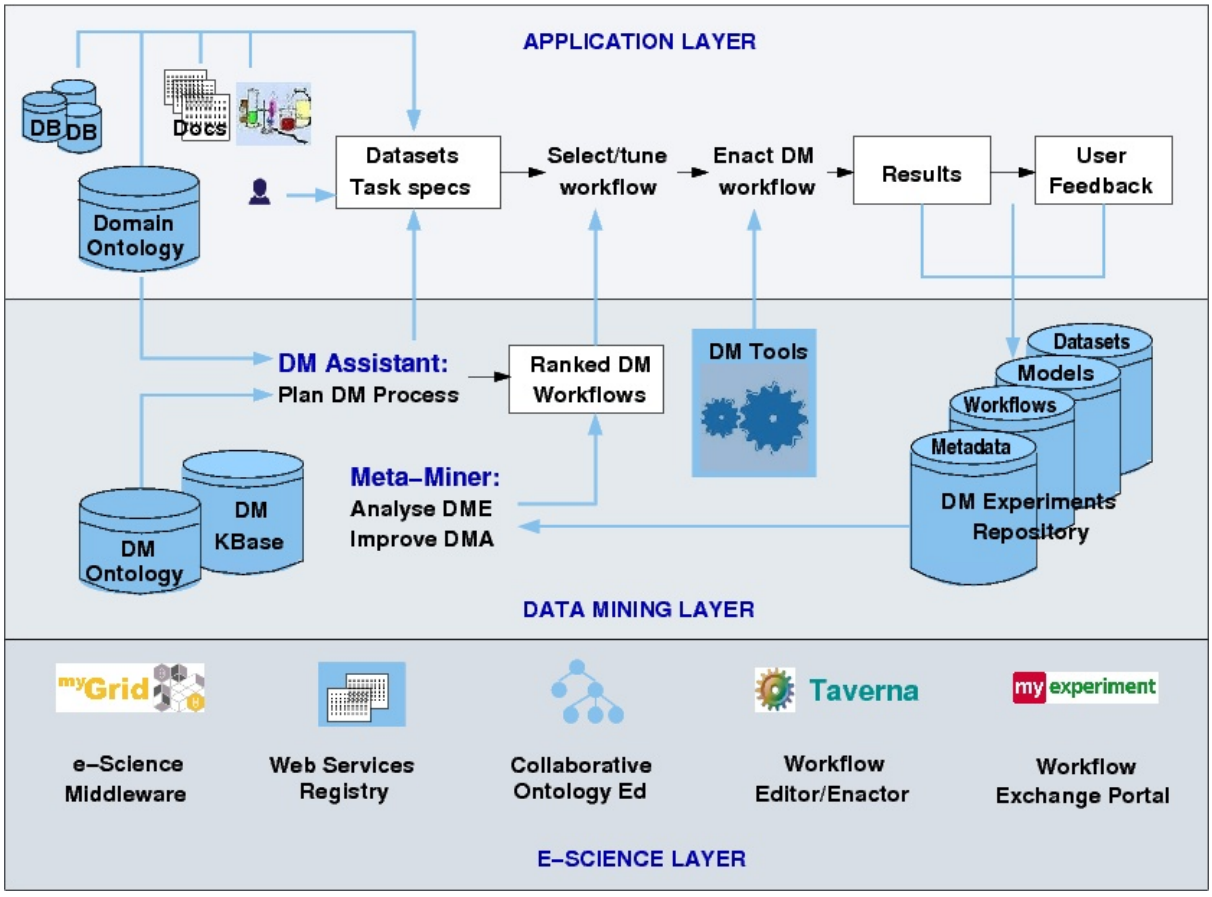
e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data-Intensive Sciences

e-LICO is a virtual laboratory built to respond to the data analysis needs of scientists at grips with massive, heterogeneous, high-dimensional data. It comprises three layers: the e-science and data mining layers form a generic knowledge discovery environment that can be easily adapted to different scientific domains by customizing the application layer.

The core of the data mining layer is a knowledge-driven discovery assistant, which relies on a data mining ontology to plan the mining process and propose workflows for a given application problem. The assistant itself is endowed with learning capabilities that allow it to improve over time. Its evolution is driven further by social networks of committed users, whose collective experience nourishes its knowledge resources and its store of tested and shared workflows. e-LICO's pilot application is a systems biology task: biomarker discovery and pathway modeling for diseases of the kidney and urinary pathways.



Summary of Activities

During the first year of the project, development of the 3 system layers was launched in parallel, with emphasis on the following work areas:

Taverna upgraded to meet the needs of e-LICO

Major upgrades were introduced into the Taverna workflow platform, leading to the release of Taverna 2, which provides improved handling of large data, better provenance tracking and support for secured web services. Taverna itself has no built-in semantics, but gathers semantics from different sources: users, shared annotated workflows from myExperiment portal, or curated registries of annotated web services such as BioCatalogue. Alongside Taverna, the e-LICO infrastructure will feature collaborative ontology building tools that allow users to pool together their expertise.

Integration of multimedia data mining functionality

RapidMiner, the leading open-source data mining software package, underwent major renovation in view of achieving e-LICO objectives. Its familiar tree-based model has been replaced by a more intuitive workflow layout that is more coherent with e-LICO's overall architecture. In addition to the 400 data mining operators offered as web services by RapidMiner and its plugins, tools for text and image processing and analysis have been integrated into the e-LICO data mining layer in the form of annotated web services.

Construction of a multi-purpose data mining ontology

One of the key activities in e-LICO is the development of a Data Mining Ontology (DMO) which will provide domain-independent technical knowledge to be harnessed throughout the knowledge discovery process. The DMO will actually be a set of modular ontologies, each built for a specific purpose. One ontology is aimed at planning and workflow construction and as such focuses on concepts such as data mining operators, preconditions and effects. Another is meant to support algorithm selection, model selection and meta-learning through a fine-grained analysis of data mining tasks, algorithms, models and datasets. A text mining ontology is being developed in collaboration with the UK National Centre for Text Mining. An initial version of these 3 ontologies is available, and an image mining ontology will be added.

Design of a planner-based Intelligent Discovery Assistant

An AI-planner based assistant has been designed to interact with the user through a graphical interface for editing descriptions of tasks, methods, and data. The assistant relies on the DM planning and workflow ontology to build hierarchical task networks and recommend a set of ranked data mining workflows that achieve the user task. Performance of these workflows will be evaluated and results stored in an experiment repository in view of subsequent analysis by a meta-miner.

User Involvement, Promotion and Awareness

The pilot user group: life scientists in COST Action EuroKUP

Cost Action EuroKUP (http://www.eurokup.org) gathers several dozen clinicians, wetlab biologists and computational biologists who share a common interest in diseases of the kidney and urinary pathways (KUP). Multiple questionnaires were sent to EuroKUP members to gather use cases and define their data analytical needs, such as biomarker discovery based on genomic, proteomic and metabolomic data. At the same time, e-LICO and EuroKUP members collaboratively build the KUP ontology and data base, specialized knowledge resources needed to make biological sense out of these high-dimensional multi-omic data.

Data mining community

Data miners are both technology providers and pilot users of the e-LICO lab. Thus e-LICO has encountered wide interest in ECML-PKDD, the premier data mining conference in Europe. Project team members presented two papers in the main conference and two others in the Workshop on Third-Generation Data Mining: Toward Service-Oriented Knowledge Discovery. Outside research circles, RapidMiner's user community will be both a test bed and a vector for disseminating e-LICO results, mainly through publication and sharing of workflows.

Semantic Web community

The e-LICO infrastructure is grounded on semantic web technologies and is developed by leading players in the field. This year they gave invited talks on semantic e-science and collaborative ontology building, organized tutorials and working sessions on OWL and OWL2, and demonstrated Taverna and BioCatalogue in ISMB-2009.

Future Work

While the first year of the project laid the groundwork of the e-LICO DM lab, the second year will be that of consolidation. The DM ontology will be widely disseminated for comments by data mining specialists. A first integrated prototype based on the planner-based DM assistant will be demonstrated and tested among EuroKUP users. The DM assistant will be extended with probabilistic reasoning capabilities as well as kernels for estimating workflow similarity. Intensive experimentation in the pilot domain will be initiated and experimental meta-data collected in the Data Mining Experiment Repository in view of meta-mining.