e-LICO Annual Report 2010

www.e-lico.eu

Project Overview

The ultimate goal of the e-LICO project is the creation of an open platform which easily allows the application of complex data analysis processes for non-analysts. To that end, the e-LICO team is building and maintaining a virtual laboratory for interdisciplinary collaborative research in data mining and data-intensive sciences. The e-lab comprises three layers:

- The e-science layer, built on an open-source e-science infrastructure, will help researchers around the world to share workflows and results, to form communities and to learn from each other. e-LICO is using the myExperiment.org portal as its main entry point to the community.
- The data mining layer is the distinctive core of e-LICO; it provides comprehensive data mining tools for various tasks and special applications. The researcher will be assisted by a knowledge-driven data mining assistant and workflow planner guiding the expert scientist through the data mining process without imposing the requirement for deep statistical knowledge on them.
- The application layer. e-LICO will be showcased in two application domains: a systems biology task (biomarker discovery and molecular pathway modelling for diseases affecting the kidney and urinary pathways) and a video recommendation task based on videos from VideoLectures.Net.

Extension of the e-LICO Consortium

As of 2010, the e-LICO consortium has grown by three new members: Josef Stefan Institute (Slovenia), Poznan University of Technology (Poland), and Ruder Boskovic Institute (Croatia) joined the team and incresed the number of members from seven to ten. The new members provide expertise, e.g., in the fields of semantic data mining, recommender systems, and multimedia mining. Algorithms developed at the respective sites have been included into the e-LICO architecture. Furthermore, they are responsible for the second application domain, a recommender system for VideoLectures.Net.

Summary of Activities

The work of the e-LICO project in 2010 was mainly focused on infrastructure: Over the first year, various components have been newly developed, improved, or adapted to the needs of e-LICO. Whereas they so far existed as individual software products, they are now integrated into the e-LICO Suite which makes these individual software products work together in a tightly integrated fashion. The products involved are the following:

- The *Taverna* workbench controls workflow enactment,
- the *myExperiment.org* community Web site hosts workflows and provides a social networking and community portal,

- the data mining solutions *RapidMiner* and *RapidAnalytics* provide data mining processes and services,
- various independent services address specialized mining and pre-processing tasks, e.g. for image mining and novel semantic data mining algorithms,
- a new data mining *workflow planner* has been developed,
- a *meta miner* is being developed to guide the aforementioned planner,
- *ontologies* are being developed as a basis for planning and meta mining.

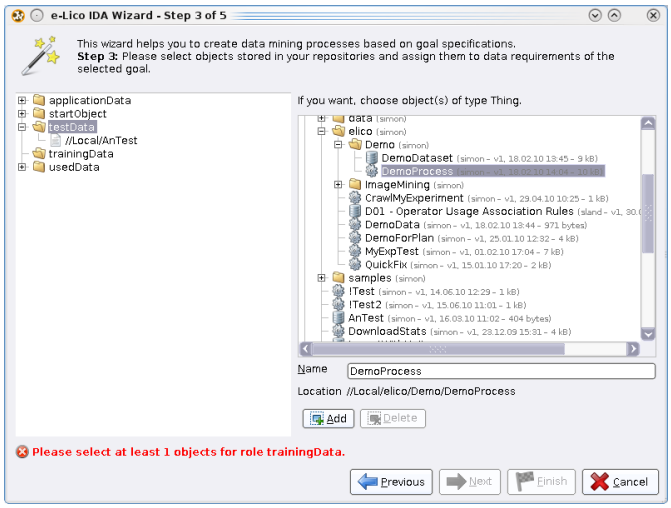
Apart from further improving the individual components, one major aspect of the work in e-LICO in 2010 was the integration of these tools that will be detailed below.

Integration of Algorithms

All of the above products have been integrated into a seamless suite: On the one hand, Taverna has a new activity type for executing data mining operators provided by RapidAnalytics. On the other hand, many of the new mining algorithms, including new image mining methods and novel state-of-the-art data mining algorithms, have been included in RapidAnalytics. Similarly, the statistics package R which has a wide acceptance in the bio domain is now also integrated as a new RapidMiner and RapidAnalytics extension. Hence, all these become available through Taverna as the e-LICO workflow enactment engine and make the suite a feature-rich toolkit. This enables the user to perform both the processing and preparation of bio-data and the actual mining task in a single user interface.

Integration of the Intelligent Discovery Assistant (IDA)

The development of the Intelligent Discovery Assistant started in year one. Its goal is to guide the expert scientist who is not experienced in statistics or data mining through the analytical process. Its core is a hierarchical planner based on ontological models of the data mining process as well as on results from meta mining and a collection of community-generated workflows. To use the IDA, the user specifies its input data characteristics as well as a task or goal to achieve, and the IDA generates a set of workflows, ranked by suitability for the particular use case.

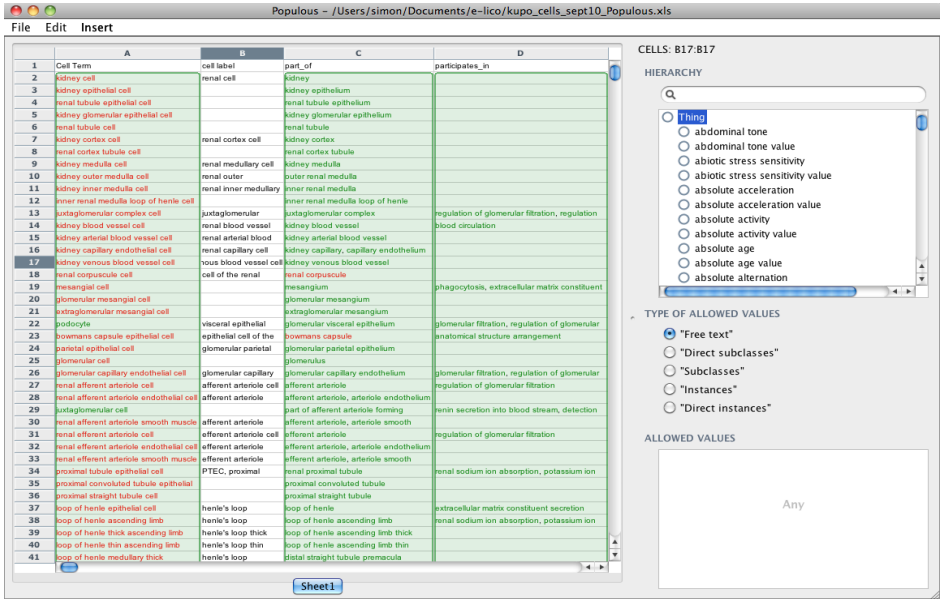


Up to now, the Intelligent Discovery Assistant was only available as a plug-in to the ontology development software Protégé. Now, it exposes an API which enabled us to build it as an extension into Taverna and RapidMiner and can be used in other products.

Thus, the planning, enactment, and sharing of data mining workflows is now possible in an integrated suite: The user selects input objects from their repository in RapidMiner or Taverna, starts the planner, and can then select from a list of proposed plans that can be directly loaded and enacted. Furthermore, the planner associates annotations with it, such that, when stored in a RapidAnalytics server, they immediately go into a case-base for later retrieval based on these annotations.

Ontology Development Tools

In year one we explored multiple tools for ontology development with a specific focus on community engagement and collaborative ontology development. In an effort to engage experts from the KUP domain in the development of KUPO, we resorted to using simple spreadsheet based templates for gathering knowledge. This approach proved successful in generating content for the KUP ontology. Our experiences lead to a set of requirements for a new generation of tools that support ontology development from spreadsheets like templates.



We implemented a prototype called *Populous* that supports embedding terms from ontologies into Microsoft Excel spreadsheets. Populous is based on the *RightField* tool and can be used to create templates for populating ontology design patterns. These templates shield the user from the underlying ontologies and technology and allow them to focus on knowledge gathering, rather than knowledge engineering. Populous supports the transformation of these templates into full blown OWL ontologies using an expressive pattern language called OPPL. The Populous approach enabled the construction of a KUP ontology by domain experts with little or no prior knowledge of OWL.

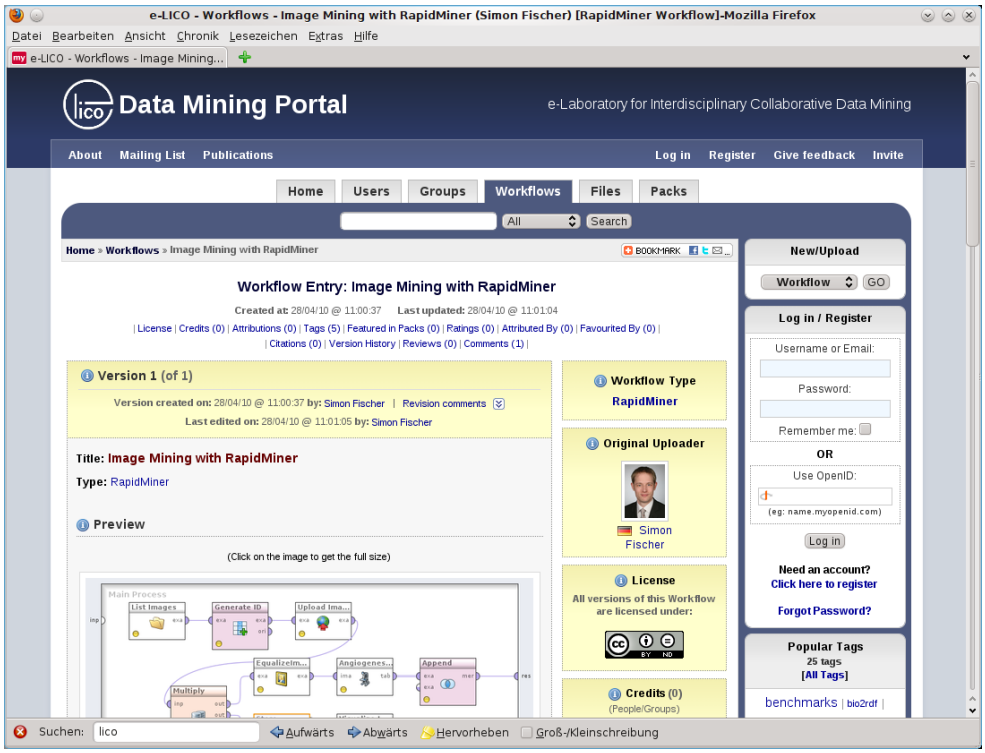
A Collaborative Ontology Development Portal has been set up within the e-LICO web site to allow for more interactive participation by interested specialists in the development of the Data Mining ontologies. Web-based tools for collaborative ontology design, such as the Cicero Argumentation Tool, are now in place and are being used to discuss open issues concerning both the content and the engineering aspects of these ontologies. Work is ongoing to link the Cicero tool and various discussion forums to an OWL Ontology Browser, such that a user can navigate not only within each ontology but also between the ontologies and the relevant issues posted on Cicero and the data mining forums. The OWL Browser will thus serve not only to explore but also to comment on and propose extensions to the ontologies.

Biological Data Collection and Infrastructure Development

During the second year of e-LICO, the collection of the biological high dimensional -omics (genomic, proteomic and metabolomic) data was completed. These datasets will be used to showcase the ability of the e-LICO platform to efficiently analyse the data and generate new research hypothesis. In parallel, a Kidney and Urinary Pathway Ontology (KUPO), along with a specialized knowledge base (KUP KB) prototype were developed to support the data-mining task. While the KUPO is used to annotate the data, the KUPKB infrastructure gives the KUP biologists the means to ask queries across many resources in order to aggregate knowledge that is necessary for answering biological questions.

Community Building

The Web site of myExperiment.org was skinned to reflect the look and feel of the e-LICO community so e-LICO researchers can identify themselves easily with their community:



Furthermore, using the portal from within the e-LICO Suite was made still easier. As with Taverna, RapidMiner users can now also share their data mining processes with a community of fellow researchers on myExperiment.org through the new Community Extension. MyExperiment will be used as the central repository for online-documentation from RapidMiner 5.1 on, which is being released in these days.

Challenge

To promote e-LICO, and to get input from the community that can be used for further analysis by the case base and meta miner, a challenge in the systems biology application domain was launched. Details about the "*e-LICO multi-omics prediction challenge with background knowledge on Obstructive Necropathy*" can be found at http://tunedit.org/challenge/ON.

User Involvement, Promotion and Awareness

As mentioned above, the scientific community is starting to notice the existence of e-LICO: The individual, very different communities of Taverna, RapidMiner, the life-sciences in general, etc. are coming together at myExperiment.org, and start forming communities and sharing workflows. E.g., there are already around 70 newly created RapidMiner workflows submitted to myExperiment.org.

The e-LICO suite was promoted in various scientific and business-oriented conferences:

In 2010, the *Workshop on Service-oriented Knowledge Discovery* (SoKD-2010) held at ECML/PKDD 2010 was organized by e-LICO members after it had already been a great success in 2009. With a focus on the use of ontologies and their use for planning, it brought much attention to the e-LICO project.

The *Planning to Learn Workshop* (PlanLearn-2010) held in conjunction with the *European Conference on Artificial Intelligence* (ECAI 2010) was co-organized by e-LICO members Abraham Bernstein and Jörg-Uwe Kietz, Zurich.

Rapid-I have organized their first RapidMiner Community Meeting and Conference (RCOMM 2010) where they also presented the new development of RapidAnalytics, the IDA, and the e-LICO project in general. In 2011, it will be held again in Dublin, Ireland. Furthermore, the results of e-LICO were presented to a business audience at the Open Source Business Intelligence Day (OSBI 2010).

Future Work

In the third year, the focus of e-LICO will be to bring the developed tools into production use. New versions of the involved tools are currently being released to the community, and further updates will be made to stabilize them and add more features. Furthermore, we will provide a bundled and well-documented package for the entire Suite to make it easily installable.

We will utilize the provided tools for actual analysis of the two use cases. A requirements specification and project plan for the systems biology domain exists and is to be executed in the final year. Furthermore, we plan to explore the use of Populous for generating templates to populate the DMOP ontology, along with extensions to support collaboration. A similar approach will be taken for the video recommender system. Furthermore, a second challenge in this application domain will be launched in 2011.

An EU FP7-ICT Project (2009-2012). Coordinator: Université de Genève (Melanie.Hilario@unige.ch)