Vanden Berghe E., M. Brown, M.J. Costello, C. Heip, S. Levitus and P. Pissierssens (Eds). 2004. p. 153-162
Proceedings of 'The Colour of Ocean Data' Symposium, Brussels, 25-27 November 2002
IOC Workshop Report 188 (UNESCO, Paris). x + 308 pp.
– also published as VLIZ Special Publication 16

# BaltCom Datawarehouse – Online data mining using MS Analysis Services

T. Jansen, H. Degel and J. Heilmann

Danish Institute for Fisheries Research
Jægersborgvej 64-66, 2800 Lyngby, Denmark
E-mail: tej@dfu.min.dk

## Abstract

BaltCom version 2.0 is an Internet based datawarehouse where the countries represented in IBSFC (International Baltic Sea Fishery Commission) upload, download, validate and analyze fisheries data. The development of the international datawarehouse was an EC funded project. In the process of assessing stocks of commercially fished species in the Baltic Sea [mainly Cod (*Gadus morhua*), Herring (*Clupea harengus*), Sprat (*Sprattus sprattus*) and Flounder (*Platichthys flesus*)] data needs to be aggregated and calculated to match the input formats required to run assessment software used in the Baltic Fisheries Assessment Working Group. Further ad hoc exploratory data mining is needed when decision makers request special analysis or when determining the coverage and the quality of the data and to expose details hidden in the data in order to explain and understand model results. It is therefore important to be able to analyse data online, examining data in different forms on any aggregation level, slicing, dicing, rolling up or drilling down the data. The need for a fast performing web based data mining application to analyse these data formats and derivates on different aggregation levels was addressed by implementing a solution using Microsoft Analysis Server and Microsoft Excel Pivot Table Services. Microsoft SQL server was selected as the database server, while the rest of the application was built on .net technology. The functionality and architecture of the solution is presented. The selection of technology is discussed. Conclusions and recommendations are given on the basis of lessons learned during development and implementation. The project indicates that it could be fruitful for the biological community to integrate some of the many existing data mining and OLAP systems in conjunction with the many new international web based datawarehouses.

Keywords: BaltCom; OLAP; Datawarehouse; Data mining; Fishery.

## Introduction and requirement description

OLAP (Online Analytical Processing) and data mining systems can be used to get information hidden in large databases extracted and displayed rapidly. Possibilities to analyse vast amounts of data through a user-friendly interface, can reveal new knowledge and facilitate the understanding of existing knowledge.

OLAP and data mining techniques have emerged in the business world through the 90's (Dunham, 2003), and are still in an early phase of development (Han and Kamber, 2001). A comprehensive review of major OLAP and data mining systems is to be found in Dunham (2003). The systems are mainly based on well known statistical methods, such as regression,

clustering, classification and neural networks, which have been widely used throughout the biological community.
In the present project only a fraction of the possibilities in such systems have been used, but the project will give indications of the usefulness of the software if used in biological sciences.

## About BaltCom

From 1995 all discard and landing data from the fishery in the Baltic Sea have been sampled during three succeeding international projects financed by the European Commission, Directorate-General FISH, Fisheries. All countries around the Baltic Sea participated. Besides being a sampling project the second and third project dealt with the development of a common database holding the data collected (IBSSP I, IBSSP II).

The overall aim of the projects was to improve the quality of the stock assessment of Baltic cod and to this end a consistent sampling of catch data in all the different countries was a major point. A very important tool for this was a common database holding all data and defining the minimum quality level. A common agreed sampling manual laying down the sampling procedure supported this. Before 1998 each country did its own sampling more or less independent and not just the sampling but also the following raising procedure from the sample level to the total national catch was made in many different ways making the transparency of the internationally aggregated data very poor. A common database does not just give easy access to information on sampling level but makes it possible to raise the sample results to the total level on a consistent and well-documented way.

The first version of the database was a plain ASCII file containing all parameter values in a fixed format. All data handling and calculating was made using SAS programming. Every half-year the data from each country were submitted to the database responsible who simply appended the new data to the old data after having performed a quality check. As the amount of data grew, it became more and more difficult to handle the database and therefore the development of a web based database was included in the last project.

## Input to fisheries assessment models and exploratory OLAP

Biological advises for managing the fish stocks are based on fish stock assessment models analysing the state of the stocks. The sampling results raised to national level and internationally aggregated to fish stock level are the input to these fish stock models. The standard models are based on the characteristics of the decay over the years of each age class existing in the fish stock. The level of each age class in a given year is given in numbers per strata (typically species, area, quarter). The declining in numbers from one year to the next is assumed to be a result of a mortality caused by the fishery (fishing mortality) and a mortality caused by other reasons (natural mortality). The fishery mortality is assumed to be the sum of the landings and the discards. Based on the catches by age group and the natural mortality by age group the stock numbers can be estimated.

The input is the following information on the catches:

- the number of individuals caught by the fishery by age group;
- the mean weight by age group in the catches.

During sampling of the catch, length frequencies of the catches are obtained and as the landings on national level are registered in total weight, it is converted into numbers by age group in a two step procedure: firstly, the total landings (in weight) are converted to numbers by length group using the length distributions obtained during sampling. Secondly, the number by length group is converted to number by age group by using an age-length key also obtained during sampling. Both the length distributions and the age-length keys are calculated based on the data held in the database. As a diagnostic of the state of the stock and in order to be able to express the stock characteristics, not just in numbers, but also in biomass; the mean weight by age group is calculated from the sampling data where individual weights are obtained.

Beside the standard input to the assessment models a long list of exploratory tabulating of the data available was needed. These are made when decision makers request special analysis or the purpose can be to determine the coverage and the quality of the data and to expose details hidden in the data in order to explain and understand model results.

It was therefore very important to be able to analyse data online, examining data in different forms on any aggregation level, slicing, dicing, rolling up or drilling down the data.

## Materials and methods

### *Selection of technology*

Microsoft SQL server was selected as the database server, while the OLAP solution was created using Microsoft Analysis Server and Microsoft Excel Pivot Table Services. The rest of the application was built on.net technology using XML standards like XML, XSD and SVG. The source code will be open.

The website can be found at the URL: http://www.BaltCom.org. The application is password protected.

### *Architecture and implementation*

The part of the solution where data are uploaded from user to database, validated, cleansed, application security and other functionality is only described briefly to give understanding of the whole solution. A detailed description is out of the scope of this paper; see Degel and Jansen (2003) for a full documentation.

On Fig. 1 the dataflow from user to database is illustrated. Users zip their ASCII-files and upload them to BaltCom by selecting tools>upload in the menu. The optional zipping of the file reduces the upload time by up to 90%. On the server the file is unzipped and converted to XML and validated against a XSD schema file. If data is imperfect, the process is stopped here and the user gets a validation report. If no errors are detected the data is saved in the database. Further data quality analysis and cleansing is offered by an outlier mining functionality. The schema file validation is described in detail in Sandbeck *et al.* (2003).

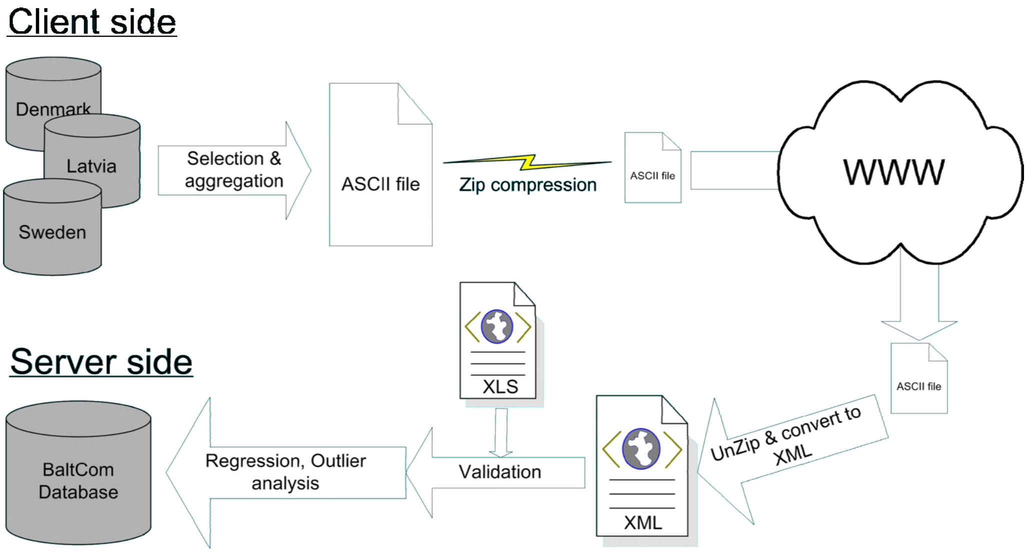


*Fig. 1. Dataflow from user to database.*

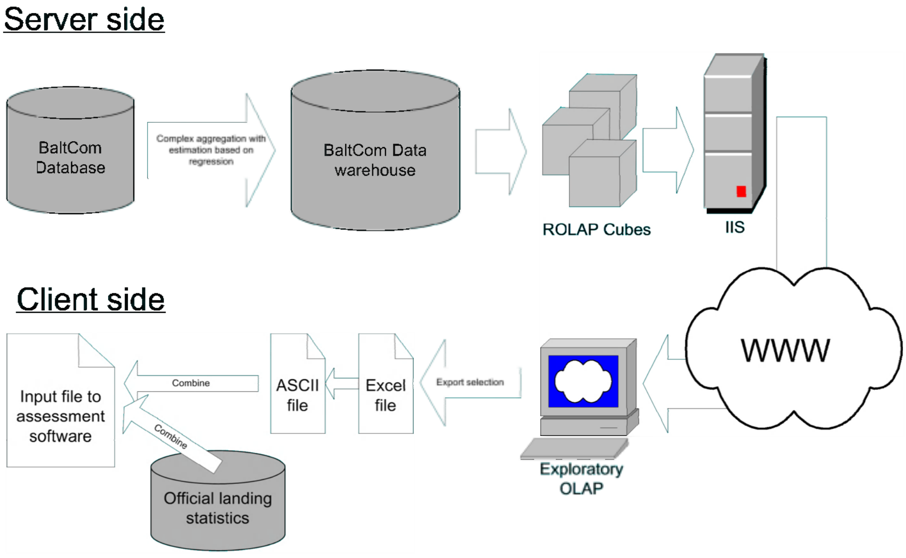


*Fig. 2. Dataflow from database to user.*

On Fig. 2 the dataflow from database to user is depicted. Calls from the web user interface to a VB.Net assembly (middle tier) updates the datawarehouse. The update procedures update the following tables in the datawarehouse: ALK (Age Length Key), SLD (Standard Length Distribution), CANUM (Catch at Age in Numbers) and MW (Mean Weight). This is done for all fishery sets [set of definitions of fisheries based on country, gear, target species, mesh size and sub division (ICES defined fishery area)]. This all happens in the data tier. The dataflow in the data tier is illustrated in Fig. 3. All the SQL calls make up a rather complex call stack hierarchy which is briefly described below. Full documentation can be found in Degel and Jansen (2003).

After ALK and SLD have been populated CANUM and MW are ready to be populated. They use age-length relations from ALK and weight-length relations are from SMALK (Size Maturity Age Length Keys). When such relations are missing for a given length, it is calculated using linear regressions. The weight-length regressions are being pulled out of the regression table, which is populated for all stratifications during the outlier analysis. The weight-length relation is modelled as Weight = Constant * $(\text{Length})^3$. The age-length regressions are being calculated runtime as linear regressions between data at ages at lengths shorter and longer. Degel and Jansen (2003) describe the set of rules defining the quality requirement of the regressions. Theoretical background for outlier analysis and regressions is found in (Sparre and Venema, 1998).

After populating the four main tables, a series of fishery set specific tables are populated with the relevant subset of data from the four main tables. These are to serve as fact tables for the OLAP (Online Analytical Processing) cubes.

A series of cubes were set up in Microsoft Analysis Server, one for each of the fact tables, and three for plain data OLAP (number of stations, length measures and otolith measures). The cubes were set up as ROLAP (Relational OLAP) real-time cubes.

The MS Pivot table and chart web components were selected as the user interface to the cubes. The component connects via msolap.asp on IIS (Internet Information Server) to msmdpump.dll. In the dll methods are being called to get data from the analysis server. Data is returned as XML back over the internet to the client. This happens without reloading the webpage on which the component resides. This OLAP cycle requires no custom development.
Exploratory OLAP can be performed at this stage, and data for preparing input to assessment models can be downloaded in the selected form by clicking the Excel icon.

## Results

Fig. 4 illustrates a pivot table with CANUM data. The user can change axes by dragging and dropping. Several dimensions can be at each axis at a time. Slicing, dicing, rolling up and drilling down are done by selecting values in the drop down lists. The pivot table is rapidly repopulated with data calculated for the new selection. On the figure the user has just clicked on ‘Size Category’ and can now select one of the values within that dimension. The same functionality is possible through the charts (see Fig. 5).

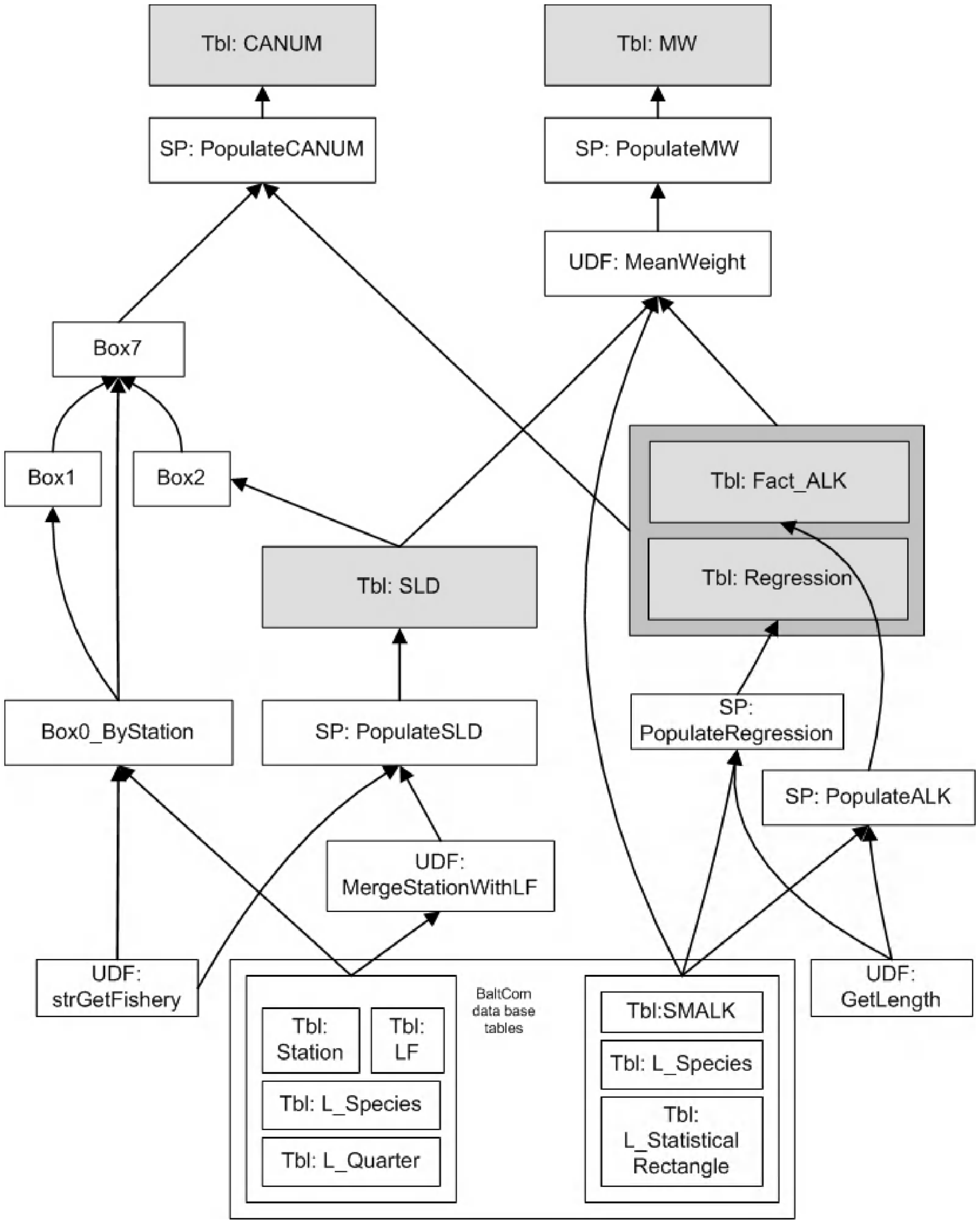
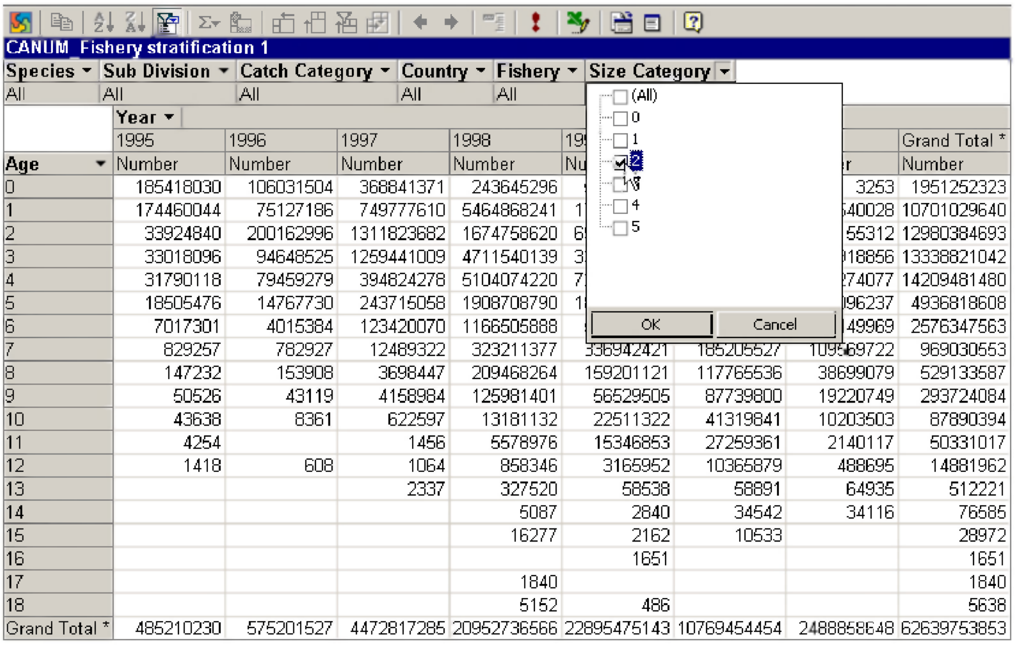


*Fig. 3. SQL call stack populating datawarehouse tables. The SQL in each unit is documented in Degel and Jansen (2003). Tbl (in white box) = Database table, Tbl (in grey box) = Datawarehouse table, UDF = User Defined Function, SP = Stored Procedure.*



CANUM_Fishery stratification 1

| Species | Sub Division | Catch Category | Country | Fishery | Size Category |
|---|---|---|---|---|---|
| All | All | All | All | All | |

□ (All)
□ 0
□ 1
☑ 2
□ 3
□ 4
□ 5

OK Cancel

| | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1995 | 1996 | 1997 | 1998 | 19 | | | Grand Total * |
| Age | Number | Number | Number | Number | Nu | | r | Number |
| 0 | 185418030 | 106031504 | 368841371 | 243645296 | | | 3253 | 1951252323 |
| 1 | 174460044 | 75127186 | 749777610 | 5464868241 | 1 | | 540028 | 10701029640 |
| 2 | 33924840 | 200162996 | 1311823682 | 1674758620 | 6 | | 55312 | 12980384693 |
| 3 | 33018096 | 94648525 | 1259441009 | 4711540139 | 3 | | 18856 | 13338821042 |
| 4 | 31790118 | 79459279 | 394824278 | 5104074220 | 7 | | 274077 | 14209481480 |
| 5 | 18505476 | 14767730 | 243715058 | 1908708790 | 1 | | 96237 | 4936818608 |
| 6 | 7017301 | 4015384 | 123420070 | 1166505888 | | | 149969 | 2576347563 |
| 7 | 829257 | 782927 | 12489322 | 323211377 | 336942421 | 185205527 | 109569722 | 969030553 |
| 8 | 147232 | 153908 | 3698447 | 209468264 | 159201121 | 117765536 | 38699079 | 529133587 |
| 9 | 50526 | 43119 | 4158984 | 125981401 | 56529505 | 87739800 | 19220749 | 293724084 |
| 10 | 43638 | 8361 | 622597 | 13181132 | 22511322 | 41319841 | 10203503 | 87890394 |
| 11 | 4254 | | 1456 | 5578976 | 15346853 | 27259361 | 2140117 | 50331017 |
| 12 | 1418 | 608 | 1064 | 858346 | 3165952 | 10365879 | 488695 | 14881962 |
| 13 | | | 2337 | 327520 | 58538 | 58891 | 64935 | 512221 |
| 14 | | | | 5087 | 2840 | 34542 | 34116 | 76585 |
| 15 | | | | 16277 | 2162 | 10533 | | 28972 |
| 16 | | | | | 1651 | | | 1651 |
| 17 | | | | 1840 | | | | 1840 |
| 18 | | | | 5152 | 486 | | | 5638 |
| Grand Total * | 485210230 | 575201527 | 4472817285 | 20952736566 | 22895475143 | 10769454454 | 2488858648 | 62639753853 |

*Fig. 4. Pivot table with CANUM data for fisheries stratification 1.*

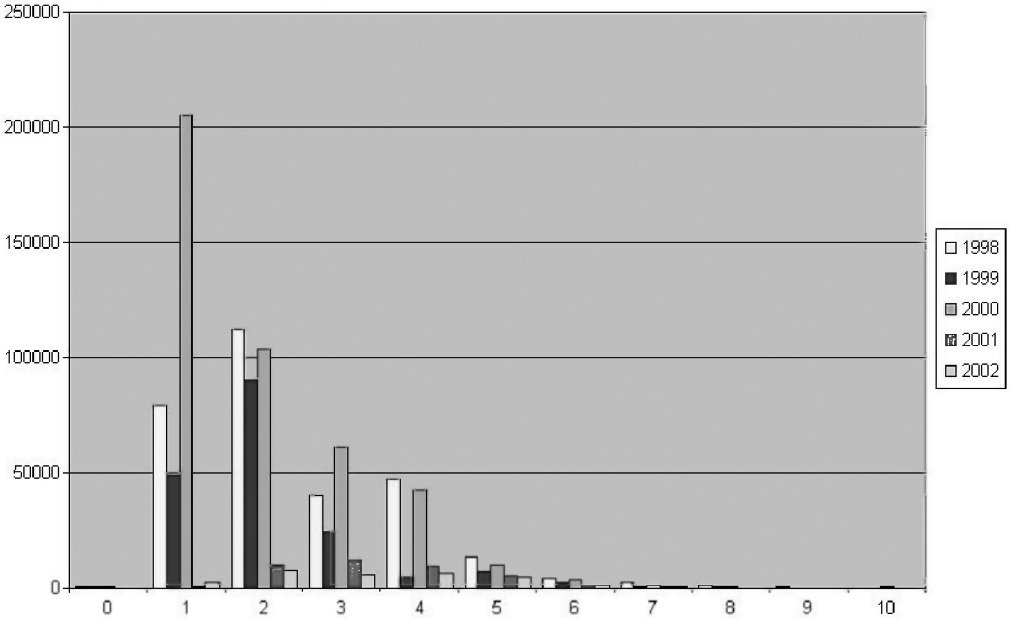


*Fig. 5. Chart with CANUM data for fisheries stratification 1. Age in years on X-axis and number of fish per landed ton.*

# Discussion

## *Selection of technology*

When selecting operating system, database server, software development language, development environment and 3$^{rd}$ part components several considerations have to be made in concern to the solutions scalability, performance, maintainability, licensing and cost of development.

The .Net platform was selected because it is based on the first and only standardized runtime environment, permitting language and platform independence; the CLI (Common Language Infrastructure). The technology was developed by Microsoft and submitted by several leading IT companies to ECMA and ISO for standardization. ECMA ratified it in December 2001 as ECMA standard 335 (ECMA, 2002), and ISO is about to publish it as ISO/IEC standard 23271 (HP, 2002).

The first open source implementation on LINUX and UNIX is on its way, Ximian (Ximian, 2002) has announced the launch of the Mono project (Mono, 2002), an effort to create an open source implementation of the .NET Framework. At this stage the ASP.Net module is still under development (Mono, 2002a). Other non-microsoft implementations SQL server 2000 was selected as the database server, because it is one of the best scaling and performing database server (TPC, 2002) and due to the ease of administration through the enterprise manager. When using SQL server the operating system is Windows 2000 server.

The rich query language T-SQL makes it possible to run complex procedural queries. This makes the data tier dynamic, well structured, transparent and the application as such performed faster than if data should get parsed back into the middle tier multiple times. However the data tier is therefore not a 100% pure data tier since some logic is implemented in this layer. Furthermore, the utilization of the proprietary T-SQL instead of ANSI SQL ties this part of the application specifically to SQL Server.
The internet information server IIS comes together with the Windows 2000 server.

Analysis server is packed together with the SQL server and is therefore an easy choice for OLAP server. However when internet support and features like real-time ROLAP and calculated cells are needed the SQL server needs to be the enterprise edition. This makes the solution rather expensive compared to a SQL server standard edition.

The SQL server is a very solid well performing piece of software. The ease of administration keeps the development and maintenance costs to a minimum. Only bad experience is the DTS (Data Transformation Services), which was used in BaltCom version 1 for the population of the datawarehouse. It is clearly a very early version that has been included in SQL server 2000, showing many signs of immature software like unhandled bugs and error messages without text.

The Analysis Server also gives an impression of not being matured. Numerous bugs and unspecified errors together with a slow administration environment raise the development costs and compromise the stability of the system. Although imperfect the administration/development environment is quite easy to use, and the most difficult part of creating the cubes is getting the syntax of MDX (Multi Dimensional Query language) right. Setting the storage mode of the cubes to ROLAP was the best solution. The only disadvantage compared to MOLAP and HOLAP is performance. The ROLAP cubes performed satisfactory, so the choice was made because ROLAP made real-time cubes possible. That way changes in the fact tables were reflected directly without processing the cubes. Only when adding new dimension values is it necessary to process the cubes. The process time is far shorter than when processing MOLAP (Multidimensional OLAP) and HOLAP (Hybrid OLAP), since it only recreates the dimensions.

The COM API (DSO) is well documented and easy to use, and was used to handle cube processing from the main application.
The client components are well performing giving the user a lot of features without any development. They however are ActiveX COM components, so they can mainly run on the windows platform. Pivot table services and chart services are a part of OWC (Microsoft Office Web Components), so OWC needs to be on the client machine. This can either be through an existing office installation or a manual component download and installation first time the website is visited. In either case a Microsoft Office license is needed.

## Conclusions and recommendations

The solution provided has covered all needs in terms of functionality. The application is performing satisfactory as well.

The developed three tier solution is smoothly scalable in terms of adding new functionality. The scalability in terms of how many users the system can handle before performance is compromised is yet to be tested.

SQL server, Visual Studio.Net and the .Net framework were successful choices. Development was very rapid, although this was the first .Net application made by the development team.
The way the .Net technology is moving these days makes it a very good choice for software development. Microsoft's .Net implementations are very comprehensive, matured and well tested. The Linux implementation by Mono is roughly one year away (Mono 2002c), and .Net applications will in theory run unmodified on Windows, Linux, HP-UX, Solaris, MacOS and FreeBSD (Mono 2002c). The COM inter-op-layer together with the Mono's CORBA inter-op-layer makes it very compliant to existing solutions whether developed in earlier Microsoft environments or in java.

The Analysis Server was satisfactory in terms of meeting the requirements of the solution, but the many unhandled bugs made the development somewhat tedious at times.

Once the client components were setup right they were very satisfactory in use, but many problems around the implementation made it a rather time consuming task. Even though some parts of the solution have now been moved to open standards, several parts are based on proprietary software tying the solution closely to Microsoft software. When developing international non-profit scientific software, one should aim at freeing the solution as much as possible from constraining licensing agreements. We recommend that the next version should if possible be based even more on open standards and open source.

The current trend of IT in the fields of biological sciences is that data are being gathered across national and organizational boundaries in larger datawarehouses or being integrated as distributed databases. In the view of the authors a prosperous next step would be the utilization of existing OLAP (Online Analytical Processing) and data mining software on top of these new structures. The present project has supported this. Such solutions would provide the fast facts and analysis required to take knowledge based decisions in a dynamic world.

## References

Degel H. and T. Jansen. 2003. BaltCom database. An internet based datawarehouse for Baltic Sea fisheries data. ICES Working Paper. Baltic Fisheries Assessment Working Group.

Dunham M. H. 2003. Data mining introductory and advances topics. Prentice Hall. 315p.

ECMA. 2002. Standard Information and Communication Systems. http://www.ecma.ch/ecma1/STAND/ecma-335.htm

Han, J. and M. Kamber. 2001. Data Mining Concepts and Techniques. Academic press. 550p.

HP. 2002. ECMA C# and Common Language Infrastructure standards. http://devresource.hp.com/specifications/ecma/index.html.

IBSSP I. International Baltic Sea Sampling Program for commercial fishing fleets I. EC Study Contract 96/002.

IBSSP II. International Baltic Sea Sampling Program for commercial fishing fleets II. EC Study Contract 98/024.

Mono. 2002. Mono Project. http://www.go-mono.com

Mono. 2002a. Mono Project. http://www.go-mono.com/asp-net.html

Mono 2002c. Mono Project. http://www.go-mono.com/faq.html

Sandbeck P., B. Cowan and T. Jansen. 2003. Use of XML technology in the Baltic Sea fisheries database. $$$

Sparre P. and S. Venema. 1998. Introduction to Tropical Fish Stock Assessment - Part 1: Manual. FOA Fisheries Technical Paper. 306/1 rev. 2. ISBN 92-5-103996-8. Online version at: http://www.fao.org/docrep/W5449E/w5449e00.htm

TPC. 2002. Transaction Processing Performance Council. http://www.tpc.org.

Ximian. 2002. Ximian. http://www.Ximian.com