

COLIM – Collaborative Information Management

VL-E / SP2.4

Ammar Benabdelkader

Hamideh Afsarmanesh

March 15, 2004



UNIVERSITEIT VAN AMSTERDAM



Presentation Outline

- Motivation
- What is COLIM?
- Focus Areas of COLIM
- Components of COLIM
- Development Environment
- Conclusions

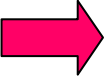




Motivation

Multi-disciplinary scientific applications of VL-E are challenged with proper management of their information

- Wide range and variety of information (bioinformatics, biodiversity, food informatics, medical diagnosis, Telescience, etc.)
- Access autonomy and security for shared data, fine-grained visibility rights
- Lack of standards for modeling, access, and retrieval
- Require assisting tools to support scientists



An emerging need to exchange and integrate information to support collaborative (multi-disciplinary) problem solving environments





What is COLIM?

COLIM - Collaborative Information Management

- A component of the VL-E middleware
- Aims to support modeling, management, and integration of information (bioinformatics, biodiversity, food informatics, medical diagnosis, and Telescience)
- Grid-based tools/mechanisms for information retrieval
- Support information access/integration for collaborative problem solving
- Assisting tools to customize the modeling and generation of data structures and schemas, from formal (ontological) definitions





COLIM – Focus Areas

Five areas of focus define the research and development in COLIM

Area 1

Analysis, design, and development of databases for scientific applications in VL-E

Area 2

Assisting scientists with dynamic ontology-based generation of their data structures and database schemas

Area 3

Analysis of Grid potential for remote information access and query processing in VL-E

Area 4

Development of mechanisms / tools for sharing and exchange of heterogeneous information from autonomous sites

Area 5

Development of the information management system, integrated with other subprograms in VL-E



Based on the migration of VL/VIMCO results to VL-E/COLIM



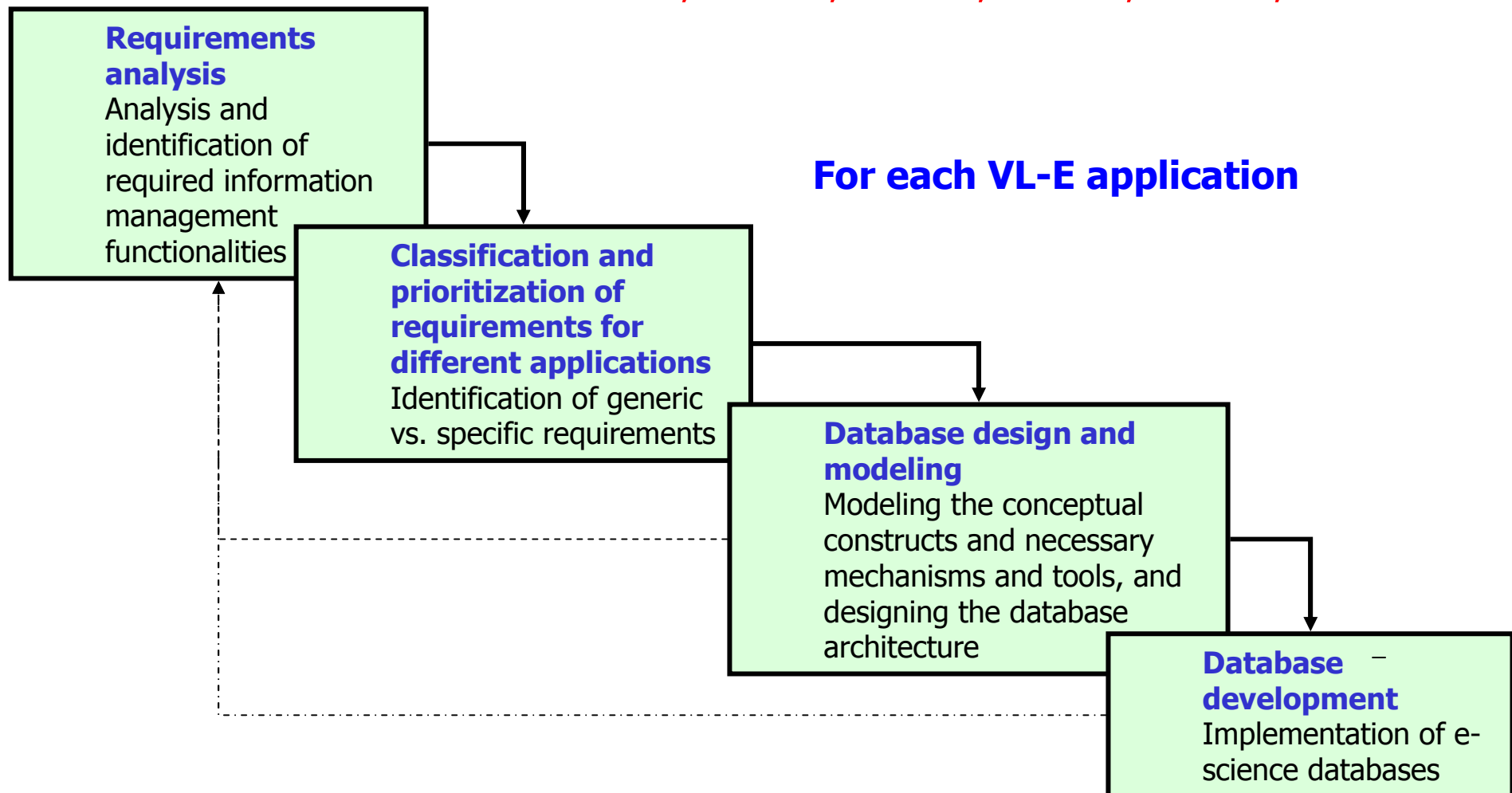
UNIVERSITEIT VAN AMSTERDAM

Area 1 –

Building VL-E Databases for Scientific Applications

Study of the VL-E applications and building their corresponding databases

in collaboration with **SP1.1, SP1.2, SP1.3, SP1.4, SP1.5, SP1.6**





Area 2 – Assisting Scientists with Dynamic/Automatic Ontology-based Generation of Database Structures

Design and development of tools to assist scientists with their dynamic/automatic ontology-based generation of data structures and schemas for their information

- Based on knowledge developed on similar tools within the EC/5FP TeleCARE project
- Definition of concepts and entities through a user-friendly ontology-based interface (e.g. based on Protégé)
- Automatic generation of required structures
- Co-work with **SP2.2** – Adaptive Information Disclosure on ontology



Area 3 – Grid Potential analysis for Information access and Query Processing

Analysis of the latest developments in Grid, to be used as the base communication infrastructure for query processing

- The Globus 'bag of services' only supported the very basic data access/retrieval mechanisms
- COLIM aims at using recent developments in the Grid for remote data access and for implementing federated information integration and query processing
- Benefits foreseen include data transfer security and optimization of query processing time



Area 4 –

Information Exchange Among Autonomous Sites

Necessary services/tools for integration and exchange of heterogeneous information from distributed autonomous sites

- Access and integration of data from both public and private organizations
- Development of adapters for some legacy systems
- Information exchange using standard formats
- Analysis of applicability/usability of existing software (e.g. IBM DB2 II) as base software for federation of databases



Area 5 – Integration of COLIM with other SPs in VL-E Middleware

- **Migration** of results generated in the previous VL project to VL-E
 - from VIMCO to COLIM, re-using VIMCO components to the maximum
- **Integration** of COLIM developments with other SPs in the VL-E middleware
 - with **SP2.1** Co-work on interactive problem solving environments
 - with **SP2.2** Co-work on ontology definition for automatic data structure definitions, for e-science applications
 - with **SP2.3** possible co-work on user interfaces for information management
 - with **SP2.5** Co-work on building the database for the core engine of the VL-E middleware



COLIM – Components

COLIM consists of five main components

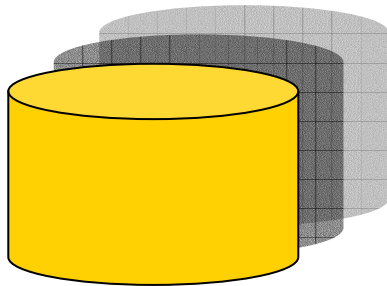
- **VL-E engine database:** Database for the core engine of VL-E middleware
- **Application databases:** Application-specific experiment-related databases
- **Base information management system (COLIM Server):** The system supporting management and manipulation of information stored in the VL-E engine database and the application databases
- **Federated information integrator and query processor:** Information integrator and query processor, using Grid as the base communication infrastructure
- **Assisting tools for scientists:** Dynamic data structure generation tool (e.g. using ontological definitions)



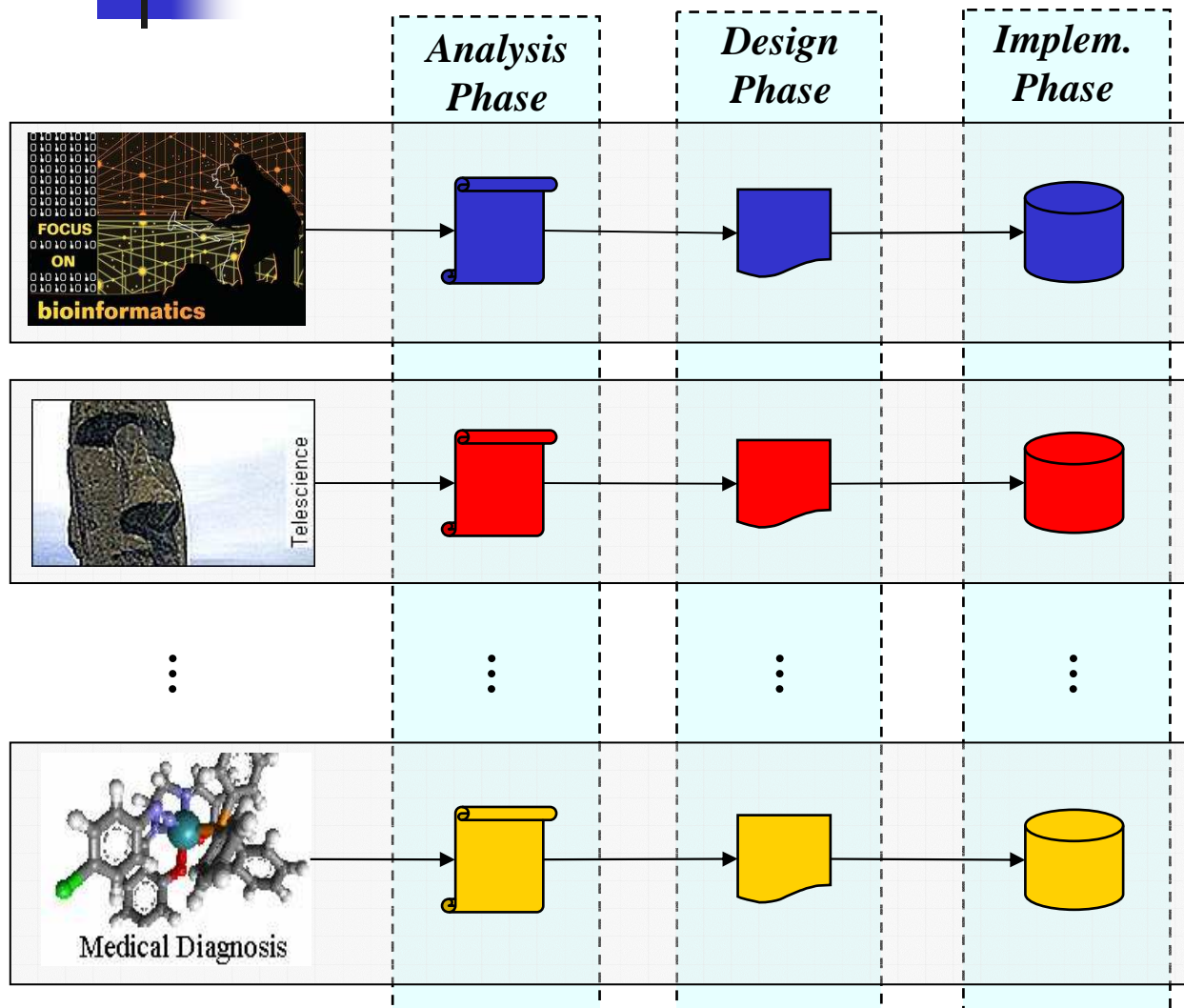
VL-E Engine Database

For the core VL-E middleware

- User information
- Roles and access rights definitions
- Active sessions
- Definition of available data sources
- Definition of available software entities
- ...



Application Databases



Databases containing application-specific experiment-related information

- Project definitions
- Experiment definitions
- Application specific- scientific information
- ...



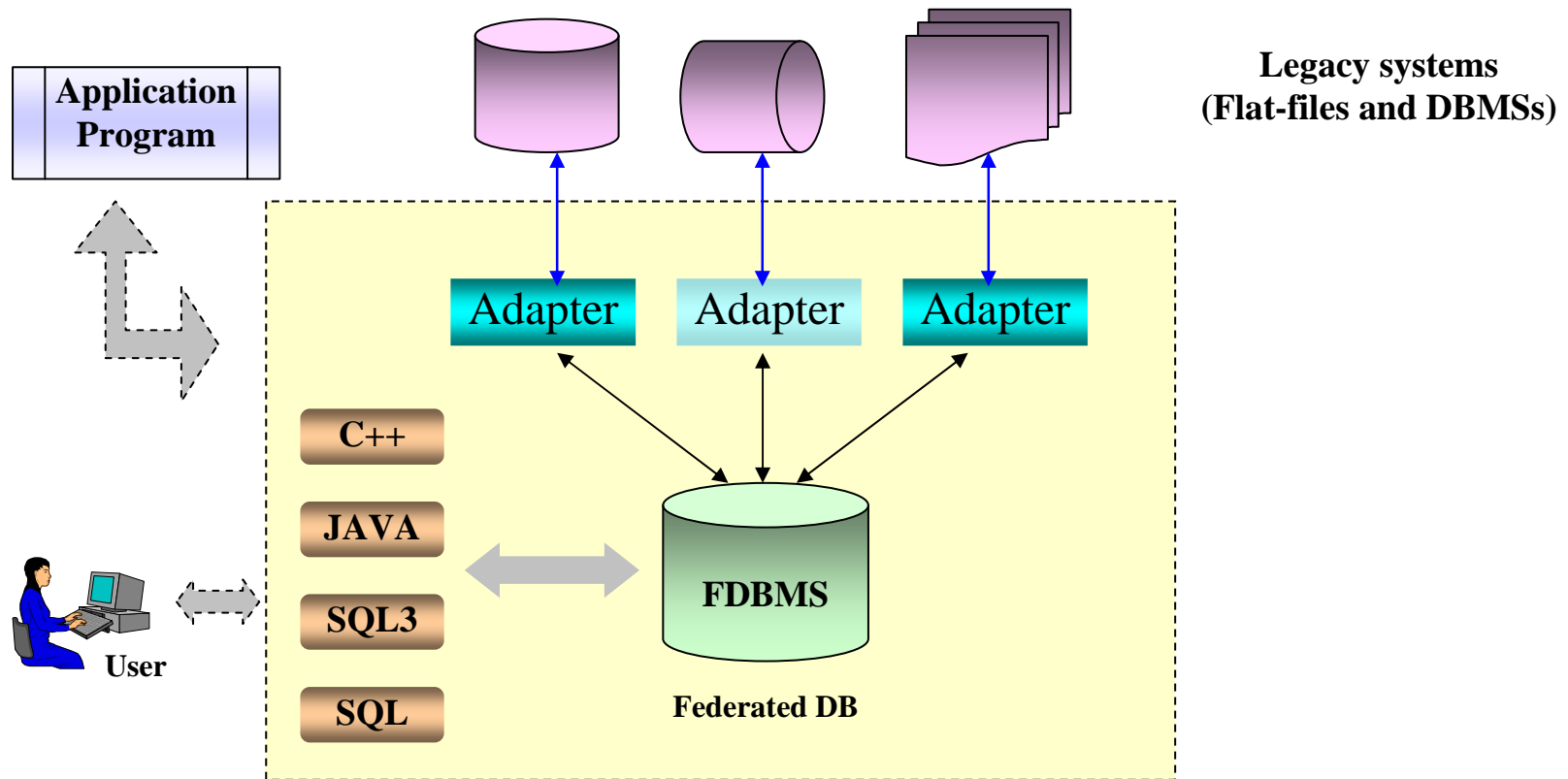
COLIM Server: Base Information Management System

The software libraries of COLIM system

- Managing and manipulation of various information
 - Application-specific experiment-related information
 - Users, roles, access rights, ...
- Common information management services shared by all scientific applications in VL-E
- Other information management libraries, e.g. from the Grid environment



Federated Information Integrator and Query Processor



Using Grid as the base communication infrastructure



Assisting Tools for Scientists

One tool is planned to support scientists with highly flexible mechanisms for automatic data structure generation based on ontological entries

- User-friendly graphical interfaces for definition of application specific data in a more user-familiar environment
- Assisting tools to automatically translate the ontology-based data definitions into data structure and database schemas





Development Environment: Initial Ideas

Development of COLIM is planned in different phases and durations

- At different stages for each component of COLIM, a demonstrator will be generated to validate/verify the achieved results
- Following development environment is planned:
 - Using Matisse 7 as the main DBMS to model, manage, and manipulate scientific data
 - The use of Oracle (or a similar relational DBMS) will be investigated
 - Using the DAS2 Cluster to run all databases and the COLIM Server
 - Using Grid technology and as much as applicable following the standards for the development of all components in COLIM





Summary

- Collaborate with other SPs to develop databases for different VL-E applications:
 - Bioinformatics-ASP, Medical Diagnosis & Imaging, Biodiversity, Dutch Telescience, and food informatics
- Design and development of VL-E engine database and COLIM Server system
- Development of assisting tools for dynamic ontology-based generation of database schemas
- To support collaborative applications, design and development of federated query processing using Grid as the base communication infrastructure
- Phase-based demonstration scenarios for validation/verification of the design and development of the COLIM components and software

