Querying structured and unstructured data - combined approach

Asen Stefanov, Atanas Palazov

Institute of Oceanology, Bulgarian Academy of Sciences, Varna, Bulgaria

Objective: The term "unstructured data" mean different things in different contexts. In the context of relational database systems, it refers to data that can't be stored in rows and columns. Here, unstructured data are understood to include e-mail files, word-processing text documents, PowerPoint presentations, JPEG and GIF image files, and MPEG video files. In the past there were two options to store unstructured data file in the database: either to store the file in the database as a BLOB column, that was good for management but too bad for performance, or to keep the file in the file system and store a link in a database- good for performance, bad for managing the files.

Now a completely new architecture inside the present DBMS for handling file or unstructured data are available. The current DBMS offers many possibilities to store file content such as images, audio, video, PDFs, spreadsheets etc. side by side with other structured information.

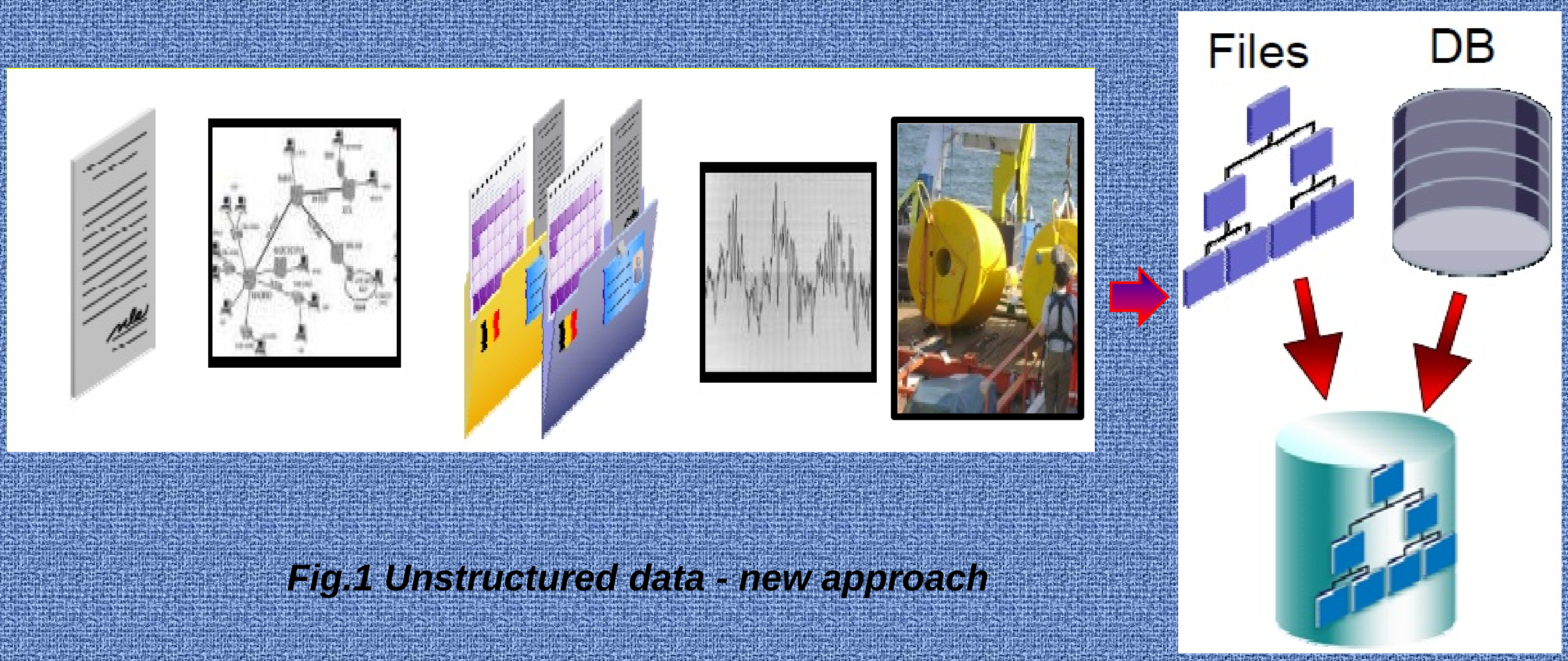
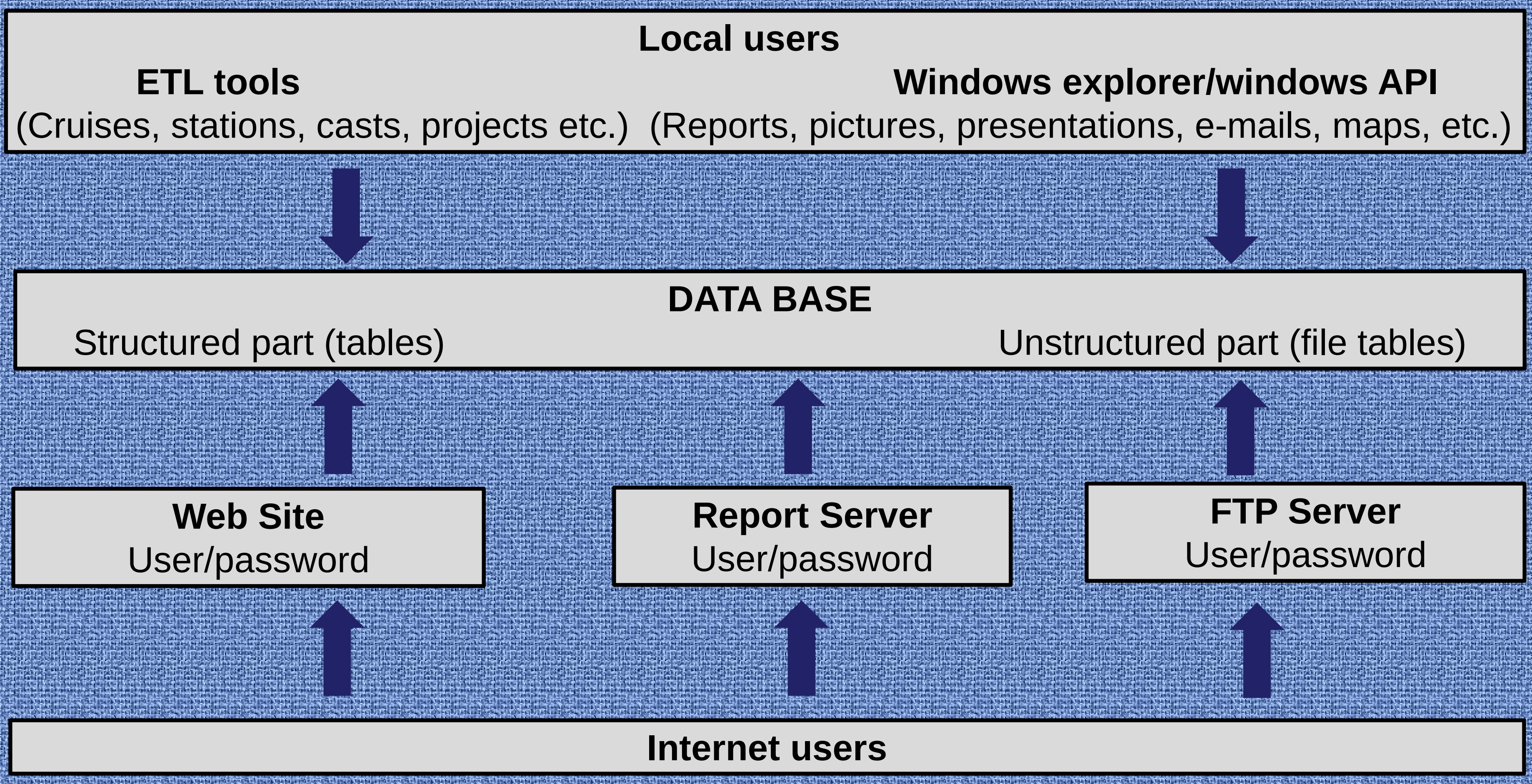


Fig.1 Unstructured data - new approach

Implementation: Today in the IO-BAS a large amount of oceanographic information is believed to be unstructured. As a result, the summary reports often lack key project data that might be relevant to decision making. "For a complete picture of the project, you have to look at both structured and unstructured data,"

IO-BAS has hands-on some experience in the still relatively new domain of unstructured data processing, and has helped the scientists to combine structured and unstructured solutions as follow:

- Initially the sources and file types of unstructured information are identified – ship's model characteristics, project documentation, contracts, cost sheets and etc.
- The various files and documents are bulk loaded, updated as well as managed in SQL in DBMS, and access is provided for standard applications as if they were stored in the file system. There are mostly text heavy, but can contain also scanned documents and images, presentations and etc.
- To access the unstructured part of data base the common network share for local users was implemented and an FTP portal was installed for Internet users.



Local users

ETL tools (Cruises, stations, casts, projects etc.)

Windows explorer/windows API (Reports, pictures, presentations, e-mails, maps, etc.)

DATA BASE

Structured part (tables)

Unstructured part (file tables)

Web Site User/password

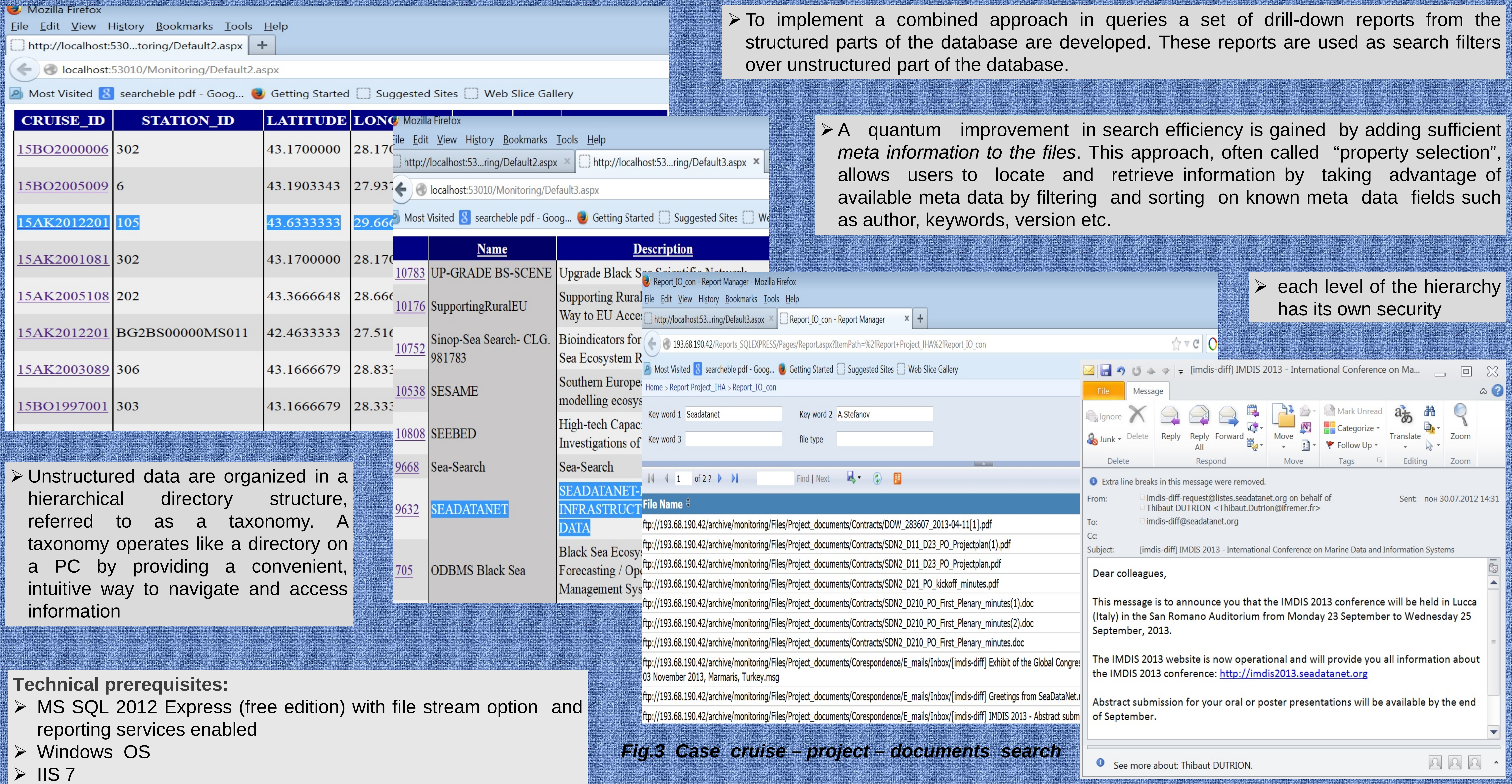
Report Server User/password

FTP Server User/password

Internet users

Fig.2 Data flow search - query- open

- To implement a combined approach in queries a set of drill-down reports from the structured parts of the database are developed. These reports are used as search filters over unstructured part of the database.
- A quantum improvement in search efficiency is gained by adding sufficient *meta information to the files*. This approach, often called "property selection", allows users to locate and retrieve information by taking advantage of available meta data by filtering and sorting on known meta data fields such as author, keywords, version etc.
- each level of the hierarchy has its own security
- Unstructured data are organized in a hierarchical directory structure, referred to as a taxonomy. A taxonomy operates like a directory on a PC by providing a convenient, intuitive way to navigate and access information


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:530...toring/Default2.aspx
localhost:53010/Monitoring/Default2.aspx
Most Visited
searcheble pdf - Goog...
Getting Started
Suggested Sites
Web Slice Gallery

CRUISE_ID	STATION_ID	LATITUDE	LONG
15BO2000006	302	43.1700000	28.170
15BO2005009	6	43.1903343	27.937
15AK2012201	105	43.6333333	29.666
15AK2001081	302	43.1700000	28.170
15AK2005108	202	43.3666648	28.666
15AK2012201	BG2BS00000MS011	42.4633333	27.516
15AK2003089	306	43.1666679	28.833
15BO1997001	303	43.1666679	28.333

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:53...ring/Default2.aspx
http://localhost:53...ring/Default3.aspx
localhost:53010/Monitoring/Default3.aspx
Most Visited
searcheble pdf - Goog...
Getting Started
Suggested Sites

	Name	Description
10783	UP-GRADE BS-SCENE	Upgrade Black Sea Scientific Network
10176	SupportingRuralEU	Supporting Rural Way to EU Acces
10752	Sinop-Sea Search- CLG. 981783	Bioindicators for Sea Ecosystem R
10538	SESAME	Southern Europe modelling ecosys
10808	SEEBED	High-tech Capaci Investigations of
9668	Sea-Search	Sea-Search
9632	SEADATANET	SEADATANET-I INFRASTRUCT DATA
705	ODBMS Black Sea	Black Sea Ecosy Forecasting / Ope Management Sys

Report_IO_con - Report Manager - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:53...ring/Default3.aspx
Report_IO_con - Report Manager
193.68.190.42/Reports_SQLEXPRESS/Pages/Report.aspx?ItemPath=%2fReport+Project_IHA%2fReport_IO_con
Most Visited
searcheble pdf - Goog...
Getting Started
Suggested Sites
Web Slice Gallery
Home > Report Project_IHA > Report_IO_con
1 of 2 ?
Find | Next

Key word 1: Seadatanet

Key word 2: A.Stefanov

Key word 3:

file type:

File Name
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/DOW_283607_2013-04-11[1].pdf
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D11_D23_PO_Projectplan(1).pdf
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D11_D23_PO_Projectplan.pdf
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D21_PO_kickoff_minutes.pdf
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D210_PO_First_Plenary_minutes(1).doc
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D210_PO_First_Plenary_minutes(2).doc
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Contracts/SDN2_D210_PO_First_Plenary_minutes.doc
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Corespondence/E_mails/Inbox/[imdis-diff] Exhibit of the Global Congres 03 November 2013, Marmaris, Turkey.msg
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Corespondence/E_mails/Inbox/[imdis-diff] Greetings from SeaDataNet.r
ftp://193.68.190.42/archive/monitoring/Files/Project_documents/Corespondence/E_mails/Inbox/[imdis-diff] IMDIS 2013 - Abstract subm

Extra line breaks in this message were removed.

From: imdis-diff-request@listes.seadatanet.org on behalf of Thibaut DUTRION <Thibaut.Dutrion@ifremer.fr>

Sent: пон 30.07.2012 14:31

To: imdis-diff@seadatanet.org

Cc:

Subject: [imdis-diff] IMDIS 2013 - International Conference on Marine Data and Information Systems

Dear colleagues,

This message is to announce you that the IMDIS 2013 conference will be held in Lucca (Italy) in the San Romano Auditorium from Monday 23 September to Wednesday 25 September, 2013.

The IMDIS 2013 website is now operational and will provide you all information about the IMDIS 2013 conference: http://imdis2013.seadatanet.org

Abstract submission for your oral or poster presentations will be available by the end of September.

See more about: Thibaut DUTRION.

Fig.3 Case cruise – project – documents search

Technical prerequisites:

- MS SQL 2012 Express (free edition) with file stream option and reporting services enabled
- Windows OS
- IIS 7

Conclusion:

It is necessary to read, integrate, and precondition the unstructured data before it can be used for the purposes of analytics in the structured environment. The benefits of combining the structured and unstructured data are:

- All information can be part of a database transaction, freeing the application from the complexity of guaranteeing atomicity, read consistency and other backup and recovery requirements
- The unstructured data can be analyzed
- The unstructured text can be accessed by direct or indirect searches
- The unstructured data can be linked to structured data and composite queries can be created
- At a certain point the saved in data base unstructured information can be transferred, for instance, into an established oceanographic relationship management system

IO-BAS, Bulgaria, Varna, First May Street 40, P.O.Box 152, Varna 9000, Bulgaria, http://io-bas.bg, e-mail: office@io-bas.bg