KDI: an Information System for Clinical and Biological Data Integration application to Cancerology

Gérôme JULES-CLÉMENT1,2,3,4, Julien ROMÉJON1,3,4, Florent YVON1,3,4, Stuart POOK1, Camille BARETTE1, Georges LUCOTTE1,3,4, Séverine LAIR1,3,4, Alban LERMINE1,3,4, Stéphane LIVA1,3,4, Grégory DUVAL1,3,4, Philippe LA ROSA1,3,4, Patrick POULETT1,3,4, Jean-Gabriel DICK1, Emmanuel BARILLOT1,3,4, Philippe HUPÉ1,3,4,5

1 Institut Curie, 2 U932 INSERM, 3 U900 INSERM, 4 Mines ParisTech, 5 UMR144 CNRS

Background

Cancer is the first cause of mortality in France and in the world (13% of overall deaths), and a major public health issue. Institut Curie scientists and clinicians contribute to the development of new treatments against cancer through innovative programs. Those programs can be compared to a mosaic composed by clinical and biological data.

KDI (Knowledge and Data Integration) initiative aims to merge clinical and biological data in a unified information system. Our strategy is to improve the global quality of the data with data management good practice (traceability, reproducibility) and provide insight to the scientist during the project duration (query). This will enable large scale retrospective studies based on the existing data.

Definitions:
- A **sample** is biological material from the patient (tissue biopsy, tumor from surgery...).
- Biotechnology enables scientist to measure sample composition (DNA, RNA, protein...). The output is data whose format is specific.
- **Data** can be structured in a relational database, or stored as flat file. Biotechnology data are managed by dedicated application with end-user interface.

Each application tackle the data complexity and heterogeneity and KDI system ensures inter-operability between application.

Data integration workflow

**Technical aspects**

- KDI core system is developed with J2EE technology. We have focus on the modularity of the application.
- **Database engineering**: relational database with entity-attribute-value (EAV) design.
- **Software architecture**: n-tiers architecture
  - **Data persistence**: Hibernate (MySQL)
  - **Service**: Spring
  - **View**: Java Serveur Faces (JSF) 2 with Trinidad and PrimeFaces
- **Inter-operability**: Web Services (SOAP)

**Perspective**

The major ongoing challenges are:
- **Takeover** the existing data into KDI information system
- Deal with the existing code and **migrate applications** to the n-tiers architecture.
- Data heterogeneity: use **Ontologies** within each application
- Use **Big Data** management system for high throughput biotechnology results?

Reference:
- [http://www.oracle.com/](http://www.oracle.com/)
- [http://www.mysql.com/](http://www.mysql.com/)
- [http://www.hibernate.org/](http://www.hibernate.org/)
- [http://www.springsource.org/](http://www.springsource.org/)
- [http://primefaces.org/](http://primefaces.org/)

**Data integration workflow**

1. Sample is collected from patient
2. Sample are processed by biotechnologies
3. Data produced is managed by dedicated application
4. Data is referenced in KDI core system
5. New data can be infered by an analysis