of input
parameters was mentioned by authors as factors that strongly affect the accuracy of operational
iceberg models.

Most of the geospatial phenomena are dynamic by nature and users always want to understand the
relationship between factors (Benn et al., 2007). In order to explore those relationships (the change of
icebergs in combination with influencing parameters), integrated techniques are required.

2.5. Data characteristics

The process and result of data integration depend on the nature of moving object, its surrounding
environment and observed data. Thus before integration, type of data to be integrated and its format
must be identified.

As mentioned before, this research intends to use iceberg trajectories as a case study of integrating
grid and trajectory data. The iceberg dataset includes more than 30 years of iceberg positions as well
as other characteristics, offered freely by NIC through the web in Excel format. The dataset contains
the following information:

- An arbitrary number identifier
- An iceberg ID
- Time: registration and observation data
- Position: sites of icebergs in each time-stamp are provide in geographical coordinates as
  latitude and longitude
- Size: the size of iceberg is given in both approximate short and long axes and in area
- Satellite: The satellite that records a particular iceberg

It is necessary to consider the temporal aspects of the iceberg dataset. The lifespan of icebergs
commonly last from a few years to 15 years. The temporal resolution is variable from 1 to 5 days.

Considering metocean data, data of metocean variables as sea-surface temperature, wind, currents can
be accessed in grid format from multiple data sources. Grid format is generally thought as a set of
numbers making up a two, three or multidimensional array of parameter values. The basic unit of the
grid is the “cell” and each cell contains one value of measured parameter. There are some different
types of grid as ERSI Coverage Format, Arc/INFO Grid, etc. however most of them consist of an
explanatory header to give information about the grid (e.g. number of rows, number of columns,
maximum and minimum values, etc.)
2.6. Uncertainty

Data are basic for decision making. Recently, due to the development and extension of data generations, there is an increasing amount of problem relating to data. This means that many decisions have to be made among increasingly complex and wide information spaces (Griethe and Schumann, 2005). GIS tools give the user full freedom to combine, overlay and analyze data from many different sources with limited attention to scale, resolution and quality of the original data and without any consideration of the accuracy characteristics of the data themselves (Gahegan and Ehlers, 2000).

When presenting some part of real world we usually have to deal with data and information. However, the used data are unfortunately not always completely perfect. Often errors, lacking values in a dataset, deviations or other kinds of uncertainty lower the quality of an analyzed data volume. Information analysts must assess not only the information presented to them, but also the confidence they have in that information. Therefore, to give a correct impression of the characteristics of the data, uncertainty problem has to be carefully considered in order not to lead to sub-optimal or even mistaken decisions (Griethe and Schumann, 2006).

2.6.1. Uncertainty concepts

The term uncertainty is used in distinct ways and fields involving statistics, finance, psychology, engineering, information science and so on. Many definitions of uncertainty have been raised but there is no agreement concerning a definition of uncertainty or a universal way on how to represent it. One of closed definitions is stated by Hunter and GoodChild (1993) as uncertainty is “degree to which the lack of knowledge about the amount of error is responsible for hesitance in accepting results and observations without caution”. Pang (2001) considers uncertainty as a multi-faced characterization about data which results, amongst others, from measurements and observations of some phenomena and from predictions made from them. Following Gahegan and Ehlers (2000), “good science requires statement of accuracy by which the reliability of results can be understood and communicated”. Accuracy can be stated as error where it is known objectively; where it is not, the term uncertainty is applied. Thus, uncertainty covers a broad range of doubt, inconsistent and error.

Tending to cover almost, if not all, possible types and sources of uncertainty in data which is not restricted to definitions used within a particular field, the National Institute of Standards and Technology (NIST) defines uncertainty to include classified uncertainty measures in these categories (Pang et al., 1997):

- Spread or statistical variations: given either by an actual distribution of the data; or estimated mean and stand deviation which can be used to calculate a confidence interval.
- Errors and differences: a difference or absolute valued error what are determined between a known correct datum and an estimate or between estimates of the data.
- Range, minimum maximum range value: an interval that data must exist in, but cannot be quantified into either the statistical or error definition.
- Noisy or missing data
The topic of uncertainty also has been receiving increasing attention in geographical sciences, in which the definition of uncertainty also varies. In a broad sense, Foody and Atkinson (2002); Zhang and Goodchild (2002) described uncertainty as terms of error, ambiguity and vagueness. While error is described for well defined objects, terms of vagueness and ambiguity are recognized as uncertainty of poorly defined objects.

More in detail, uncertainty is defined as a composition of different concepts and occurs at different levels of abstractions (Gahengan, 1998; Gahegan and Ehlers, 2000):

- Temporal error: describes the occurrence of uncertainty within time
- Positional error: describes the degree of uncertainty in space
- Completeness: represents a more abstract concept concerned with coverage
- Consistence: describes the reliability of data
- Attribute accuracy: describes measurement accuracy process of data collecting

### 2.6.2. Sources of uncertainty

The causes of uncertainty – some yet undefined doubt in the data – are diverse. There could be errors in measuring or simulation runs, statistical variations, unavoidable, or performance reasons intended losses in the transformation or even in the presentation of the data. Generally, uncertainty can be initiated from the data investigation or data acquisition through transformation and interpolation operations to the final integrated data set.

- Uncertainty in acquisition: A set of data can be acquired from instrument measurement, numerical model or human data entry; and most of them have a statistical variation. In case of instrument measurement, the longer time required to perform the measurement, the more reliable the results produced. From this method, experimental variability can be found whether it is obtained by a machine or by a scientist. On the other hand, for human data entry, the variability is caused by two factors: (1) perception among individuals and (2) minor differences in performing a task repeatedly (Gahengan, 1998).

- Uncertainty in transformation: Transformation is a process to convert data from one to another unit data measure by involving some algorithm. For instance raw data, such this type of data cannot be used directly for other process such as rescaled and re-sampled; data transformation has to be done first in order to do such those processes in which this process can be done automatically without intervention of the operator. The transformation process is to be considered as a potential source of some uncertainty because it changes in the original form of the data.

- Uncertainty in interpolation: In GIS environment, commonly the data comes from different sources using different formats. Further this data might have different spatial and temporal resolution; thus interpolation is required to estimate the data value of the differences. However, the interpolation causes uncertainty on the estimated value.

### 2.6.3. Uncertainty sources of the case study

Uncertainties in iceberg and metocean data integration are introduced in different stages, from data acquisition, data format transformation to the integration process. It may include all of uncertainties
about completeness, consistence, attribute accuracy, temporal and spatial error as mentioned above. Metocean and iceberg data are both obtained using remote sensing techniques. In relation to satellite observations, these data may contain the uncertainties about spatial and spectral uncertainties. The weather situation may affect the quality of the observed data and lead to the problem of missing data. In addition, position and size of iceberg are extracted from satellite images, grid data are often from satellite data, the measurement in these processes may also lead to uncertainties of data.

To integrate trajectory data with two or more other different data sets, they should have the same temporal resolution or the data sets that users want to integrate with their trajectory data have to have higher temporal resolution. Otherwise, an interpolation process should be applied to fill the gaps in the data sets. Furthermore, a grid data set usually has missing data, especially at Antarctic area. Hence, once again an interpolation process should be applied. All of these problems also cause the uncertainties in the integrated data.

Uncertainty cannot be certain. However, based on the available of metadata and reference data set users should be informed about the degree of uncertainty. The degree of uncertainty in process of interpolation is also can be estimated base on the applied methods. For the issue of uncertainty caused from interpolation process, it will be determined more detail in section 4.3

2.7. Conclusion

This chapter highlighted the use of trajectory and grid data. To determine trends, patterns and relationships between moving objects and factors influencing their movement, integration is necessary. For the case study of weather and iceberg data, three groups of users have been distinguished and their requirements can be based on the posed questions. Furthermore, the data uncertainty and the cause of uncertainty are discussed. The next chapter will describe the techniques for integrating trajectory and grid data.
3. Data integration techniques

3.1. Introduction

In the last chapter, we discussed the use and user requirement for integrating grid and trajectory data. A brief explanation of integration concept was also given. For the purpose of integrating grid and trajectory data in a web environment, this chapter is going to discuss the possible techniques and choices. Data models of Open Geospatial Consortium, Inc (OGC) and Unidata/UCAR from Geo-Information Science community and Earth Science community respectively are considered as potential standards to be used in this study. For choosing the more suitable one, we need to distinguish the general goals, data formats, availability of data and software of these two organizations. Based on the above choice, a data access protocol, a software package for data integration and a conceptual model of intended system are to be determined.

3.2. The choice of a data system

Continuous data such as metocean data (temperature, wind, humidity etc.), Normalised Difference Vegetation Index (NDVI), and so on are commonly stored in grid formats by both two main communities, namely Earth Science (ES) and Geo-Information Science (GIScience) communities. The structures, exchange mechanisms, and analysis methodologies of grid data have been developed over years by several organizations such as JPL (Jet Propulsion Laboratory), ERSI (Environmental Systems Research Institute), OGC (Open Geospatial Consortium), Unidata/UCAR. However, the developing technologies of these organizations have been devoted to their own realm.

Earth Science data are commonly stored in grid file formats including netCDF, HDF, GRIB and BUFR. Although there are differences among these formats, multidimensional grid is the main data structure of all the data (Caron et al., 2004). One important characteristic of grid data in ES community is that time is a dimension of the data structure (see Figure 3-1). This characteristic allows well representing spatio-temporal relation of continuous and complex observed phenomena in the real world.

However, regarding to the Earth location aspect, ES data are normally kept as simple as possible (Di et al., 2003). Generally, location is only recorded by coordinate in latitude and longitude. Reference datum is no mandatory. Hence, in the data structure, the reference datum is often not specified (URL.9). Most of the time, the role of spatial reference is only to specified the units of the coordinate variables (latitude, longitude). When the data lacks information about the datum, WGS84 datum is considered as the reference datum of the data set.
In GIScience community, vector and raster structure are used in either two or three dimensions to model real world phenomena. Phenomena and their location on the Earth are respected equally. Spatial reference that connects the position of the feature to the real world is important to GIScience. The spatial location of a feature needs to be referenced to a coordinate reference system (datum and coordinate system) to specify its position in the real world. However, considering the time aspect in geospatial data, it is only seen as an attribute. Due to this circumstance, static phenomena can be well represented but spatio-temporal relations are hard to be handled (see Figure 3-2).

The development of grid data model for continuous data is considered and still being continue in many organizations in many organizations of both Earth Science and Geo-Information Science communities. However, this development in each organization is in different degree of maturity. Unidata, with over 20 years experience in this field, includes almost all grid data providers and contains the largest grid data resources. Moreover, comparing two data structures of these two communities, Earth Science data structure includes time as a dimension. This is especially advantageous for the purpose of integration with trajectory data. Therefore, for this research, we decide to concentrate on multidimensional grid data of the ES community.
In the Earth Science community, Unidata is considered as the major organization that leads the developments of relevant technologies including technologies for grid data. Unidata is one of programs in the University Corporation for Atmospheric Research (UCAR) Office. The mission of Unidata is to offer data services, tools, and cyber leadership that strengthen Earth system science, increase educational opportunities, and enlarge participation. Being a diverse community of over 160 institutions with the common goal of sharing data and tools to access and visualize that data, Unidata has been practicing its missions for over 20 years (URL.7).

Unidata provides a set of services including client/server software and data management toolkits such as OPeNDAP protocol, OPeNDAP client software and libraries to build up client/server software (see Figure 3-3(left)). OPeNDAP stand for “Open-source Project for a Network Data Access Protocol”, it is a data access protocol commonly used in Earth science community. OPeNDAP will be introduced more detail in the next section. There are many servers that allow access to grid data using OPeNDAP protocol: NOAA, MODIS, TPAC etc. There are various OPeNDAP clients as available: CDAT, Pydap IDV and so on. Some OPeNDAP clients will be introduced and reviews in section 3.4.

![Figure 3-3: Unidata and OGC architectures](image)

In the Geo-Information Science community, the Open Geospatial Consortium, Inc (OGC) is the primary participant. OGC is an international industry consortium of 385 companies, organizations including government agencies and university together develop standards for interface. The aim of making standard for interface is to make spatial data and services interoperable with all kinds of
application. The OGC is considered as a global forum for developers and users cooperate to develop standards for geospatial interoperability (URL.10).

OGC also provides standards and a framework of web services such as Web Map services, Web Feature Services, Web Coverage Services and Web Catalog Services (Figure 3-3(right)). OGC Web Services allow distributed geo-processing systems to communicate with each other across the Web using familiar technologies such as Extensible Markup Language (XML) and Hypertext Transfer Protocol (HTTP).

For exploiting the strength of ES data in multidimensional grid format and the large availability of datasets in existing servers in the ES community, Web Coverage Service was added a standard-based interface to provide a gateway that help Web Coverage Service clients can access data sets from Earth Science community. There are some of OGC WCS libraries for working with coverage services in various programming and scripting languages such as OWSLib package in Python (Domenico et al., 2009).

Earth Science and Geo-Information Science communities have the same objective of making spatial data freely and widely available through web services. However, they have different concept regarding the support of the services. OGC only provides the definition and standards for the architecture of the web services. OGC is not involved with client-server methodologies and is not concerned with the implementation of the service. Therefore, we opt Unidata architecture as the best model to build up an integration system of trajectory and grid data since there are various protocols, supported tools and client software available.

Among them, the OPeNDAP is the one of most popular protocol with many available OPeNDAP servers, therefore, we decide to use the OPeNDAP protocol and grid data from OPeNDAP servers.

### 3.3. OPeNDAP (Open source Project for a Network Data Access Protocol)

The key issue in the distribution of science data is the incompatibility of data formats and web service interfaces. While the interoperability issue of web service interface is still being tried to handle, the problem relating to data formats has been addressed since the establishment of OPeNDAP.

The practice in the scientific data analysis is that data are stored in large amounts in different file formats, and each of them could be read and written only by its own specialized Application Programming Interface (API). A part of them uses common standards, while others remain in less standard formats. Most of these formats are incompatible with others. For instance, the netCDF and HDF formats are two file standard formats that are widely used in the realm of earth sciences. Nevertheless, with different file structures, netCDF file cannot be read by using HDF format and vice versa. These become obstacles for scientist to freely share their data and impede correspondence with central data archives as well.

OPeNDAP is a client-server protocol used for accessing and sharing of grid data. It allows users to access data from remote sources and then to store these data in the file format that users want. It also offers both a server software, to make data available on the remote servers, and a client software to
access those data. The main objective of OPeNDAP is to have a computer program that can operate on data provided by different institutions and stored in various format on a remote computer.

The OPeNDAP protocol defines method of communicating between an OPeNDAP client and server. It contains four components (URL.8):

- Firstly, an intermediate data representation is used for data transferring from the remote sites to the client. It consists of two main components: a set of simple data types (e.g. bytes, integers, strings, floating point numbers and a specialization of a string for holding an URL (Universal Resource Locator)).
- The next component is a format for the ancillary data. This format is called the Data Descriptor Structure (DDS). This format provides a description of the shape and the size of the different data types that are stored in several data sets. This format help to process of translating data set into the intermediate representation, then from the intermediate representation into the intended data model.
- The third component is a procedure for retrieving data and ancillary data from remote platforms.
- Finally, an Application Programming Interface (API) that consists of OPeNDAP classes and data access implements the protocol.

To describe the data that are sent by the OPeNDAP, two different kinds of message are used. The first one is the Data Descriptor Structure (DDS) that was mentioned above. The other one is the Data Attribute Structure (DAS). DAS contains all the other information that are necessary for client to recognize the data. The DDS contains information about the syntactic metadata that related with the structure of the data types, while DAS contains information about the meaning of data values.

According to (URL.8), OPeNDAP support various web services as following:

- Data Attribute service: returns a text file that describes the structure all data attributes of the specified dataset. This service will be triggered whenever server is requested with a Uniform Resource Locator (URL) including “.das” at the end.
- Data Descriptor service: returns a text file that describes the structure of all variables in the specified dataset. This service will be activated whenever server is requested with a URL that ends with “.dds”.
- OPeNDAP Data service: returns the actual data of specified dataset in a Multipurpose Internet Mail Extensions (MIME) document. This service will be activated whenever server is requested with a URL including “.dods” at the end.
- ASCII Data service: returns a representation of the specified data in ASCII format. This format makes the data available to most of recent web browsers. This service will be triggered whenever server is requested with a URL including “.asc” or “.ascii” at the end.
- WWW Interface service: produces an HTML form that store information from dataset. With this form, user can create a URL with which to ask for OPeNDAP data. WWW Interface service will be activated whenever server is requested with a URL that refers to a file or a directory and includes “.html” at the end.
- Information service: returns information about the dataset and server. This document is in HTML form and it may contain information about both data server and the dataset what
referred in the URL. A request with a URL including “.info” at the end will make this service active.

- Version service: returns information about version the OPeNDAP server software what is being used on the server. A request with a URL including “.ver” at the end will start this service.

3.4. OPeNDAP clients

3.4.1. OPeNDAP clients

This section is going to select an OPeNDAP client software for purpose of integrating data via web service. There are some of widely used OPeNDAP clients as introduced below:

Ferret: is a software for visualizing and analyzing large and complex gridded data. It provides the capabilities to access extensive remote databases by using OPeNDAP. Ferret is mostly used by the oceanographers and meteorologists. The most important thing of Ferret that others software do not have is the flexibility, geophysical formatting, intelligent connection with its database. Furthermore, it can manage large calculations and process in 4 dimensions.

GrADS: stands for The Grid Analysis and Display System. It provides tool for accessing, manipulating and visualizing earth science data. GrADS provides two models for gridded data and station data. It support many file formats (e.g. GRIB, NetCDF, HDF, BUFR, etc). GrADS provides mechanism for dealing with 5-Dimensional data environment including latitude, longitude, vertical level, time and grids. A data descriptor file is used in the data set. GrADS can handle regular grids, non-linearly spaced grids, Gaussian grids or variable resolution grids. It also can handle different data sets, overlay them graphically with correct spatial and time.

IDV: The Integrated Data Viewer is a Java- based software which developed by Unidata. It has capability to analyze and visualize many types of data as satellite images, grid data, radar data, shapefile, GeoTIFF etc.

MATLAB: according to URL.11, “MATLAB is a high level technical computing language”. It provides an interactive environment for developing algorithm, visualizing data, analyzing data and numeric computation. By using Matlab Structs Tool, Matlab is provided the way to read any OPeNDAP data. Matlab is a commercial package.

Pydap: is a Python library that implements the OPeNDAP. Pydap can act as a server for distributing the data. It also can act as a client to retrieve data from the remote servers on OPeNDAP servers. Hence, Pydap provides the abilities to scrutinise and operate a data set without considering it is local data or remote data (URL.12).

The selection of suitable client was based on the following criteria:

- Client should be extendable - should have scripting functionality
- Client should be able to read trajectory data
- It should be possible to integrate into client/server architecture
- It should support reading and analysing grid data
- Priority will be given to open source, non-commercial client software

Concerning the characteristics and ability of each OPeNDAP clients above, they satisfy most of given criteria, especially the capability to analysing grid data. However, only MATLAB and IDV are supported to read and analyse trajectory data (and other GIScience data as shapefiles, GeoTIFF, etc.). These clients are also supported by many build-in functions from their libraries (much more than others). Since MATLAB is a commercial package while IDV is a free and open-source software, we decided to choose IDV as the best client software for this research. The introduction about IDV will be given in the next section.

### 3.4.2. Integrated data viewer (IDV)

The hindering factors for data integration are the diversity of data format, and spatio-temporal resolution of datasets from multiple data sources. Consequently, the decision makers have to find methods to get a unified data format.

With regard to trajectory and grid data, the issues related to the data formats are become more complex. Trajectory data (or geospatial data in common) are two or three dimensional parameters; while metocean data like temperature, wind speed and wind direction, are usually considered as multidimensional variables. For performing study in movement behaviours or the relationship between moving objects and surrounding environment, GIS software is required where it allows to perform data analysis and to display those data. Conventionally, before grid data can be integrated with trajectory data using GIS software, users have to manually download, reference these data. In addition, GIS tools are not supported to deal with the variety of time and multidimensional data.

Due to the condition above, Unidata developed Integrated Data Viewer (IDV) - an application to meet the analysis needs of both Earth and Geo-Information science communities. IDV is equipped with tools that have capability for analyzing and visualizing where they allow users to combine metocean and other Earth science data along with societal data in GIS formats. IDV also provides features of both standard meteorological display and analysis programs with GIS tools. By using these, a large amount of meteorological data such as satellite imagery, gridded data, surface, upper air and radar data; and other Earth science data can be displayed at the same time with GIS data (shapefiles, Arc/Info, GeoTIFF and trajectory).

One of the main essential features of IDV is the ability to locate and work with geo-referenced datasets located on remote servers. It is performed by taking advantage of network access. Since the increase in the number of datasets that are accessible to the earth science community related to data from new sensors, higher spatial and temporal resolution data, it become impractical to distribute these datasets in real-time using conventional methods (Murray et al., 2003).

Through the framework developed by THREDDS (Thematic Real-time Environmental Distributed Data Services) project, the IDV users can access, and display datasets from a network of thematic servers. Using the metadata catalogs of the THREDDS framework, the IDV provides users with
options either selecting entire data sets, or selecting small subsets of data (spatial and temporal). The integrated nature of the IDV allows users to bring in data from multiple servers through a variety of protocols (e.g. FTP, OPeNDAP, ADDE, HTTP) and combine them into a single display.

IDV uses the VisAD library for data and display models as well as other Java utility packages. VisAD (Visualization for Algorithm Development) “is a Java component library for interactive and collaborative visualization and analysis of numerical data” (URL.6). VisAD provides a number of specific data structures like images, grids and tables. The VisAD data model defines a set of classes that can be used to build any hierarchical numerical data structures. Metadata including units, sampling topologies, error estimates and coordinate transformations is integrated into each data object and is carried beside through mathematical operations (Murray et al., 2004). Because the metadata contained in VisAD data objects include geolocation information, IDV can automatically reproject data from different source into a unified projection.

IDV allows users to create simple formulas by IDV Scripting Language or use the Jython language to execute calculations that are more complex. Jython is an implementation of Python programming language to run on Java environment.

3.5. Conclusion

In order to find the most suitable data model, data access protocol and client software, this chapter introduced two organizations, Unidata in Earth Science community and OGC in Geo-Information Science community. Due to the advantages of data structure, data availability and service supporting, grid data from Unidata participants was selected to use. Following this choice, OPeNDAP – a data access protocol and IDV client software are chosen and introduced. The capability for real application of these choices will be verified in the next chapter when we implement a prototype for the data integration system.
4. Designing and implementing the prototype for data integration system

4.1. Introduction

Following the choices of using grid data from Earth Science community, OPeNDAP protocol and IDV software as introduced in chapter 3, the present chapter discusses the solutions for building up a data integration system. The approach of this data integration system can be separated into two main parts. The first one is to construct a data integration mechanism that is done by using IDV software and Jython scripting language. The second one is to make this mechanism available as a web service. By this way, it can facilitate users from different areas on the world with their distinct research fields. Through graphical user interface, users can just send trajectory file, grid data source links and their requirement to server and server will return them an integrated dataset.

4.2. Desktop-based integration mechanism

4.2.1. Data integration

For purpose of integrating data by using IDV, trajectory file needs to be stored in local directory and data source links need to be indicated. Normally, one grid data source link can point to several data fields (so called list of data choices). For example, wind direction source link has the following form:

dods://motherlode.ucar.edu:8080/thredds/dodsC/modelsNc/NCEP/OCEAN/Global_5x2p5deg/OCEAN_Global_5x2p5deg_20091214_0000.nc

This link points to the list of data choices: u-component of wind at surface of the earth, v-component of wind at surface of the earth, significant height of combined wind waves swell at surface of the earth, etc.

From a data source link, its list of data choices can be obtained by using makeDataSource and getDataChoices methods from IDV library:

```python
data_source= makeDataSource(grid data source link, type of data)
list_of_data_choice= DataSource.getdataChoices(source)
```

The optional type parameter type of data in the method makeDataSource is used to specify the type of data. The purpose here is to get grid data from OPeNDAP servers, so the type of parameter should be “DODS.GRID”.

Assuming that we have text files that contain information about grid data source links, preferred data choices, chosen interpolation method and time accuracy, and also trajectory data file in the local directory, these files can be read by using Jython script as below:
file = open('Path/To/text_file_name.txt', 'r')
links = file.readlines()

The input trajectory data should be formed in a specific type. The trajectory dataset should be stored in text comma separated value (CSV) file format. To make IDV understand the trajectory dataset, the metadata has to be specified. The trajectory file needs two extra lines that define the field names within the file and the types of the fields. One more requirement for trajectory data is that there must be latitude, longitude and time field in the data. Input trajectory data should be like this:

(index) -> (ID, Time, Latitude, Longitude, Name (Text), Length)
ID, Time[dateformat], Latitude[unit="deg"], Longitude[unit="deg"], Name(text), Length[unit="km"]

1,1995-01-04 00:00:00,-76.50,-43.90,"A22A",66.672
2,1995-01-11 00:00:00,-76.25,-40.72,"A22A",66.616
3,1995-01-17 00:00:00,-75.46,-43.60,"A22B",66.521
4,1995-01-25 00:00:00,-56.20,44.20,"B22A",90.482

Trajectory data in a specify directory in local directory can also be read by using makeDataSource and getData methods as below:

trajectory_source = makeDataSource(trajectory data link, type of data)
trajectory_data = getData(data source name, data choice name)

One to be noticed here, the optional type parameter “type of data” in makeDataSource method should be “FILE.POINTTRACKTEXT”.

Depending on the particular characteristic of each environmental factor surrounding moving object, its value (e.g. temperature degree, wind speed, NDVI) may change slightly or largely following temporal dimension. Therefore, the time accuracy needed for each data field is different. In this prototype, with each data field, we give three options for users to select time accuracy: 1) time accuracy at second, 2) time accuracy at hour, and 3) time accuracy at day. For example, if the time accuracy at hour is selected, trajectory data will be integrated with grid data at the same hour of day (in case if grid data are available at that hour). If there are more than one value of data grid are recorded at the same hour, the first one will be used.

Normally, data fields such as temperature, wind speed or current speed can be represented by only one data choice. Thus, the value of grid data of those fields at the same location and same time with trajectory data can be extracted. However, cases of wind or current directions are different. The directions of wind or current are conventionally represented by two horizontal vector components, the “U” component represent the East-West component while the “V” component represents the North-South component. To get the directions, it is required to generate a direction grid from U and V component grids. This step is done by creating a grid with name “direction”:

direction = dirn( vecr (u-component grid, v-component grid))
The `dirn` method will return a grid of direction and the value of each grid cell is relative to the North. After creating the direction grid (for example: direction grid with name “direction” as above), we can use data extraction method method to get the direction value at given location and time.

Nevertheless, there is a difficulty when applying the data extraction method. Generally, grid data are heavy thus getting the whole grid dataset can lead to full memory problem in computer. The solution is that instead of getting the whole grid dataset at once, we get a subset of the grid corresponding to each trajectory record.

The size of iceberg is large (only icebergs with area are bigger than 10 nautical miles are stored in database (URL3)). Therefore, an iceberg may overlay with a number of grid cell. Since we have information about approximate values of iceberg’s dimensions, but we do not have any information about iceberg’s orientation, we assume that all icebergs are square with side length is equal to the dimension length of the iceberg and the coordinates of the iceberg are extracted from its centre.

The grid subset will be limited by grid cells what may overlay with assumed area of iceberg and the time as observation time. We know the coordinate at the centre of iceberg but we do not know its location in the grid cell (the centre of iceberg may locate at the edge of grid cell), hence the smallest area of subset data will be area of nine grid cells (one cell for the location of iceberg centre and eight for its neighbours).

If `glength` is the dimension length of iceberg and `relatlong[0]`, `relatlong[1]` are latitude resolution, longitude resolution values respectively, each time a record in trajectory data is read out, we can set and get subset grid by using script as below:

```python
dataSelection = DataSelection() 
steplat=int(glength/float(2*relatlong[0]))+1
steplong=int(glength/float(2*relatlong[1]))+1
jumplat= steplat * relatlong[0]
jumplong=steplong * relatlong[1]
dataSelection.setBounds(lat+jumplat,long-jumplong,lat-jumplat,long + jumplong)
times = ArrayList()
times.add(time)
dataSelection.setTimes(times)
data_subset= data_source.getData(datachoice,None,dataSelection,None)
```

After acquiring the grid subset, the value of grid data at specified location and time can be obtained by using data extraction method as mentioned above but this time, it is for only data subset. In case of wind or current direction, the grid subset of each U and V component grids need to be achieved before applying `verc`, `dirn` and data extraction methods.

With large moving object such as iceberg, only the values of grid data at centre point may not be enough to give a relationship indication between influencing factors and moving object. Therefore,
minimum, maximum and average values of grid data corresponding to the area of moving object are also acquired.

4.2.2. Data interpolation and uncertainty

4.2.2.1. Data interpolation

Frequently, a grid has missing cell values, therefore the interpolation methods are necessary for estimating missing values that lie in between the known cell values. There are many interpolation methods for weather data such as optimal inverse distance, cubic splining, kriging or cokriging etc. “These methods differ in their assumptions, local or global perspective, and deterministic or stochastic nature” and the phenomena itself (URL.14). The key for selecting an interpolation method is the understanding about data. Based on the nature of particular data and application domain, the corresponding interpolation methods should be applied.

However, for this research, we do not have the ambition to go deeply into interpolation issue and support users by all possible interpolation methods. For normal data fields such as temperature, wind speed or current speed, this prototype will offer only two methods for data interpolation. These two methods are: inverse distance weighting functions and interpolate from grid values of the record before and the record after the time of missing value. Inverse distance weighting function is aspatial interpolation methods. The foundation of it is the assumption that in space, points what closer to the centre point are more likely to have similar values than points what are further.

As the same assumption as explained in section 4.2.1 that the all icebergs are square with side length of square is equal to the dimensional length of iceberg and the coordinate of an iceberg is extracted at its centre. In case of the grid value at the centre point of iceberg is missing, we apply the inverse distance weighting functions to extract values of the grid that overlay with that iceberg in order to interpolate the missing value of the grid (see Figure 4-1).
Figure 4-1: Inverse distance weighting. In green, the assumed iceberg size. In dashed line, grid cells that iceberg overlay. In blue, the centre of measured grid cells.

Assuming that the distance from measured grid $i$ to the iceberg centre point is represented by $d_i$. Commonly, the weight factor applied in inverse distance weighting is the distance squared, and the formula is to estimate $m_0$ is:

$$m_0 = \frac{\sum_{i=1}^{n} \frac{m_i}{d_i^2}}{\sum_{i=1}^{n} \frac{1}{d_i^2}}$$

Where: $m_i$ is data value of neighbour cell $i$. This value is obtained by using “data extraction method” method (see section 4.2.1).

d_i is the distance between the centre of missing value cell and neighbour cell $i$. The distance $d_i$ is calculated as:

$$d = \text{radiusEarth} \times 2 \arcsin \left( \sqrt{\sin^2 \left( \frac{\phi_i - \phi_j}{2} \right) + \cos \phi_i \cos \phi_j \sin^2 \left( \frac{\lambda_i - \lambda_j}{2} \right)} \right)$$

Where: $\phi_i, \lambda_i$ are latitude and longitude in radian of the first location

$\phi_j, \lambda_j$ are latitude and longitude in radian of the second location

The Earth radius value used in this prototype is 7371.1 km

Interpolate from grid value of the record time before and the record time after: if the value of a grid cell is missing, the cell values of the record time before and the record time after at the same location will be used to interpolate. To get the record time before and record time after the time of missing value, we get the list of time of each data choice and find the index of time of missing value first. If TimeIndex is time index of time of missing value, we have time index of the record time before and record time after are (TimeIndex− 1) and (TimeIndex +1) respectively. Since the list of date time in IDV (object in this list is “Date” type) cannot be applied for Jython script to get index of object in the list, we make a copy of this list but objects in this list is “string” type. Once, we have time index of record time before and record time after time of missing value, we can set the “dataSelection”, get data subset and get grid data value as explained in section 4.2.1.

For direction data (vector data such as wind direction, current direction), due to the different of its nature, we support user to find the value of the nearest neighbour (in assumed area of iceberg) of missing point or/and the value of both record time before and the record time after the time of missing value. To find the value of the nearest neighbourhood, we use data extraction method method from IDV library with “samplingMode” parameter should be 100.

One disadvantage by using this method is that we do not know the location of this neighbour, therefore the uncertainty degree of interpolated values cannot be determined.
4.2.2.2. The uncertainty of interpolated data

As explained in chapter 2, uncertainty may be caused from many sources: from data collection through processing to its representation. The interpolation process gives a guess of a probable grid cell value, therefore the obtained cell value after interpolation is just an approximated value. The interpolation methods are always less than perfect and thus they bring a certain quantity of uncertainty to the estimated grid value. In other words, each estimated grid value is associated with a degree of uncertainty. To help user understand the quality of the interpolated dataset, the uncertainty degree of each estimated cell value is needed. To calculate the uncertainty degree, we built up a method that is based on two assumptions: firstly, measured values closer to the estimated value will have more influence than those are farther away and secondly, the more measured values are used to estimated, the more accuracy of the estimated value is.

The method to estimate uncertainty degree:

\[
U = \frac{S}{\sum d_i}
\]

Where:
- \( U \) - Uncertainty degree
- \( S \) - Total number of measured cells
- \( d_i \) - distance of measured cell \( i \) (1 cell size is used as 1 unit of distance)

For instance, the uncertainty degree of \( m_0 \) (Figure 4-1) can be estimate as:

\[
U_{m_0} = \frac{7}{\sqrt{3^2 + 3^2 + 3 + \sqrt{2^2 + 2^2} + 3 + \sqrt{3^2 + 2^2} + \sqrt{2^2 + 3^2} + \sqrt{3^2 + 3^2}}}
\]

These steps are done by using IDV software and Jython script.

All value of trajectory data fields and integrated, interpolated value are recorded in a variable. After finishing the whole integration and interpolation processes, this variable will be written to a CSV text file in format by using `writeFile` method:

\[
\text{IOUtil.writeFile(“Path/To/File.txt”, variable.toString())}
\]

CSV format is chosen for the format of integrated data set to return to users because it is a common and simple format. It is supported by all most all database management systems and programming languages.
4.3. Building up data integration web service

The previous section introduced data integration mechanism what is generated by using IDV and Jython script. To generate data integration mechanism, we assumed that trajectory data file, text files containing grid data source links, preferred data choices, interpolation method and time accuracy are already stored in local directory. However, the objective of this research is to support diverse users from research fields and different data integration needs through web environment. Users should not be required about knowledge of IDV or IDV programming, they are only required for the understanding about data source location they need. Therefore, we build up a web service base on the functionality of IDV and Jython language (see Figure 4-2). To embed IDV on the server side, two problems need to be solved. The first problem is how IDV software and scripts could be automatically evoked on a server when there are requirements from users. Another is how IDV could acquire trajectory data file, grid data source links and options for data interpolation, time accuracy that are specified by users through the user interface.

According to the manual, IDV and Jython script can be called through the command line as follows:

```
runIDV -script <Jython script>
```

Therefore, the important requirement for using server side scripting language is the ability to run a command line. Based on this requirement and the powerfulness of Python itself, Python was chosen as the server-side scripting language for our web service.

To build up the server, Apache HTTP Server software has been chosen. Apache is a popular, robust, commercial-grade and freely-available source code for an HTTP web server.

To incorporate Apache and Python, the mod_python module is used. Mod_python is an Apache module what embeds the Python interpreter in the server. Thus, it support Apache server the ability to execute Python code (URL13).

The model of web service system is illustrated in steps as in Figure 4-2 and the model of system in steps is illustrated as in Figure 4-3.
Figure 4-2: The conceptual model of data integration system via web service
Figure 4-3: The system model in steps
4.3.1. Graphical User Interface

The Graphical User Interface (GUI) allows users to provide their input to the system and to get back the response from the server (output). The GUI consists of HTML forms to acquire data source links, type of data (grid or image), and trajectory data file, preferred method for data interpolation and time accuracy (Figure 4-6), data choices (Figure 4-7) from users.

With the assumption that users are researchers, scientists or engineers who have knowledge about their fields, we let users find and choose data source links based on their own purposes. Data source link form permits users to give any kinds of OPeNDAP grid data such as metocean data (temperature, wind, current etc.) or NDVI etc. which they would like to integrate with their trajectory data. From literature review in chapter 2, we identified five main factors that affect iceberg behaviour namely sea surface temperature, wind speed, wind direction, current speed and current direction. Therefore, in this prototype, we limit the number of input box for data source links at seven (wind and current directions need four links). For other applications, this number can be increased or decreased. One more requirement, since metadata of grid data are not usually given enough, the spatial resolution information (geospatial longitude and geospatial latitude resolutions) will be asked to give by users.

Normally, as mentioned before, one grid data source link can point to several data fields. Thus, users need to specify which “data choice” they need from the grid data source link. With each data source link that is submitted by users, IDV and Jython script will be evoked to find list of data choices. After that, this list of data choices will be displayed on “data choices” form (Figure 4-7) and user can put their preferred choice in input box.

To get direction value (in case of wind and current directions), both two data choices that refer to U component and V component needs to put in input box in “data choices” form. To distinct with other data choices (not U or V component), user should type “UV” before type that data choice. The data choice for U component should be put first and the data choice for V component should be put afterward. Grid data which are pointed by these U and V components need to have the same latitude, longitude resolutions and time list (list of time records). The value of time record in time list and resolutions will be check by Jython script when IDV is evoked. If they are not equal, the integration and interpolation for these data choices will not be called to do. For example, with two data choices: 
* u-component of wind direction
* v-component of wind direction

user should type in the input boxes as:

UV
u-component of wind direction
UV
v-component of wind direction

Another form permits users to navigate and upload trajectory data file from their own local directory to the server. In this research, the iceberg trajectory data will be used.

The GUI also gives users options to choose the type of data interpolation method and time accuracy based upon the users’ preference in case web services return missing values for specified trajectory points (Figure 4-6). Users can choose methods for data interpolation such as a) interpolation by neighboring cell values for the same time and b) interpolation for the same cell location but between a record before and a record after the record of missing value; and c) they can select both these two
methods. About time accuracy, as explained in previous section, there are three options that are offered a) time accuracy at second; b) time accuracy at hour; and c) time accuracy at day.

After submitting the request through forms, users will retrieve the integrated dataset in a comma-separated values (CSV) form (Figure 4-3-step 19, 20).

### 4.3.2. Server side Python script

Python script was integrated with Apache by mod_python. It takes on three main roles: a) connecting client and server sides (or in other words is receiving input from users and sending output back); b) writing data and requirement from user to text files (these text files will be read by using IDV and Jython script); and c) evoking IDV and Jython script through command line.

Once, users upload trajectory data file and provide the source link, choose the interpolation method and time accuracy, Python script reads these data and requirements as variables (Figure 4-3-step 1) and writes them into distinct text files on the server (Figure 4-3-step2).

After writing text files in a specified directory, through the command line, Python script will change the default directory path on the command line to a directory path of IDV and then call IDV software and Jython script to run (Figure 4-3-step 3, 4). These steps are done by using Python script in Windows operating system as below:

```python
import os
os.chdir("C:\Program Files\IDV 2.7u2")
os.system("runIDV –script readlink.py")
```

The Jython script “readlink.py” will read data source links in text file (which was written in above steps) in server (Figure 4-3-step 5) and connect to OPeNDAP servers to get lists of data choices of these data source links (Figure 4-3-step 6). Also, lists of data choices are written in a text file by Jython script (Figure 4-3-step 7). The text file that contains lists of data choices will be read by Python script and show for users through HTML form (Figure 4-7) (Figure 4-3-step 8, 9). Here, once again, Python script read user requirement (user’s preferred data choices) as variables, write these variables into a text file on server and then call IDV and Jython script to do next steps (Figure 4-3-step 10,11,12,13).

After processes of integration and interpolation are done (Figure 4-3-step 17), the integrated dataset is write into a text file on server. This integrated data are read and returned to user through GUI by using Python script:

```python
def return_data(req):
    req.content_type="text/plain"
    req.sendfile("Path/To/integrated_file.txt")
```
After returning the last output, Python script is used to delete all of text files what were written from processes before. For this step, we use script as:

```python
import os
os.remove('Path/To/File.txt')
```

### 4.4. Results

The results of implementing the prototype for data integration system will be introduced as an application for iceberg case study.

The case study area of this research is Antarctica- the Earth’s southernmost continent. Antarctica is surrounded by the southern Pacific, Atlantic and Indian Oceans. On average, it is the coldest, driest and windiest continent. The total area is 14 million kilometres square. However, 98 percent of this area is covered by ice with average at least 1.6 kilometres in thickness. This ice amount is 90 percent of the world’s ice and contains about 70 percent of the world’s fresh water (URL.16).

Since the nature of Antarctic area, there are no human residents permanently reside there. Only cold-adapted plants and animals including penguins, seals and some types of algae can inhabit. However, there are quite a lot of researchers and scientists dwell at some research stations scattered around the continent. It is because the important role of Antarctica and Antarctic iceberg on climate regulation. Studying iceberg behaviours in relation with surrounding environmental factors can help to answer many remained questions about climate change, sea vegetation, etc.

As introduced before, Antarctic iceberg data are freely provided through web environment by NIC (National Ice Centre) from 1978 to present (see Figure 4-4).

#### Antarctic Icebergs

<table>
<thead>
<tr>
<th>Iceberg</th>
<th>Date Updated</th>
<th>Image Source for Update</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Size (NM)</th>
<th>Archived Image</th>
<th>Date of Archived Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-22A</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>78.07 S</td>
<td>41.42 W</td>
<td>60 X 21</td>
<td>A22A</td>
<td>2009929</td>
</tr>
<tr>
<td>A-27</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>78.07 S</td>
<td>41.39 W</td>
<td>10 X 08</td>
<td>A27</td>
<td>2009929</td>
</tr>
<tr>
<td>A-45D</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>71.21 S</td>
<td>59.26 W</td>
<td>21 X 15</td>
<td>A45D</td>
<td>2009154</td>
</tr>
<tr>
<td>A-43K</td>
<td>2009971</td>
<td>MODIS</td>
<td>52.15 S</td>
<td>14.38 W</td>
<td>22 X 06</td>
<td>A-43K</td>
<td>2008316</td>
</tr>
<tr>
<td>A-5</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>63.10 S</td>
<td>57.41 W</td>
<td>14 X 04</td>
<td>A5</td>
<td>2008165</td>
</tr>
<tr>
<td>A-1</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>77.64 S</td>
<td>58.10 W</td>
<td>14 X 07</td>
<td>A1</td>
<td>2008029</td>
</tr>
<tr>
<td>A-8</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>77.56 S</td>
<td>46.42 W</td>
<td>10 X 10</td>
<td>A8</td>
<td>2008029</td>
</tr>
<tr>
<td>A-58</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>72.28 S</td>
<td>76.41 W</td>
<td>13 X 01</td>
<td>A58</td>
<td>2008352</td>
</tr>
<tr>
<td>A-59</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>78.18 S</td>
<td>72.56 W</td>
<td>13 X 04</td>
<td>A59</td>
<td>2008352</td>
</tr>
<tr>
<td>A-60</td>
<td>2009971</td>
<td>ENVISAT</td>
<td>78.21 S</td>
<td>72.59 W</td>
<td>13 X 02</td>
<td>A60</td>
<td>2008352</td>
</tr>
<tr>
<td>B-09B</td>
<td>2009970</td>
<td>MODIS</td>
<td>67.14 S</td>
<td>182.21 E</td>
<td>51 X 20</td>
<td>B09B</td>
<td>2009970</td>
</tr>
<tr>
<td>B-15A</td>
<td>2009936</td>
<td>MODIS</td>
<td>53.38 S</td>
<td>159.08 W</td>
<td>40 X 07</td>
<td>B15A</td>
<td>2008352</td>
</tr>
<tr>
<td>B-15B</td>
<td>2009970</td>
<td>ENVISAT</td>
<td>68.08 S</td>
<td>77.35 E</td>
<td>38 X 10</td>
<td>B15B</td>
<td>2008352</td>
</tr>
<tr>
<td>B-11F</td>
<td>2009970</td>
<td>ENVISAT</td>
<td>65.77 S</td>
<td>80.14 E</td>
<td>20 X 07</td>
<td>B11F</td>
<td>2009928</td>
</tr>
<tr>
<td>B-10G</td>
<td>2009970</td>
<td>ENVISAT</td>
<td>61.28 S</td>
<td>80.14 E</td>
<td>20 X 11</td>
<td>B10G</td>
<td>2008928</td>
</tr>
<tr>
<td>B-11H</td>
<td>2009970</td>
<td>MODIS</td>
<td>66.03 S</td>
<td>100.04 E</td>
<td>13 X 02</td>
<td>B11H</td>
<td>2009970</td>
</tr>
</tbody>
</table>

Figure 4-4: Antarctic iceberg data provided by NIC
Since there are some error and duplications in datasets from NIC, we use a dataset which was downloaded from NIC and cleaned by Ipopo (Ipopo, 2009). To use for our web service, this data set needs to be restructured to the format as IDV can read (as explained in section 4.2). A snapshot of the restructured data set is showed below (Figure 4-5):

![Figure 4-5: Iceberg trajectory dataset](image)

About influencing factors of icebergs, as identified in chapter 2, there are five main weather factors that affect to iceberg calving and movement. Due to the inavailability of grid data for Antarctic area, we could find data sources for only two in five factors: sea surface temperature and wind direction from TPAC. TPAC stands for Tasmanian Partnership for Advanced Computing, and “it is a member of the Australian Research Collaboration Service (ARCS) who aim to provide long-term eResearch support services for the Australian research community” (URL.15). Three data links of grid data in 1995 (wind direction data need two data links for U and V components) are used:

- **Grid data source link of U component of wind in 1995:**

- **Grid data source link of V component of wind in 1995:**

- **Grid data source link of sea surface temperature in 1995:**

For integrating these influencing factors with iceberg trajectory data, we use prototype of web service as introduced in chapter 4. Firstly, these data links, spatial resolutions are put in input boxes as in Figure 4-6.
Data integration service

Please provide OPeNDAP grid data source links:

<table>
<thead>
<tr>
<th>Data source 1:</th>
<th>Data source 2:</th>
<th>Data source 3:</th>
<th>Data source 4:</th>
<th>Data source 5:</th>
<th>Data source 6:</th>
<th>Data source 7:</th>
</tr>
</thead>
<tbody>
<tr>
<td>link: ...</td>
<td>link: ...</td>
<td>link: ...</td>
<td>link: ...</td>
<td>link: ...</td>
<td>link: ...</td>
<td>link: ...</td>
</tr>
<tr>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
<td>spatial resolution: lat</td>
</tr>
<tr>
<td>long</td>
<td>long</td>
<td>long</td>
<td>long</td>
<td>long</td>
<td>long</td>
<td>long</td>
</tr>
</tbody>
</table>

Please select interpolation method:

- Inverse distance weighting

Time accuracy at: Day

From each given data link, IDV connect to OPeNDAP server and get a list of data choice corresponding to each data link:

- The list of data choices from data source link of U-component of wind in 1995:
  4xDaily u-wind at sigma level 995
- The list of data choices from data source link of V-component of wind in 1995:
  4xDaily v-wind at sigma level 995
- The list of data choices from data source link of sea surface temperature 1995:
  Geographic longitude of T_cell centers, Geographic latitude of T_cell centers, noname

These lists of data choices is displayed on another HTML form, by doing that, it can help user find data choice easier. Data choices of data source link 1 and 2 point to U and V components of wind direction, thus to make system understand that user want to get direction of wind, in front of each of data choice user need to put “UV” (Figure 4-7).
Data integration service

Please provide data choice:

Data source 1: dods/ingortal.ifatis.edu.au-1/thredds/dodsC/library/iceolp/surface/wwnd.s95/wwnd.s95.1995.nc
List of data choices: [4xDaily u-wind at sigma level 995]
Data choice: 4xDaily u-wind at sigma level 995

List of data choices: [4xDaily v-wind at sigma level 995]
Data choice: 4xDaily v-wind at sigma level 995

List of data choices: [Geographic longitude of T-cell centers, Geographic latitude of T-cell centers, None]
Data choice: None

Web service server returns an integrated dataset in CSV format (see Figure 4-8).

<table>
<thead>
<tr>
<th>Latitude</th>
<th>Longitude</th>
<th>Time</th>
<th>length Iceberg</th>
<th>data 1 max</th>
<th>data 1 min</th>
<th>data 1 average</th>
<th>direction</th>
<th>data 1 max</th>
<th>data 1 min</th>
<th>data 1 average</th>
</tr>
</thead>
<tbody>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-01-04 00:00:00</td>
<td>66.67 A2A</td>
<td>-15.993</td>
<td>-15.993</td>
<td>-15.993</td>
<td>183.493</td>
<td>358.577</td>
<td>3.122111</td>
<td>127.079</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-01-11 00:00:00</td>
<td>66.67 A2A</td>
<td>-6.496751</td>
<td>-2.927652</td>
<td>-3.649869</td>
<td>104.7268</td>
<td>225.4914</td>
<td>11.88981</td>
<td>131.0215</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.4</td>
<td>1995-01-17 00:00:00</td>
<td>66.67 A2A</td>
<td>-15.78204</td>
<td>-7.88324</td>
<td>-12.131</td>
<td>292.6195</td>
<td>299.5644</td>
<td>48.814688</td>
<td>200.4441</td>
</tr>
<tr>
<td>-76.5</td>
<td>-44.2</td>
<td>1995-01-25 00:00:00</td>
<td>66.67 A2A</td>
<td>-5.47764</td>
<td>-2.18189</td>
<td>-3.8323</td>
<td>233.837</td>
<td>42.51044</td>
<td>149.5956</td>
<td></td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-02-02 00:00:00</td>
<td>66.67 A2A</td>
<td>-20.61946</td>
<td>-9.5992</td>
<td>-16.9646</td>
<td>313.6767</td>
<td>39.0857</td>
<td>153.4937</td>
<td></td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-02-08 00:00:00</td>
<td>66.67 A2A</td>
<td>-10.88844</td>
<td>-5.22415</td>
<td>-8.703</td>
<td>104.6477</td>
<td>146.3107</td>
<td>204.775</td>
<td>100.7844</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-03-14 00:00:00</td>
<td>66.67 A2A</td>
<td>-21.64503</td>
<td>-10.7312</td>
<td>-16.353</td>
<td>176.2686</td>
<td>259.6951</td>
<td>140.3098</td>
<td>157.8428</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-02-22 00:00:00</td>
<td>66.67 A2A</td>
<td>-18.0462</td>
<td>-7.11462</td>
<td>-12.137</td>
<td>299.6046</td>
<td>241.3296</td>
<td>161.5631</td>
<td>150.2396</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-03-01 00:00:00</td>
<td>66.67 A2A</td>
<td>-24.50572</td>
<td>-22.5609</td>
<td>-23.532</td>
<td>143.5817</td>
<td>83.2538</td>
<td>2.945594</td>
<td>34.69547</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-03-10 00:00:00</td>
<td>66.67 A2A</td>
<td>-17.17564</td>
<td>-15.1281</td>
<td>-15.12</td>
<td>216.4373</td>
<td>220.4692</td>
<td>190.0935</td>
<td>217.6493</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-04-04 00:00:00</td>
<td>66.67 A2A</td>
<td>-51.2715</td>
<td>-44.9649</td>
<td>-56.255</td>
<td>221.9871</td>
<td>225.1849</td>
<td>71.565099</td>
<td>142.2114</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-04-10 00:00:00</td>
<td>66.67 A2A</td>
<td>-20.95339</td>
<td>-25.54</td>
<td>-29.147</td>
<td>228.3640</td>
<td>242.3189</td>
<td>191.31</td>
<td>221.0315</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.9</td>
<td>1995-04-17 00:00:00</td>
<td>66.67 A2A</td>
<td>-35.3667</td>
<td>-31.7421</td>
<td>-38.581</td>
<td>140.7583</td>
<td>147.45</td>
<td>111.3706</td>
<td>127.0301</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.7</td>
<td>1995-05-07 00:00:00</td>
<td>66.67 A2A</td>
<td>-18.75891</td>
<td>-10.1374</td>
<td>-16.337</td>
<td>183.7892</td>
<td>220.6593</td>
<td>240.7512</td>
<td>189.9042</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.7</td>
<td>1995-05-30 00:00:00</td>
<td>66.67 A2A</td>
<td>-32.6876</td>
<td>-28.6782</td>
<td>-37.599</td>
<td>32.5367</td>
<td>193.6271</td>
<td>207.8973</td>
<td>125.3113</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.7</td>
<td>1995-06-06 00:00:00</td>
<td>66.67 A2A</td>
<td>-32.1044</td>
<td>-26.9177</td>
<td>-37.804</td>
<td>32.0383</td>
<td>211.7594</td>
<td>824.4621</td>
<td>167.7749</td>
</tr>
<tr>
<td>-76.5</td>
<td>-43.7</td>
<td>1995-06-13 00:00:00</td>
<td>66.67 A2A</td>
<td>-12.80557</td>
<td>-11.4632</td>
<td>-14.109</td>
<td>162.4744</td>
<td>170.8034</td>
<td>137.8784</td>
<td>153.0235</td>
</tr>
<tr>
<td>-76.7</td>
<td>-43.8</td>
<td>1995-06-23 00:00:00</td>
<td>66.67 A2A</td>
<td>-14.64826</td>
<td>-14.3462</td>
<td>-15.346</td>
<td>28.4862</td>
<td>117.7152</td>
<td>155.9296</td>
<td>61.8541</td>
</tr>
<tr>
<td>-76.7</td>
<td>-43.8</td>
<td>1995-06-28 00:00:00</td>
<td>66.67 A2A</td>
<td>-27.9857</td>
<td>-26.7553</td>
<td>-29.782</td>
<td>28.4264</td>
<td>75.9636</td>
<td>110.7723</td>
<td>29.05468</td>
</tr>
<tr>
<td>-76.7</td>
<td>-43.8</td>
<td>1995-07-11 00:00:00</td>
<td>66.67 A2A</td>
<td>-23.21201</td>
<td>-21.8024</td>
<td>-26.218</td>
<td>23.5981</td>
<td>215.8764</td>
<td>220.6809</td>
<td>140.7107</td>
</tr>
</tbody>
</table>

Figure 4-8: Integrated dataset (opened in Microsoft Excel)

To illustrate attributes sea surface temperature, wind direction of integrated dataset, we use Google Earth software to visualize iceberg trajectories. Icebergs are represented by arrow symbol. Wind direction attribute is represented by arrow’s orientation and sea surface temperature attribute is represented by colour saturation. According to the integrated dataset, values of sea surface temperature vary from – 53.620 to 1.713 C degree, we classify these data into five groups corresponding to five colour saturations (see Figure 4-9).
4.5. Evaluation

The main objective of the study have fulfilled by making a data integration mechanism available as a web service. There is an issue relating to the performance of system needs to be concerned. It is the time consumption for processing of integration mechanism on server. While was designing the prototype, two scenarios were raised. The first scenario, the whole grid data are obtained and integration, interpolation processes will be done with this data grid, the advantage of this scenario is the system have to connect to server and download grid data only once. The disadvantage is that it requires a powerful server with large memory and the process of integration or interpolation will have to work with a big grid, this will consume more time for processing (comparing to process with small data grid). The second scenario, instead of obtain the whole data grid, only subset data (which based on assumed area of iceberg and record time) is achieved. The advantage and disadvantage of this scenario are opposite with the first one. It does not require much memory of server, however it needs to download data many times (depends on the number of trajectory records).

Unfortunately, for this prototype, we did not have chance to try both of these scenarios. This prototype was build on a personal computer (Intel Core 2 Duo CPU 2.26GHz) with 512 megabytes are used for Java memory, therefore we had to select the second scenario. There were 289 iceberg trajectory records and three data sources links were used (sea surface temperature, U and V components of wind direction). The processing time was about 680 seconds (11 minutes) in total. The time consumption for downloading one data subset was 0.75 second. The time consumption for getting grid value at specified location and time was 0.0009 second (using data extraction method, see section 4.2). When applying iceberg trajectories and grid data (sea surface temperature, U and V components of wind direction), the server had to download data subset 867 times and the method data extraction method was used 5780 times.
4.6. Conclusion

The prototype of a client-server system for integrating grid and trajectory data was implemented and fulfilled in this chapter. The system includes three main parts. Firstly, graphical user interface part containing HTML forms allows users upload their trajectory data file and put grid data source link to request integrating process. The second part is the middleware, Python script that is integrated with Apache server by mod_python module is used for connecting information between client and server. The last one is the integration and interpolation processes on server side. This part is done by using IDV and Jython script. Some interpolation methods such as inverse distance weighting, nearest neighbour, interpolation by value at different record time, and issues of uncertainty in applying interpolation method are also considered. The integrated data set is visualized in Google Earth to illustrate result of integration process. The performance of prototype is evaluated at the end of chapter.
5. Conclusions and recommendations

5.1. Introduction
The relations between behaviour of a moving object and its surrounding environment are inseparable - they have influence on each other. The development of tracking devices, satellite technology, grid data exchange mechanism and web services bring unprecedented opportunity for researchers, scientists to discover these relations. However, the incompatibility between data formats and spatial-temporal resolutions is a major barrier for the data integration. The main objective of this research was to find a solution to support users integrating grid and trajectory data via a web service. The previous chapters described that lead to fulfilment of this objective. The present chapter discusses the research conclusions based on research objectives and questions, and finally, recommendations are proposed for further researches.

5.2. Conclusions
The conclusions of this study can be summarized based on the research questions as follows:

1. a) Who are the user and what are the requirements for integrating different spatial temporal datasets?

In chapter two, we identified and classified users of icebergs and influencing factor into three main groups as follow:
- Institutions and groups that involve in data collections and monitoring
- Navigators, navigation planners and engineers
- Scientists such as oceanographers and climatologists

These groups pose different questions as:
- When, where and how does iceberg calving occur?
- What are the expected speeds of iceberg movement in different areas?
- When and where does the calving process occur frequently?
- What is the effect of weather factors on iceberg movement?

These questions commonly concern about the behaviours of icebergs (movement, calving) in the relationship with surrounding environment in spatial-temporal dimensions. To address these questions, the main requirement is users need to be supported by enough data sets of iceberg and influencing factors.

b) What kind of metocean data needs to be derived?
There are many factors influence on iceberg behaviours, we identified wind speed, wind direction, current speed, current direction and sea-surface temperature as five main factors and need to be derived.

2. **a) Where can climate data be extracted from?**

Grid data exchange mechanisms are developed by both Earth Science and Geo-Information Science communities. However, there are three main advantages of Unidata mechanism (a dominant organization in Earth Science community). Firstly, data structure containing time as a dimension allow easy handling and representation of spatial temporal relations of continuous and complex phenomena observed in the real world. Secondly, Unidata, with over 20 years experience in this field and including almost all grid data providers, contains the largest grid data resources. Thirdly, Unidata provides a set of services including client/server software and data management toolkits such as OPeNDAP protocol, OPeNDAP client software and libraries to build up client/server software. Therefore, for this research, we decide to use multidimensional grid data from participants of Unidata/UCAR organization such as TPAC (Tasmanian Partnership for Advanced Computing).

**b) What is the remote data access protocol?**

This research used OPeNDAP (Open-source Project for a Network Data Access Protocol) protocol to access grid data from remote OPeNDAP servers. OPeNDAP is a common client-server protocol in Earth Science community. It allows users to access unknown format data on remote server and then store these data in the format they want.

3. **How are the data formatted? What are the spatial and temporal resolutions of the data sets?**

This study involved type of data sets, namely trajectory and weather grid data. The input trajectory data set should be stored in text comma separated value (CSV) file format. The file needs two extra lines that define field names within the file and types of the fields. There must be latitude, longitude, and time field in the data. One more requirement for iceberg trajectory data is that the field of dimension length of iceberg needs to be included. The temporal resolution of icebergs is irregular and varies from 1-5 days.

Grid format is generally thought as a set of numbers making up a two, three or multidimensional array of parameter values. The basic unit of the grid is the “cell” and each cell contains one value of measured parameter. There are some different types of grid as ERSI Coverage Format, Arc/INFO Grid, etc. however most of them consist with an explanatory header to give information about the grid (e.g. number of rows, number of columns, maximum and minimum values, etc.). Grid data is usually provided at daily temporal resolution, the sea-surface temperature and wind direction grid data which were used for integrating with iceberg trajectory data are provided 4 times a day by TPAC.

4. **a) What kind of spatial statistical climate data value is needed (max, min or average)?**

With a large moving object such as iceberg, the values of grid data at only centre point may not be enough to give a relationship and indication on the factors influencing movement of an iceberg. Therefore, minimum, maximum and average values of grid data corresponding to the area of moving object are need to be acquired.
b) Which integration methods to be applied for integrating climate data in relation to iceberg data?
The integration process was done by using IDV and Jython script. With each iceberg trajectory record, subsets of grid data will be obtained. The sizes of these subsets are based on the area of iceberg with assumption that all icebergs are square with side length equal to the dimension length of the iceberg and the coordinates of the iceberg are extracted from its centre.

c) Which interpolation methods to be applied for integrating climate data in relation to iceberg data?
For normal grid data, we give user options to select two types of interpolation: The first one is inverse distance weighting functions (one of spatial interpolation methods). The foundation of this method is the assumption that in space, points closer together are more likely to have similar values than points more distant. The second one is interpolation from grid values of the record time before and the record time after the time of missing value.

For grid data contain vector (such as wind direction), users are supported to find the value from the nearest neighbour or find both values of record time before and record time after the time of missing value.

5. Which method to be applied for determining the uncertainty and error quantification of integrated data?
Since we could not find any reference dataset to quantify the quality of integrated data set, most of attributes in integrated dataset have not qualified. For interpolated data, an uncertainty degree is estimated and included in output file to return to users.

6. How to evaluate the prototype?
The prototype - a web service with capability of integrating trajectory data with grid data from remote servers has been fulfilled research objective. Considering the time consumption of prototype performance, two scenarios relating to methods for downloading data are proposed (see section 4.5). Due to the limitation of hardware, we could not implement both these scenarios. The time consumption for the implemented scenario is recorded. However, another scenario still needs to be implemented to compare. From there, the best scenario can be selected.

5.3. Recommendations for further research
The fundamental research problem has solved, we propose some recommendations for the development of better web service of data integration as follow:

It is well known that iceberg’s behaviour is influenced by many factors, especially: current speed, current direction, sea surface temperature, wind speed, wind direction, however, due to the data inavailability for Antarctic area, only grid data of sea surface temperature and wind direction were used in this prototype. Thus, to understand iceberg’s behaviour, additional data of wind speed, current speed, current direction are required.
There is a problem that may happen for client-server system when applied in real use. In case there are many people users using our web service at the same time, text files that were created to save data information and user’s options may be deleted or data will be sent to wrong users. This problem can be solved by using different text file names with different users. These text file names can be named base on cookie or user’s names when they log in. Due to the time limitation, we leave this problem for further work.

Two scenarios for downloading grid data are introduced. To select the better one, both these scenarios should be test on a real server computer. The criteria should be the time consumption and also the stability of server in case many users request service at the same time.

The web service for data integrating could be more useful when helping user to find or prompt data source links. When was designing the prototype of system, we tended to create a catalog of data sources or using catalog from some data services such as THREDDS. By this way, it will facilitate user when finding data sources link and metadata as well. However, to be implemented, it requires more time than our limitation. Therefore, we suggest this step for continuing researches.

Data interoperability is always an important issue. Considering the trend in both Earth science and Geo-Information science communities, they are moving closer to each other on the issues of standardization. It is a promising trend for this prototype to extend its capability to support user integrating many other types of data from both these communities.
6. References


IPOPO, J. 2009. Visualizing spatio-temporal data sets with different temporal resolutions. MSC, ITC.


URLs


