e-DBI:
e-science Database Integrator

A. Benabdelkader, ***V. Guevara***
Science Park 107, 1098 XG,
Amsterdam, The Netherlands

BiG Grid
the dutch e-science grid

UvA Universiteit van Amsterdam

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

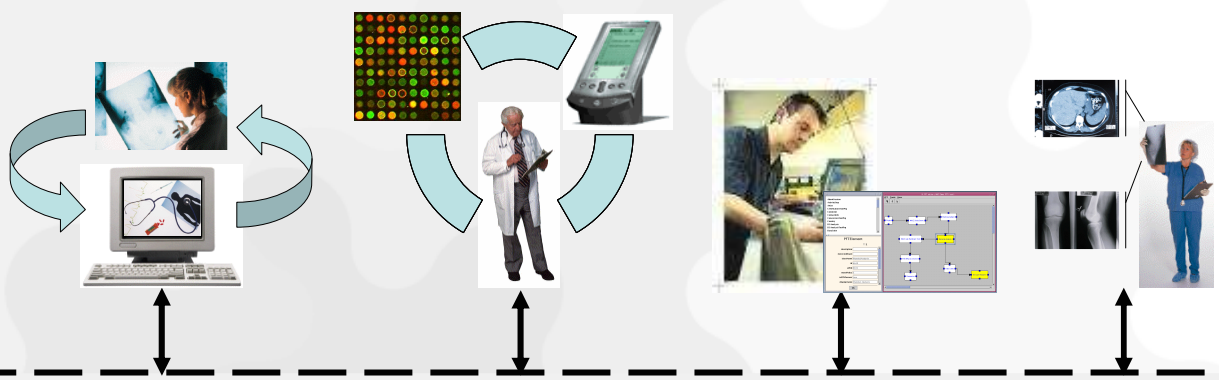
a new way of performing collaboration in scientific research

*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

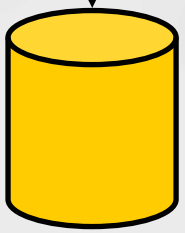
e-Science framework



Future applications, experiments., etc.

e-Science "pluggable" infrastructure

Middleware from Grid-Services to science applications



Scientific Databases

Future services

vl·e

e-Science Challenges

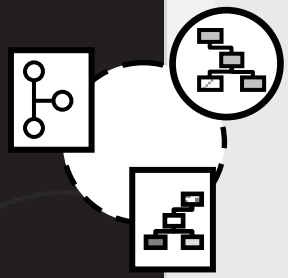
Data size

- 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

VL-e project: Virtual Laboratory for e-Science

Multidisciplinary virtual laboratory environment for collaborative experimental science

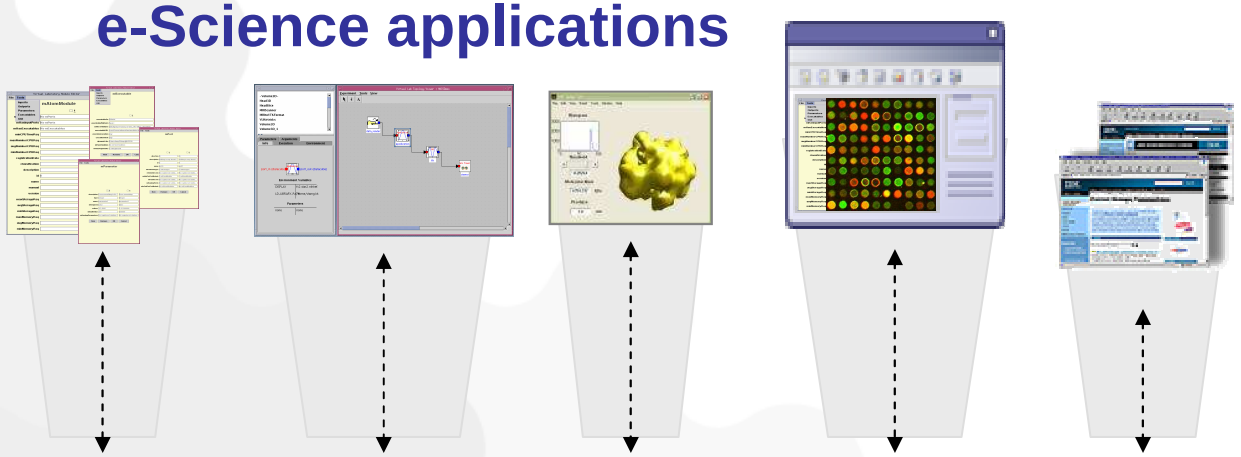
(Dutch BSIK-OC&W / ICT-EZ project)

- 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*

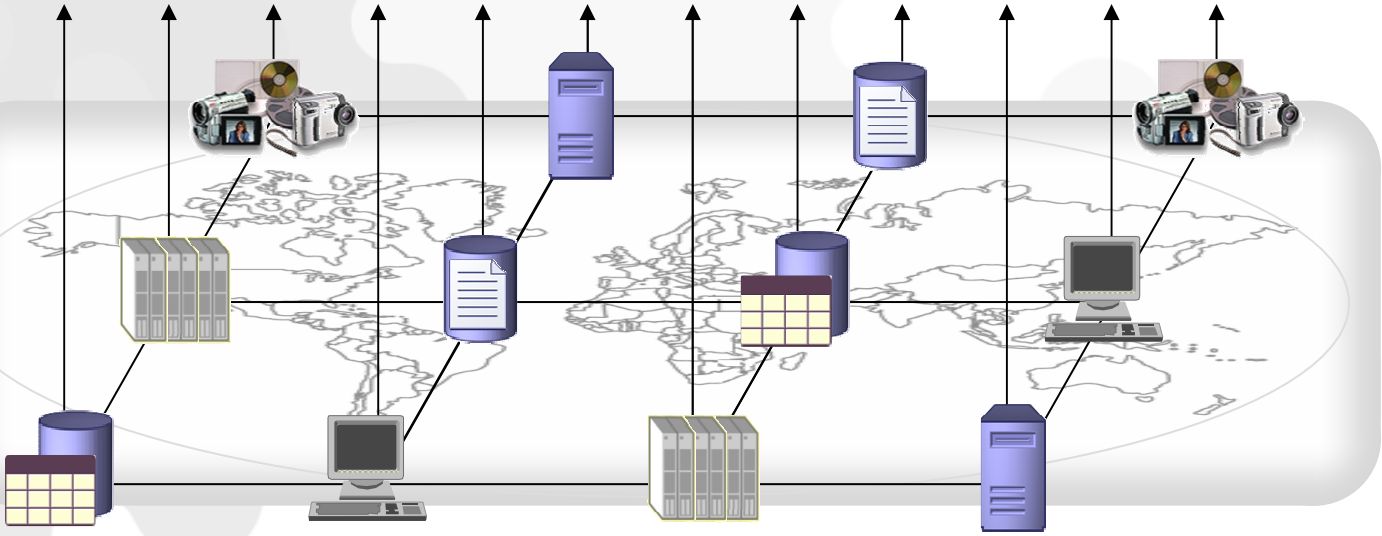
VL-e research areas

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

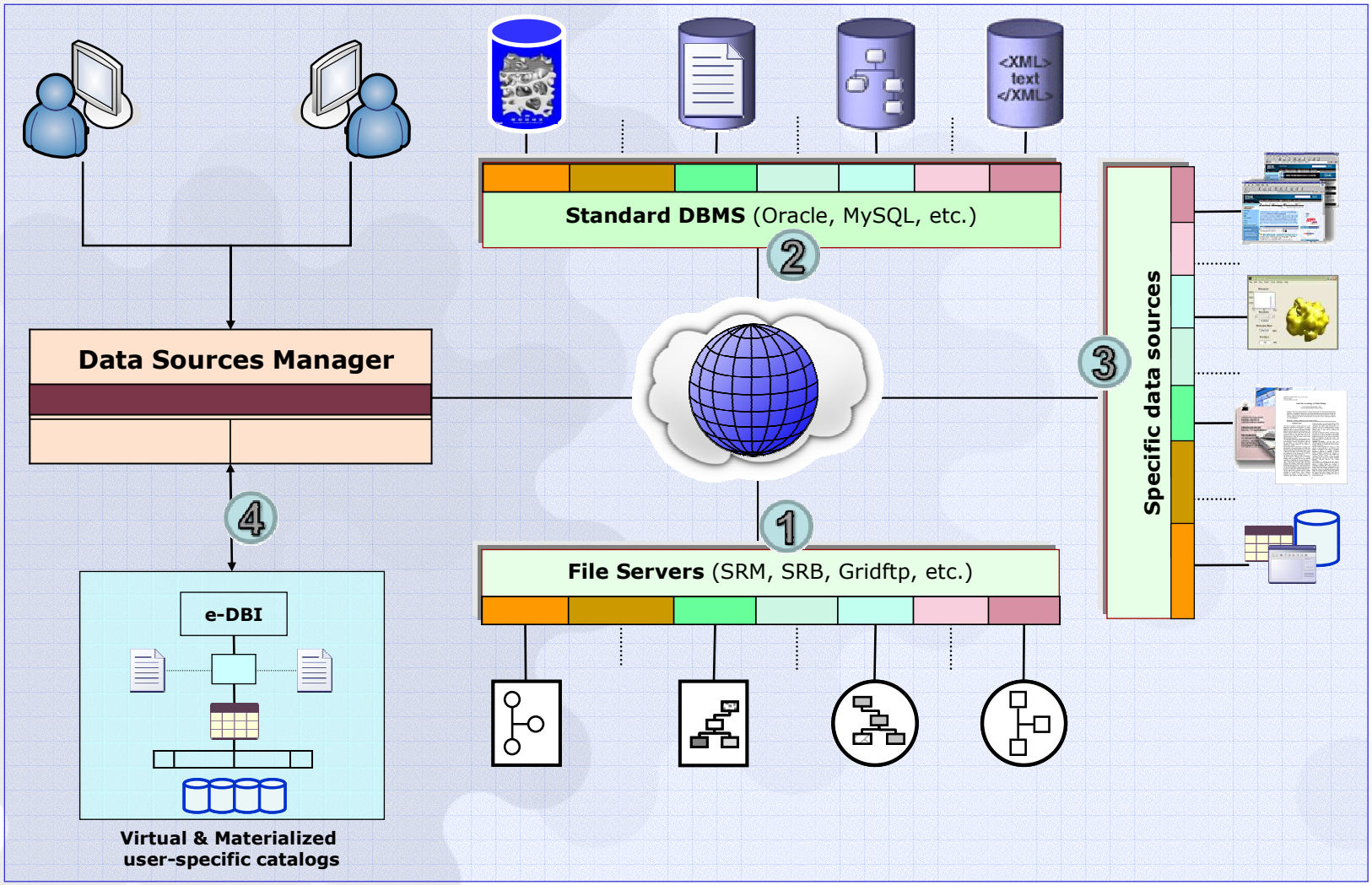
Implementation:

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

vl·e

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:

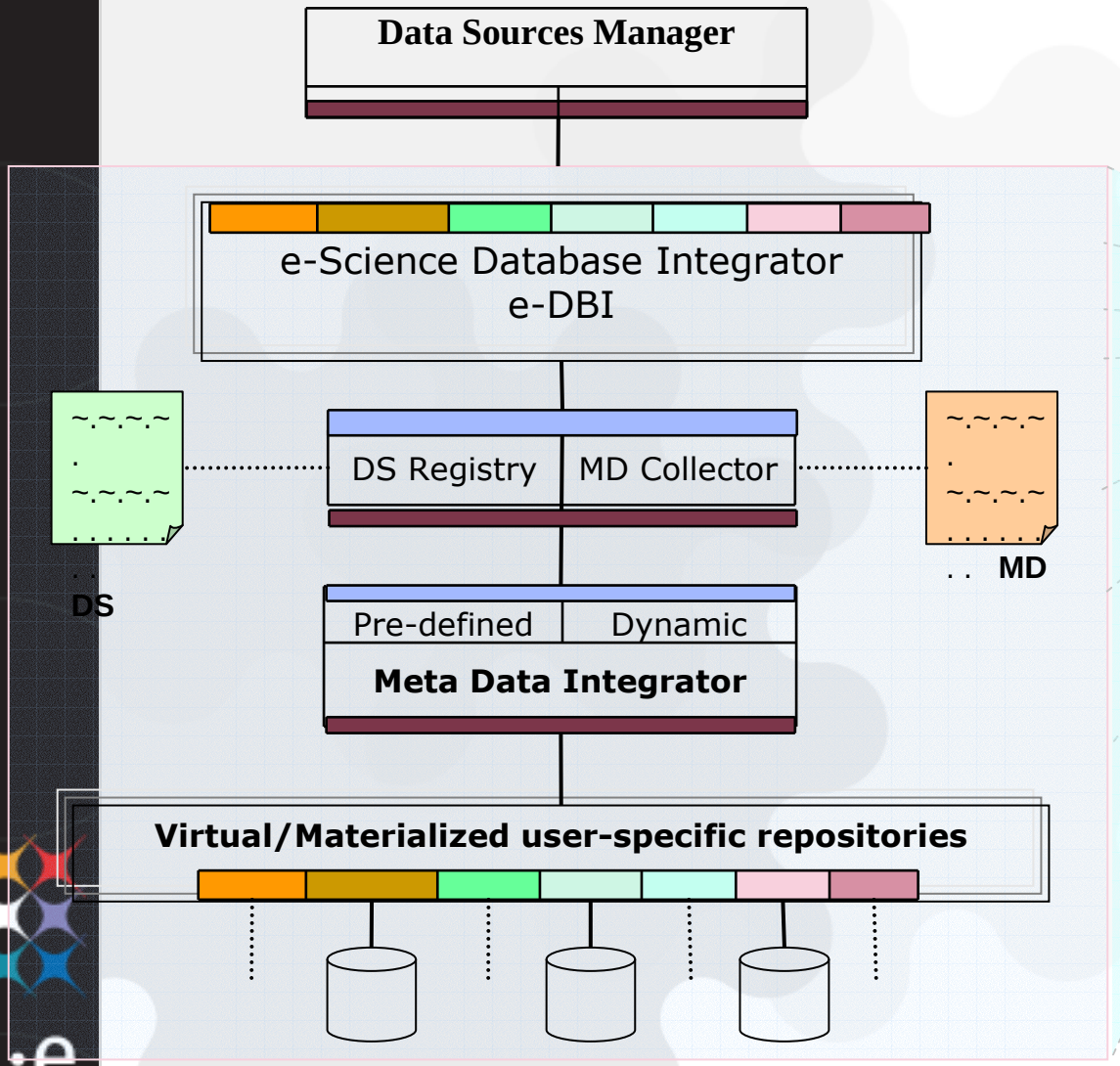
Specific data sources, proprietary data format used by specific scientific applications. The support of this type of data is only provided if highly and strongly requested by the applications themselves.

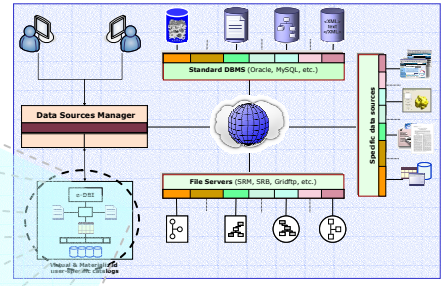
4th Level:

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.

vl·e

Data Integration Layer

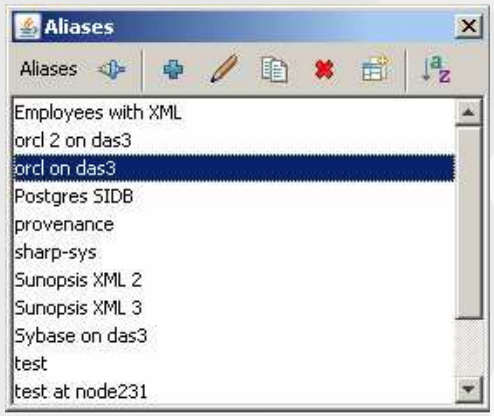


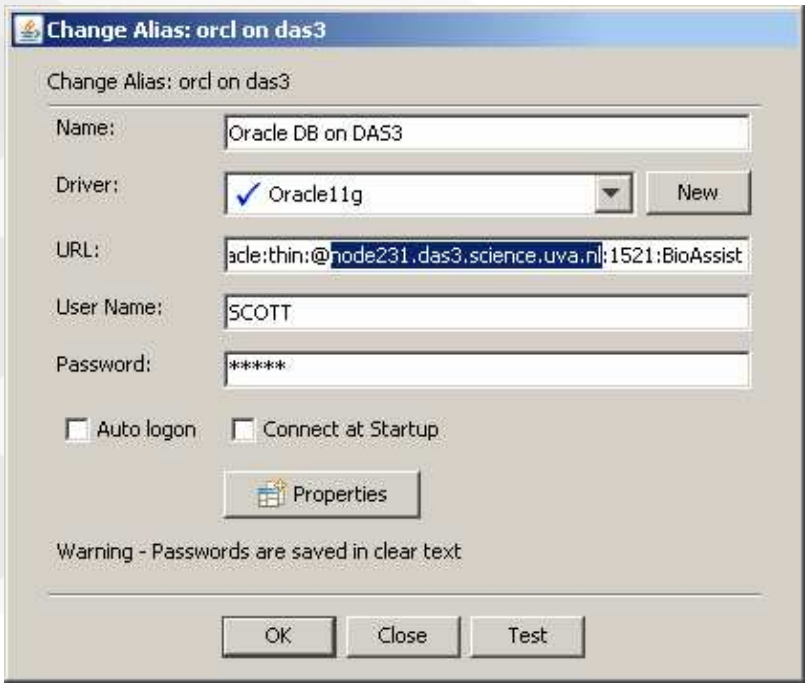


vl·e

e-DBI – DS Registry

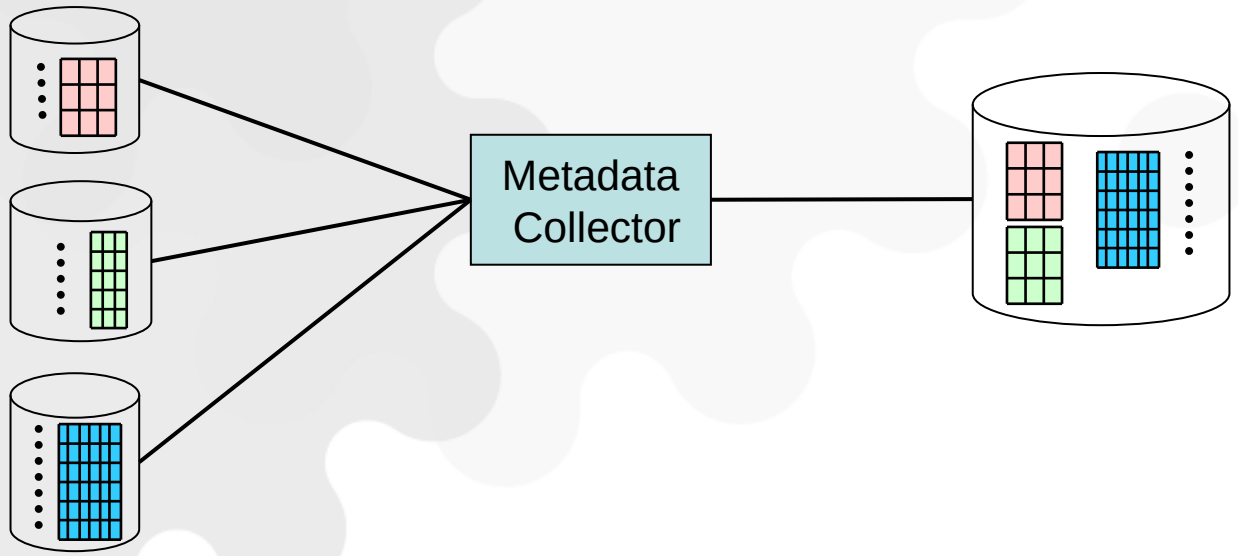
Description: e-DBI Data Source Registry allows the user from the application to **register the data sources** that will be used during the integration process. Information to be registered includes: DS name, host, port, driver, user name, and user password.





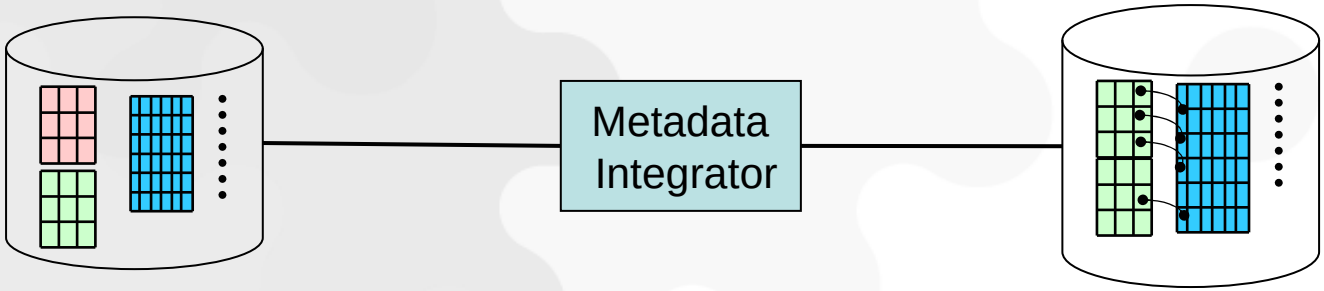
e-DBI – MD Collector

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.

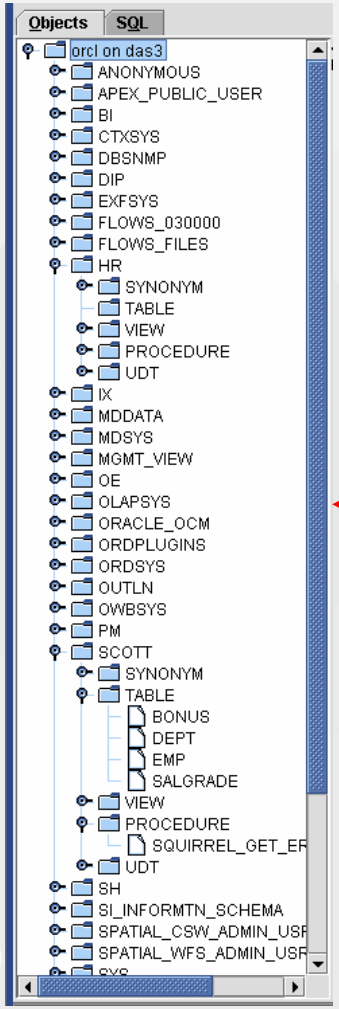


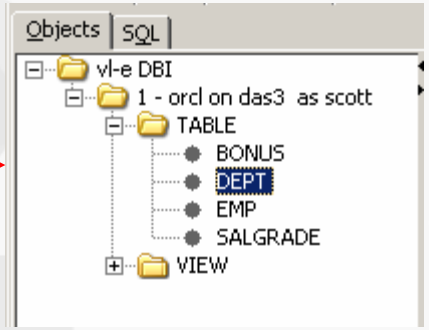
e-DBI – Principles

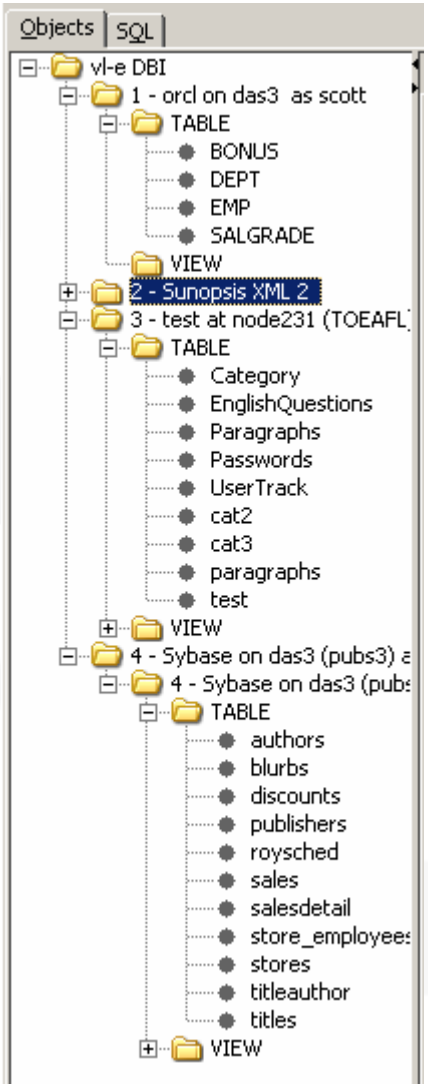
- **e-DBI build on top of Squirrel SQL**
  - Squirrel SQL provides seamless access to databases through JDBC
  - Squirrel SQL provides details information about the data sources
- **Focus on convenience and user-friendlyness**
  - Make Squirrel SQL more convenient for data integration and for e-science.
    - 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





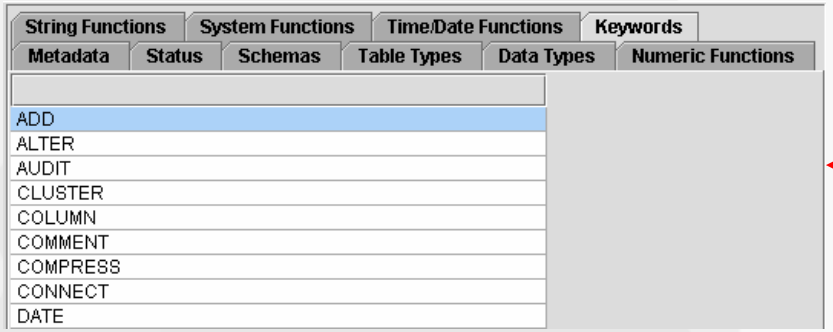


User Interface Adaptation

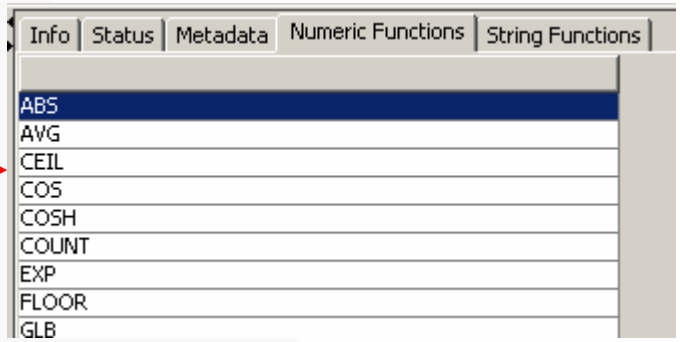
e-DBI vs. Squirrel SQL

Simplification

Connection metadata simplification

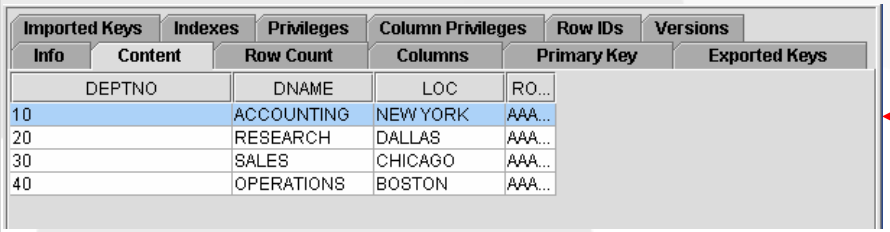


Squirrel SQL

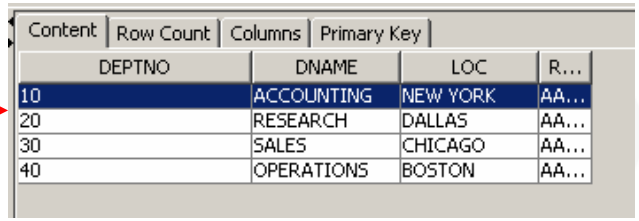


e-DBI

Table data/metadata simplification

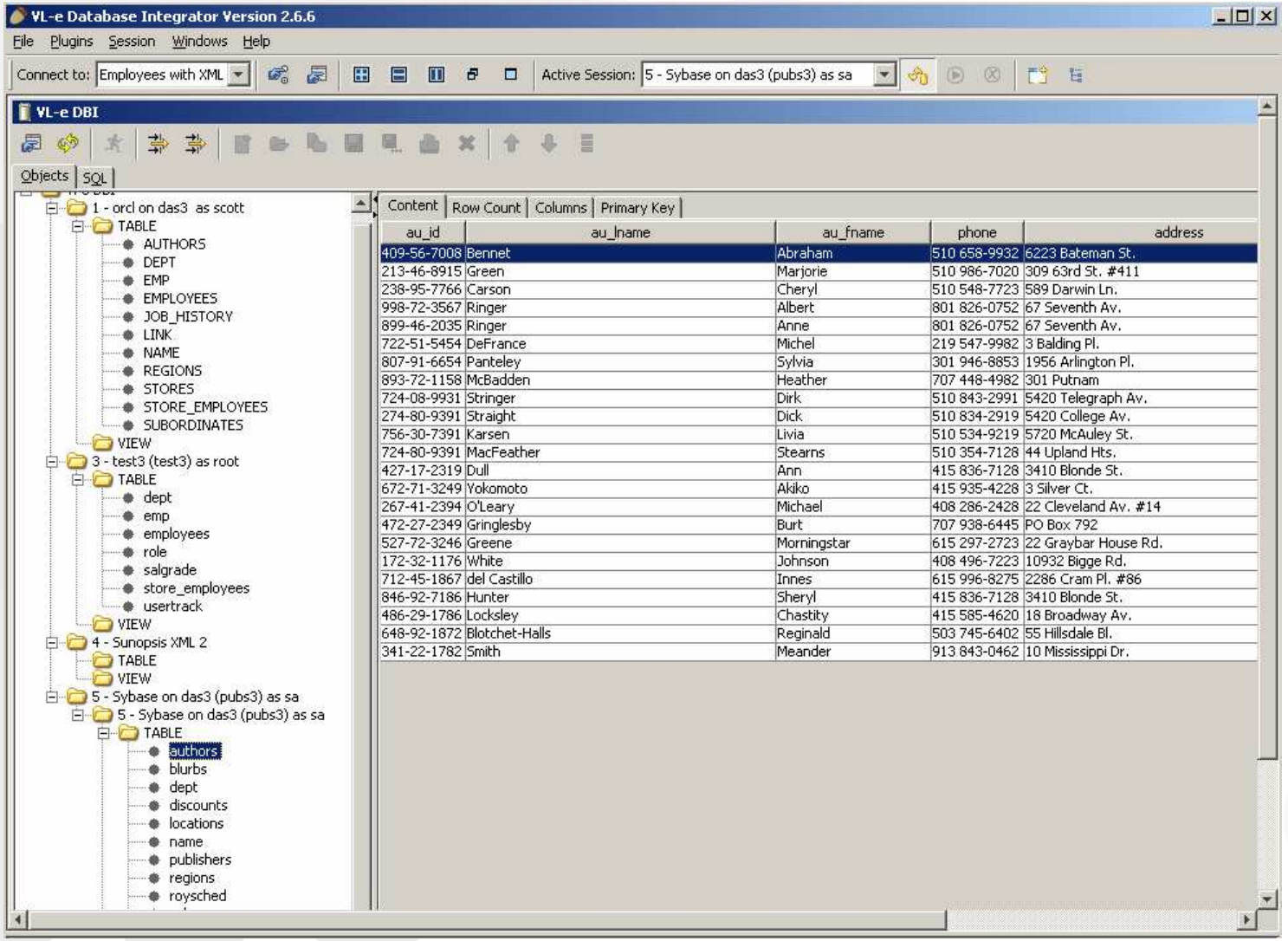


Squirrel SQL



e-DBI

e-DBI Interface




Data Integration in the Life Sciences Workshops
DILS 2009, University of Manchester, United Kingdom
20th-22nd July
DILS09

Thank you!


BiG Grid
the dutch e-science grid


UvA
Universiteit van Amsterdam


vl-e
virtual laboratory for e-science