

Data Integration in the Life Sciences Workshops

DILS 2009, University of Manchester, United Kingdom

20th-22nd July

DILS09



e-DBI: e-science Database Integrator

A. Benabdelkader, V. Guevara

Science Park 107, 1098 XG,
Amsterdam, The Netherlands



vl·e

virtual laboratory for e-science

Presentation Outline

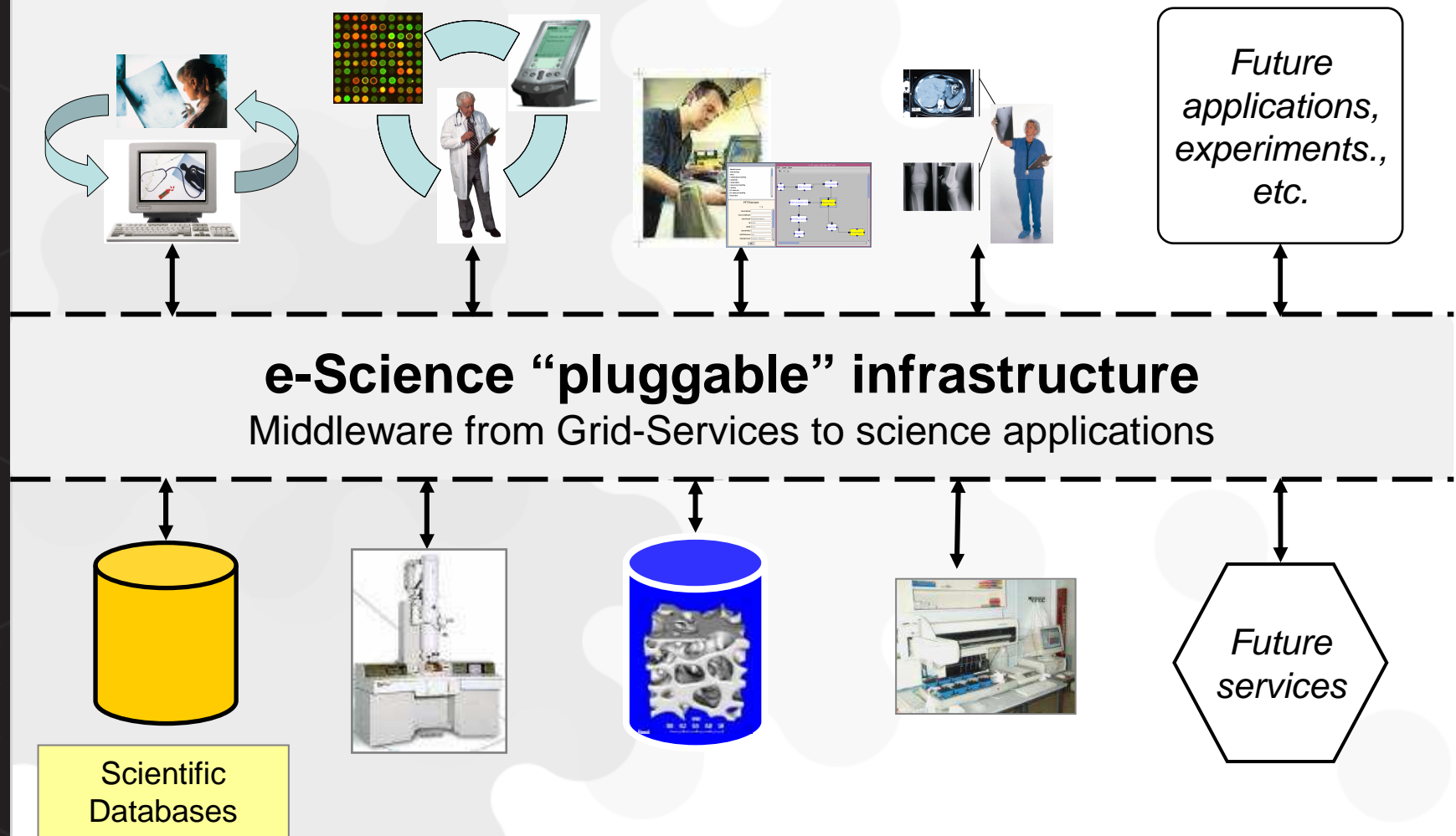
- Introduction
 - ∇ Scientific collaboration
 - ∇ Information management challenges
 - ∇ VL-e project
- Data management approach
 - ∇ Data Structure Generation
- e-science Database Integrator

e-Science Paradigm

*a new way of performing collaboration in scientific research
by sharing of **computing resources** and **information**
among a large number of scientists*

- **Large amounts of data** are generated by either *simulations* or *'networked' instruments* (i.e. instruments that are connected to storage and computing facilities through computer networks)
- Many steps in experiments are **automated** (e.g. re-plating biological sample by using a pipetting robot)
- **Information and communication technologies** (ICT) are extensively used throughout the entire experiment life-cycle, from experiment design and execution to results analysis and interpretation

e-Science framework



e-Science Challenges

Data size

- In biology, sequence databases double in every 14 months
- In physics, 100s of MB of data is generated by a single experiment



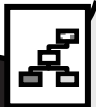
Security

- Access rights and visibility levels per experiment
- Robustness and data integrity



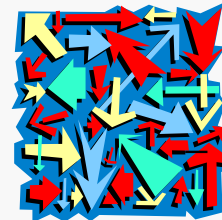
Data heterogeneity

- Wide variety of types of scientific information (diagnosis, readings, etc.)
- Various representations / formats (images, 3D reconstructions, etc.)
- Various access mechanisms



Complex environment

- Long and complex experimentation procedures
- People with different expertise



Lack of standards

- Different modeling and representation of information
- Specific solutions for some of the main problems
- Wasted efforts



Need for collaboration

- Sharing of resources (data, hardware, software, etc.)
- Collaborative work



VL-e project:

Virtual Laboratory for e-Science

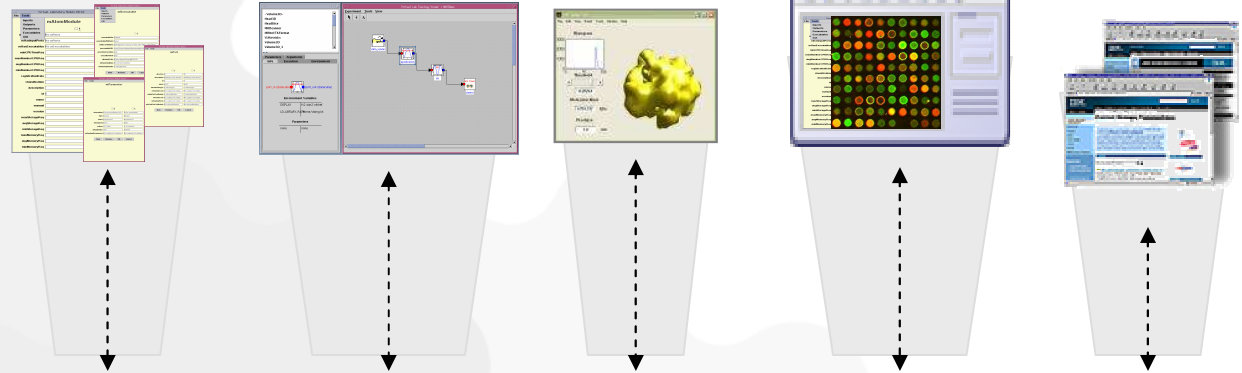
Multidisciplinary virtual laboratory environment for collaborative experimental science

(Dutch BSIK-OC&W / ICT-EZ project)

- Enable scientist to **define**, **execute**, and **monitor** their **collaborative experiments** by providing:
 - ∇ **location independent** experimentation
 - ∇ **familiar** experimentation environment
 - ∇ **assistance** during experimentation
- Designing, developing & integrating middleware to **bridge the gap** between the technology push of the high performance networking and the Grid, and the application pull of a wide range of scientific experimental applications
 - *High Energy Physics*
 - *Medical Imaging*
 - *Bio-Informatics*
 - *Food Informatics*
 - *Bio-Diversity*
 - *Dutch Tele-Science Laboratory*

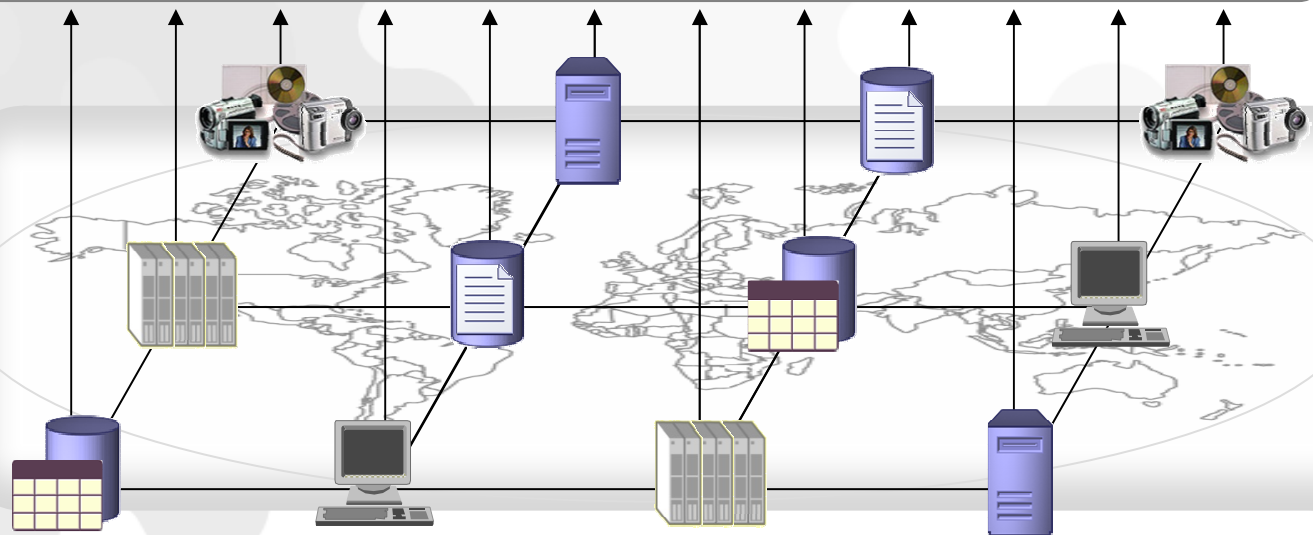


e-Science applications



VL-e middleware and generic facilities

Large-scale distributed systems



Scaling up & validation

Data Management Approach

*Provide a general framework for data management that support the management and the integration of data including **large data files**, **standard databases**, **ontologies**, and **data provenance**.*

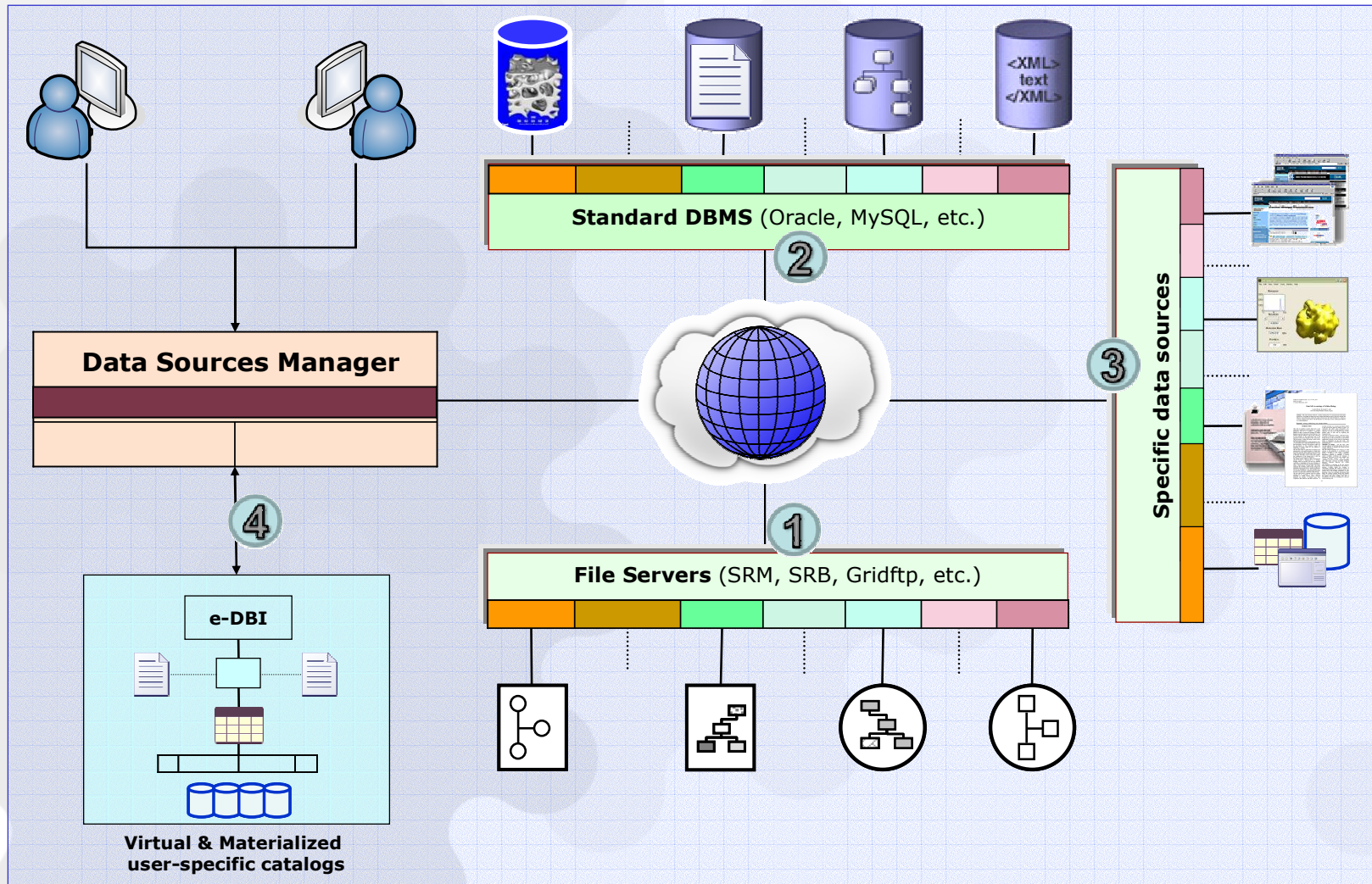
Functionality:

- To allow the storage and sharing of large data files
- To allow the annotation of scientific data with metadata and data provenance
- To allow the integration of data and metadata from different sources of information

Implementation:

- Follow a convenient implementation approach:
 - ∇ Make use of existing technologies (file servers, DBMS, XML, JDBC, etc.)
 - ∇ Enforce the use of open source and standard tools
 - ∇ Develop user-friendly interfaces
 - ∇ Hide system complexity (facilitating adoption)
 - ∇ Provide extensible and multi-platform solutions
 - ∇ Provide multi-environment solutions (desktop, server, grid-enabled, etc.)

Data Management: High-level architecture



Data Management:

Levels of integration

1st Level:

File Servers, consisting of secure online repository where scientific applications can store, organize, and share their data files

2nd Level:

Standard Databases, consisting of structured data and metadata. Metadata at this level mostly make references to external data files at the file servers

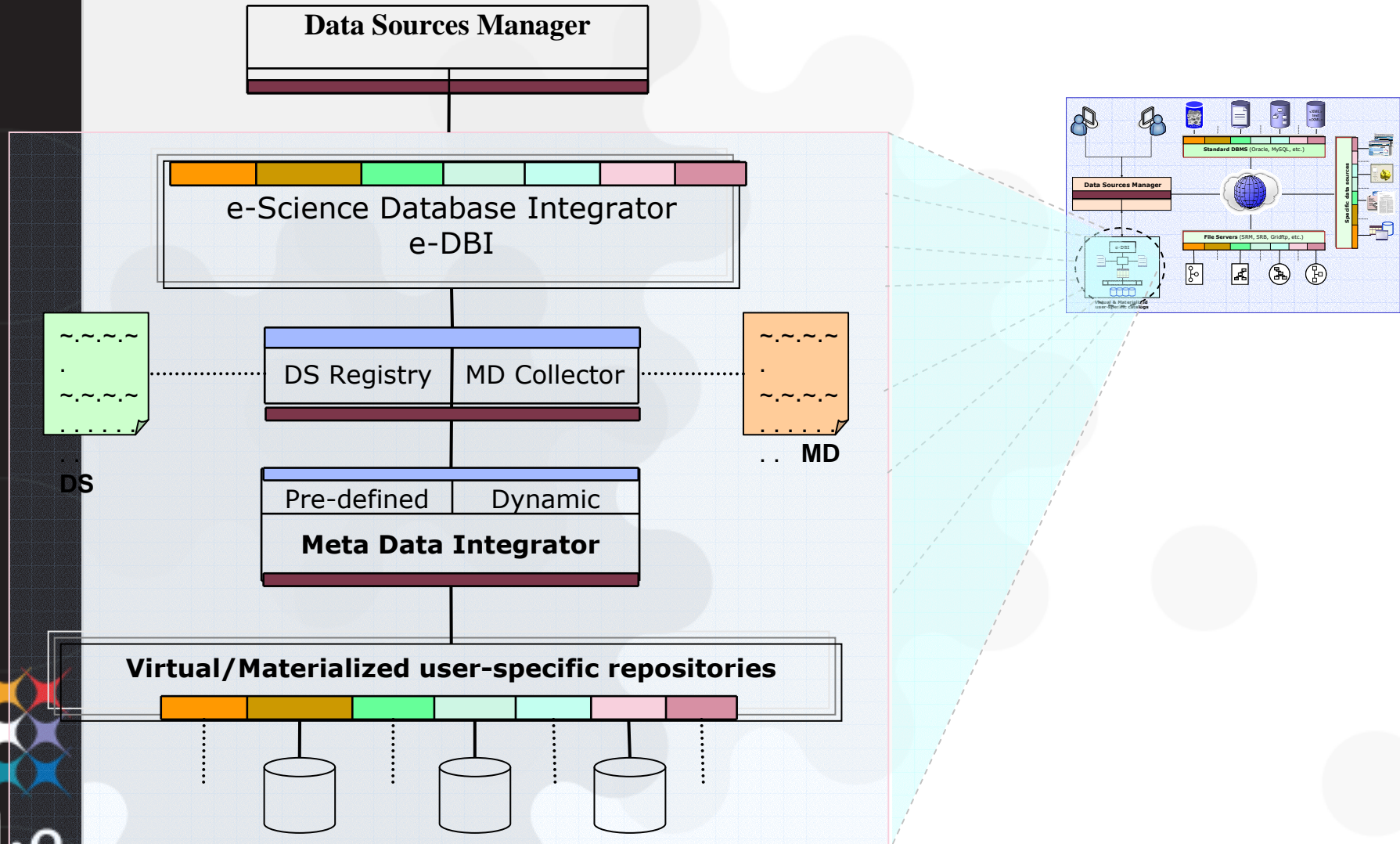
3rd Level:

Specific data sources, proprietary data format used by specific scientific applications. The support of this type of data is only provided if highly and strongly requested by the applications themselves.

4th Level:

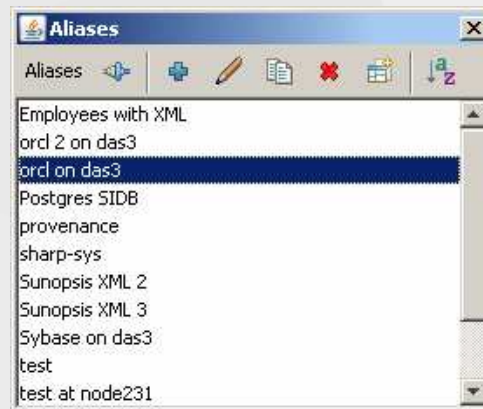
Data Integration Layer using the federated approach, with support of data warehousing, will be build based on the registered data sources and facilitated by the metadata information. In addition, knowledge integration and extraction tools could be also build at this level.

Data Integration Layer



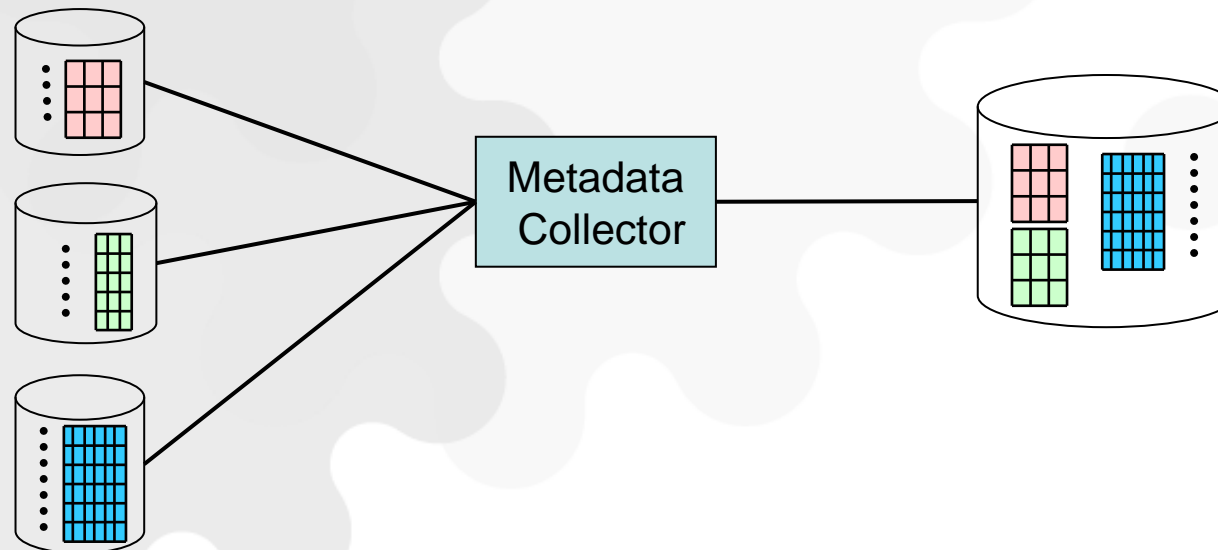
e-DBI – DS Registry

Description: e-DBI Data Source Registry allows the user from the application to **register the data sources** that will be used during the integration process. Information to be registered includes: DS name, host, port, driver, user name, and user password.



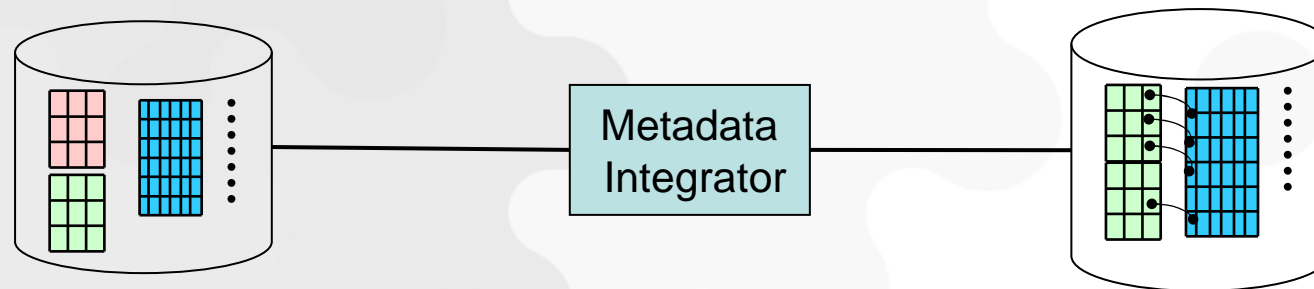
e-DBI – MD Collector

Description: e-DBI Meta Data Collector allows the user from the application to **identify the sub set of meta data** to be used for integration. In addition, MD Collector allows a limited meta data conversion to be applied against the single data sources, namely: renaming, conversion, aggregation, and type casting.



MD Integrator

Description: e-DBI Meta Data Integrator allows the user from the application to perform **MD integration** from the different data sources based on the set of metadata gathered through the MD collector. MD Integrator will allow a full integration of meta data from the different source, including data merging and data aggregation.

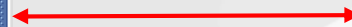
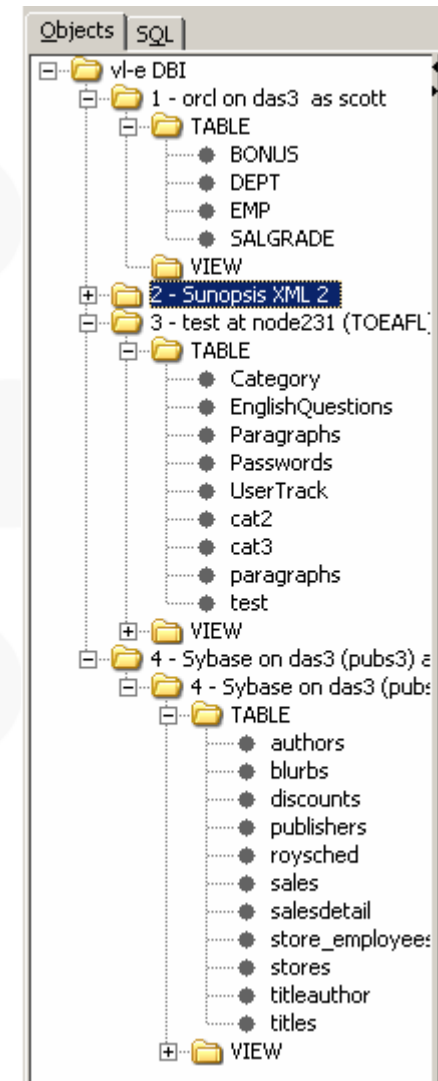
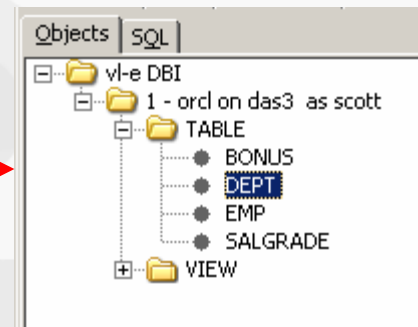
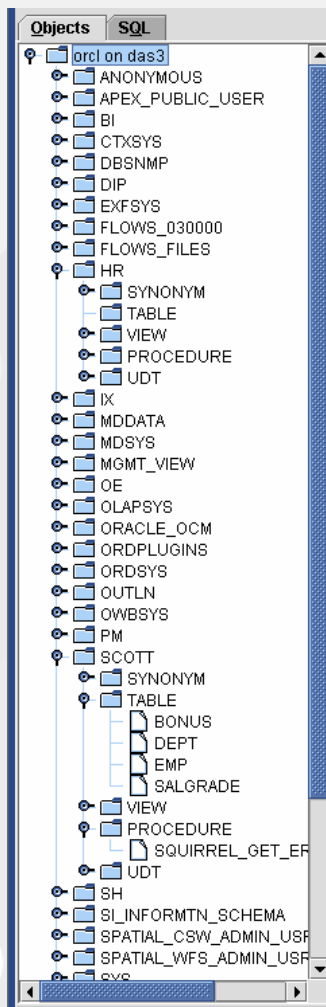


e-DBI – Principles

- **e-DBI build on top of Squirrel SQL**
 - ∇ Squirrel SQL provides seamless access to databases through JDBC
 - ∇ Squirrel SQL provides details information about the data sources
- **Focus on convenience and user-friendliness**
 - ∇ Make Squirrel SQL more convenient for data integration and for e-science.
 - ∇ Adaptation: arrangement to the interface
 - ∇ Simplification: hide unnecessary details from the scientist
- **Implementation of Data Integration Functionalities**
 - ∇ Allow the scientist to create a virtual database of his/her choice and to integrate data from multi-format data sources.
 - ∇ Scientist could filter the data
 - ∇ Scientist could reformat the data
 - ∇ Scientist could enhance the VDB structure
 - ∇ Scientist could refresh the VDB data

e-DBI vs. Squirrel SQL

User Convenience



User Interface Adaptation

e-DBI vs. Squirrel SQL

Simplification

Connection metadata simplification

String Functions	System Functions	Time/Date Functions	Keywords
Metadata	Status	Schemas	Table Types
Data Types	Numeric Functions		
ADD			
ALTER			
AUDIT			
CLUSTER			
COLUMN			
COMMENT			
COMPRESS			
CONNECT			
DATE			

Squirrel SQL

Info	Status	Metadata	Numeric Functions	String Functions
ABS				
AVG				
CEIL				
COS				
COSH				
COUNT				
EXP				
FLOOR				
GLB				

e-DBI

Table data/metadata simplification

Imported Keys	Indexes	Privileges	Column Privileges	Row IDs	Versions
Info	Content	Row Count	Columns	Primary Key	Exported Keys
	DEPTNO	DNAME	LOC	RO...	
10	ACCOUNTING	NEW YORK	AAA...		
20	RESEARCH	DALLAS	AAA...		
30	SALES	CHICAGO	AAA...		
40	OPERATIONS	BOSTON	AAA...		

Squirrel SQL

Content	Row Count	Columns	Primary Key
DEPTNO	DNAME	LOC	R...
10	ACCOUNTING	NEW YORK	AA...
20	RESEARCH	DALLAS	AA...
30	SALES	CHICAGO	AA...
40	OPERATIONS	BOSTON	AA...

e-DBI



e-DBI Interface

The screenshot displays the VL-e Database Integrator Version 2.6.6 interface. The window title is "VL-e Database Integrator Version 2.6.6". The menu bar includes "File", "Plugins", "Session", "Windows", and "Help". The "Connect to:" field shows "Employees with XML" and the "Active Session:" field shows "5 - Sybase on das3 (pubs3) as sa".

The main interface is divided into two panes. The left pane, titled "VL-e DBI", shows a tree view of database objects. The right pane, titled "SQL", shows a table view of the "authors" table.

The table view shows the following data:

au_id	au_name	au_fname	phone	address
409-56-7008	Bennet	Abraham	510 658-9932	6223 Bateman St.
213-46-8915	Green	Marjorie	510 986-7020	309 63rd St. #411
238-95-7766	Carson	Cheryl	510 548-7723	589 Darwin Ln.
998-72-3567	Ringer	Albert	801 826-0752	67 Seventh Av.
899-46-2035	Ringer	Anne	801 826-0752	67 Seventh Av.
722-51-5454	DeFrance	Michel	219 547-9982	3 Balding Pl.
807-91-6654	Panteley	Sylvia	301 946-8853	1956 Arlington Pl.
893-72-1158	McBadden	Heather	707 448-4982	301 Putnam
724-08-9931	Stringer	Dirk	510 843-2991	5420 Telegraph Av.
274-80-9391	Straight	Dick	510 834-2919	5420 College Av.
756-30-7391	Karsen	Livia	510 534-9219	5720 McAuley St.
724-80-9391	MacFeather	Stearns	510 354-7128	44 Upland Hts.
427-17-2319	Dull	Ann	415 836-7128	3410 Blonde St.
672-71-3249	Yokomoto	Akiko	415 935-4228	3 Silver Ct.
267-41-2394	O'Leary	Michael	408 286-2428	22 Cleveland Av. #14
472-27-2349	Gringlesby	Burt	707 938-6445	PO Box 792
527-72-3246	Greene	Morningstar	615 297-2723	22 Graybar House Rd.
172-32-1176	White	Johnson	408 496-7223	10932 Bigge Rd.
712-45-1867	del Castillo	Innes	615 996-8275	2286 Cram Pl. #86
846-92-7186	Hunter	Sheryl	415 836-7128	3410 Blonde St.
486-29-1786	Locksley	Chastity	415 585-4620	18 Broadway Av.
648-92-1872	Blotch-Halls	Reginald	503 745-6402	55 Hillsdale Bl.
341-22-1782	Smith	Meander	913 843-0462	10 Mississippi Dr.



Data Integration in the Life Sciences Workshops

DILS 2009, University of Manchester, United Kingdom

20th-22nd July



Thank you!



vl·e

virtual laboratory for e-science