Enabling biomedical data analysis workflows: the Multi-Knowledge collaborative platform

Michele Amoretti, Francesco Zanichelli, Gianni Conte

Information Engineering Department, University of Parma, Via Usberti 181/a,
43100 Parma, Italy
{michele.amoretti, francesco.zanichelli, gianni.conte}@unipr.it

Abstract. The objective of the Multi-Knowledge project is the development and validation of a collaborative IT platform for knowledge management, allowing geographically dispersed groups of researchers, dealing with different data sources as well as technological and organisational contexts, to create, exchange and manipulate new knowledge in a seamless fashion. The ambition is also to define a methodological framework that can easily be extended to include additional sources of knowledge and expertise (biomedical data, images, environmental data), and can be applied to wider sectors of medical research. After two years of work, the Multi-Knowledge platform is almost complete and the second pilot experiment is being carried out. In this paper we describe the Multi-Knowledge project, starting from user requirements which have driven the development process, then going into details of the different modules which compose the platform, and finally illustrating the experiments which are being conducted across different sites.

Keywords: collaborative platform, knowledge extraction, e-health

1 Introduction

The Multi-Knowledge project [1], which is funded by the European Commission in the context of the Sixth Framework Programme for Research and Technological Development (thematic area Information Society Technologies), arises from the data processing needs of a network of medical research centres, located in Europe and USA, performing research activities in the field of metabolic and cardiovascular diseases. These needs are mostly related to the integration of three main sources of information, namely clinical data, patient-specific genomic/proteomic data (in particular information acquired by means of microarray technology), and demographic data.

In this context the main objective of the Multi-Knowledge project is related to the development and validation of a collaborative IT platform for knowledge management, allowing geographically dispersed groups of researchers, dealing with different data sources as well as technological and organisational contexts, to create, exchange and manipulate new knowledge in a seamless fashion. The project also aspires to define a methodological framework that can easily be extended to include additional sources of knowledge and expertise (biomedical data, images, environmental data), and can be applied to wider sectors of medical research. After two years of extensive work, the Multi-Knowledge platform is approaching its full realization and the second pilot experiment is currently in progress.

Critical and difficult issues addressed in the project concern the management of data which are heterogeneous in nature (continuous and categorical, with different order of magnitude, different degree of precision, etc.), origin (statistical programs, manual introduction from an operator, etc.), and coming from different data environments (from the clinical setting to the molecular biology lab). The Multi-Knowledge service-oriented architecture enables workflow design and execution based on novel operating procedures to manage and combine heterogeneous data and make them easily available for data analysis.

The paper is organized as follows. In section 2 we describe user requirements which have driven the development process of the MK platform. In section 3 we discuss relevant literature on computer supported cooperative work (CSCW). In section 4 we describe the MK platform, focusing on its functional properties which are made available by the synergy of several modules. In section 5 we describe the experiments which are being conducted across different sites. Finally, in section 6 we summarize results and contributions.

2 User Requirements

As dictated by modern software engineering, before tackling the design and implementation of the MK platform, we identified actors and analyzed the socio-organizational context in which they operate. This process led us to the definition of a specific set of use cases.

An actor is a subject that performs an activity. In the MK context, there are two types of actors: human beings, and non-human beings (like a system or an organization as a whole). When an actor is executing an activity he is playing a role. For instance, a woman (human actor) may play the role of a doctor and, at the same time, of a patient. In general, we defined an actor to be tuple

<**role**, **scope**, **responsibilities**, **background & skills**>

In the domain addressed by the MK project, a human actor may have one of the following roles:

- Practicing Physician
- Biomedical Researcher (clinical or basic)
- Biostatistician
- Principal Investigator
- System Administrator
- Security Manager

In the same context, non-human actors may be:

- Pharma Industry
- Clinical Research Organization

- Clinical Research Center

Describing each actor type in depth, with related responsibilities, background and needs, is not possible because of space limits. We prefer to show and discuss the activity scopes which have been identified and associated to actor types (see table 1).

Table 1. Activity scopes covered by the Multi-Knowledge platform.

Scope	Meaning
Experiment	In the Multi-Knowledge terminology an *experiment* begins with the definition of the final goals and of the target patients, and ends with the analysis of the results. An *experiment* may occur several times and in different manners.
Pharma research	A research activity of a pharmaceutical industry, like the testing of a new drug, from the profiling of the tests sets till the analysis of the final results.
Large scale study	A multi-center large scale clinical research study, typically aimed at obtaining approval from regulatory agencies or conducted in relation with the pharma research context (see above).
Multi-Knowledge project	The usage of the Multi-Knowledge platform to enhance the conduction of experiments and research studies.

As we briefly described in section 1, the Multi-Knowledge project stems from the data processing needs of a network of medical research centres performing research activities in the field of metabolic and cardiovascular diseases. In this research context, a central role is played by technologies that support gene expression measurement. Gene expression is a highly complex and tightly regulated process that allows a cell to respond dynamically to environmental stimuli and to its own changing needs. The expression profiles of all cells in an organism at a given time t can be thought of as the expression profile of the organism at t. While the genome of an organism is relatively invariant, the expression profiles of cells, of cell populations, and of organisms are highly variable, changing over time as a function of developmental, environmental and other conditions. For this reason, the core of the MK project lays on the integration of information coming from different levels of reading of the disease process, from clinical to genomic/proteomic, to better characterize a pre-disease phenotype highly suggestive to evolve in clinically manifest diseases.

Given this premise, use cases have been grouped into five sets. A first important one (*Data Integration*) relates to the introduction and integration of heterogeneous data into the Multi-Knowledge platform. This data is generated by diverse sources: medical history of the subjects and their first degree relatives, laboratory data, data from physical examinations, emerging biochemical markers (e.g. related to insulin resistance and CV risk), instrumental findings (*e.g.* flow-mediated vasodilatation test results), genomic/proteomic data from microarray chips. In this context, a specific use case is devoted to data cleaning and normalization, performed by biostatisticians. A second set of use cases (*Data Analysis*) relates to on-line statistical processing and data mining: identification of differentially expressed genes, determination of a bynary or higher partition of the sample set to be studied for differential expression, assessment of differential expression of genes for a quantitative trait such as blood pressure or IMT, identification of functional enrichment in over or under expressed genes by means of statistical models, development of classification signatures, etc. A third set of use cases addresses the general requirements of the workflow system adopted within the MK platform: workflow definition, w. initiation, w. processing, workflow process recording and monitoring, w. p. repeatability and versioning, definition and management of research paths, experience feedback, support co-design of gene or protein microarray chips to conduct specific experiments, and off-line analysis. In the same set we put the use cases which are related to experiment conduction, illustrated in figure 1. Another set of use cases is related to system administration functions: login, add user, definition of security policy, distributed data backup, infrastructure management. The last set of use cases is related to system security management and enforcement of data privacy and protection, according to international rules and laws that the Multi-knowledge system was demanded to comply with: compliancy with European Union rules for data protection, compliancy with the requirements of e-Privacy legislation regarding communications infrastructure, enforcement of compliancy with interoperability security constraints, authentication and authorization, data encryption, patient anonymization, etc.

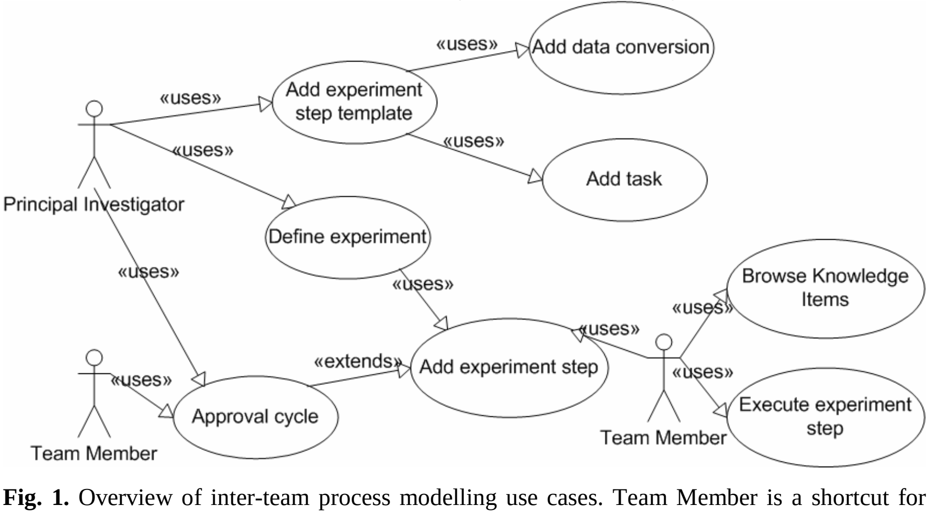

Fig. 1. Overview of inter-team process modelling use cases. Team Member is a shortcut for indicating both Biomedical Researcher and Biostatistician roles.

3 Related work

From the technical point of view, the MK platform is a Computer-Supported Cooperative Work (CSCW) system. Many researchers of information systems criticize current CSCW systems based on their usability and users' satisfaction. The

Multi-Knowledge project has been entirely conducted with the supervision of end users, i.e. biomedical researchers, biostatisticians and also practicing physicians. This approach, suggested for example by Ruppel et al. [2], lead the consortium to obtain a better-quality application, and better acceptance, which is particularly important in the case of collaborative systems.

The first feature demanded to the MK platform is knowledge sharing support. As suggested by Kindberg et al. [3], we distinguished between several types of knowledge: data, domain, users (their competences, their needs). The premise of MK-supported activities is that researchers from different organizations and institues agree on sharing the (anonymized) data they have on their patients, and to exchange specialized knowledge. The issue of exchaning patient-related data is being partially solved by the increasing adoption of electronic medical records (EMRs), instead of traditional paper medical records (PMRs). Bringay et al. [4] observed that practitioners still prefer to use the PMR to collaborate, because electronic medical documents do not allow reproducing some practices of collaboration carried out with the PMR, in particular one practice: annotations which are used as support for the collaboration. In the research context of the Multi-Knowledge project, EPRs are required since clinical data must be automatically integrated with genomic/proteomic data (in particular information acquired by means of microarray technology), and demographic data.

Knowledge sharing is just one kind of cooperative activity the MK platform was demanded to support. The other categories, according to Bardram’s classification [5], are: organization of work, planning and scheduling, and communication, with the general objective of creating new knowledge. In the first phase of the project we explicitly detailed these activities, with particular emphasys on the identification of different roles for human actors (as we summarized in section 2). Role-Based Collaboration (RBC) theory is a natural approach to integrate the theory of roles into the CSCW systems [6,7,8].

From the specification of user requirements, the importance of the Process Modelling component (including Knowledge Extraction) of the Multi-Knowledge system definitely emerged. Workflow management technology is not used in healthcare as often as in other domains. Healthcare workflows have “transactional” elements, such as admitting a patient or taking a blood glucose measurement, but focusing on individual transactions obfuscates the most important element of healthcare workflows—the need to flexibly promote and maintain the highest possible standard of care for patients [9]. In the field of biomedial research, Workflow Management Systems (WMS) are seen as a viable solution for the creation and deployment of new flexible and extensible data integration and analysis network tools [10,11]. Some WMS have been proposed [12,13] and are now under careful testing aimed at the verification of their actual ability to cope with the data integration issue. While their potentiality is clear, some limitations are now arising, including both network issues (e.g., quality of service, speed, access restrictions) and practical issues (e.g., long running jobs, huge input/output).

In this context, Multi-Knowledge can be compared to COCOON [14], which is aimed at activating regional semantics-based healthcare information infrastructures with the goal of reducing medical errors, and ARTEMIS [15], whose objective is to develop a semantic framework for the healthcare domain, building upon a peer-to-peer architecture in order to facilitate the discovery of healthcare services. Examples of such services are those listed by the Biological Web Services (BWS) page [16]. Among all, GeneCruiser [17] is a Web Service for the annotation of microarray data, developed at the Broad Institute (a research collaboration of MIT, Harvard and its affiliated hospitals). GeneCruiser allows users to annotate their genomic data by mapping microarray feature identifiers to gene identifiers from databases, such as UniGene, while providing links to web resources, such as the UCSC Genome Browser. It relies on a regularly updated database that retrieves and indexes the mappings between microarray probes and genomic databases. Genes are identified using the Life Sciences Identifier standard. A more complex example of Web Service-oriented architecture providing transparent access to biomedical applications on distributed computational resources is the National Biomedical Computation Resource (NBCR) [18], which is based on Grid technologies such as Globus Toolkit. NBCR users are allowed to design and execute complex biomedical analysis pipelines or workflows of services.

Compared to these initiatives, the Multi-Knowledge project is a step forward since its objective is the creation of *collaborative environments in which many kinds of actors participate in the workflow execution*. Experiment steps are defined by and conducted under the responsibility of a research team including biomedical researchers, bioinformaticians, practicing physicians, and coordinated by a principal investigator.

4 The Multi-Knowledge platform

The Multi-Knowledge distributed platform (see figure 1), based on the client/server paradigm, integrates the following modules:

- Portal
- Workflow Designer and Execution Engine (MK-WF)
- Data Collection and Normalization (MK-DCNS)
- Data Analysis Tool (MK-DA)
- Visualization Tool (MK-VIZ)
- Report Generator and Manager (MK-REP)

Instead of describing each module separately, we prefer to analyze three funtionalities of the MK platform, and to discuss how and which modules are involved in each of them.

4.1 Multi-Knoweldge Portal

The Portal module is the entry point for different kind of users: practicing physicians involved in clinical data collection, biomedical researchers charged with genomic data normalization, principal investigators and biomedical researchers interested in browsing and downloading workflow descriptions and experiment reports. Moreover, the Portal allows to obtain the MK-DA tool (which includes the Report Generator), the Workflow Designer tool, and the MK-VIZ tool.

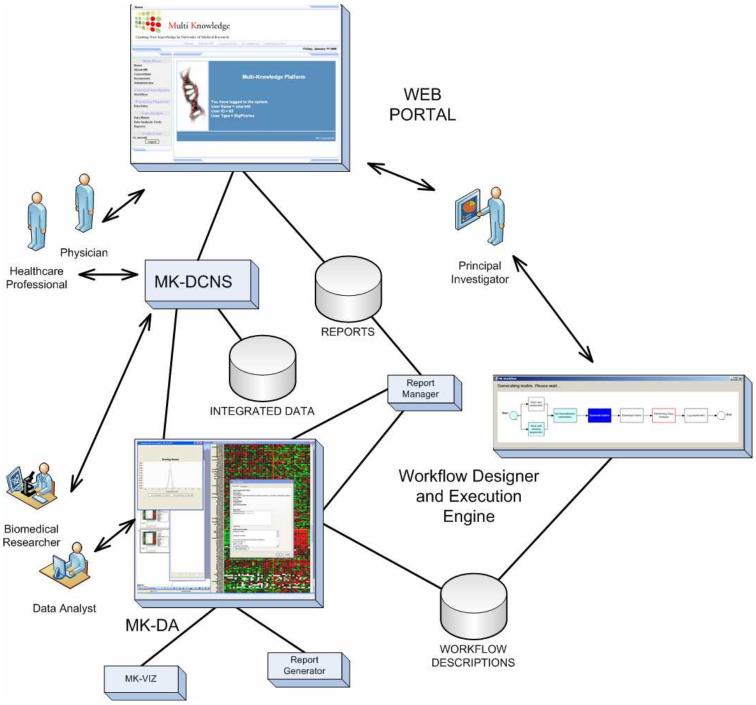

Fig. 2. The Multi-Knowledge platform, with different kinds of users participating in the knowledge extraction process.

The Portal is based on Joomla, which is a popular content management system and Web application framework (evolution of Mambo), but integrates also Java Server Pages (JSP) and ASP.NET pages. From this point of view, the Multi-Knowledge Portal is a pioneering artifact. The main issue has been the definition of a strategy for session sharing among Joomla PHP pages and the JSP/ASP pages of the other MK modules, which are deployed in different Web application engines. By achieving session sharing, it has been possible to introduce Single Sign-On (SSO), *i.e.* a method of access control that enables a user to authenticate once and gain access to differently deployed Multi-Knowledge modules. Another feature is role-based access, for which *e.g.* users with principal investigator access level are allowed to access most Multi-Knowledge modules, but they are not allowed to associate patients' clinical data with their vital statistics, for which only only practicing physicians are responsible.

4.2 Multi-Knowledge Workflows

In section 2 we summarized the use cases which address the general requirements of the workflow system adopted within the MK platform. As this is a crucial part of the project and one of the most innovative, a supplemental analysis level is needed, with respect to the previous subsections.

A Multi-Knowledge research experiment consists of a set of Experiment steps, defined by and conducted under the responsibility of a research team, coordinated by a Principal Investigator, that aims at achieving an established scientific goal. We have illustrated how the MK platform supports data collection and normalization, as well as data analysis, visualization and reporting. Now we focus on the orchestration of a whole experiment.

Each of the data analysis step may generate new knowledge elements that contribute to create and successively expand an *experiment-related body of knowledge (EBoK)*. Based on an analysis of the EBoK (performed from their different scientific point of views) research team members can propose the execution of additional experiment steps or to further carry on the process. This means that we may see the process as a spiral in which every cycle allows a better focus on the scientific objective which was initially stated for the experiment.

It is important to note again that the crucial objective of the Multi-Knowledge Process Modelling component is to guarantee the seamless integration of the different contribution brought in by the different research team members. This is not an always easy task because, for instance, some data analysis steps may be proposed/requested by researchers that do not have the specific expertise to conduct them while, vice versa, those who have the expertise to conduct them may not be able to understand the full, specific implications of the resulting knowledge.

This is better clarified if we compare the main interaction patterns among the researchers. There are two major classes of interaction patterns, represented in the following sequence diagrams: the *supervised experiment* and the *unsupervised experiment*, illustrated in figure 3.

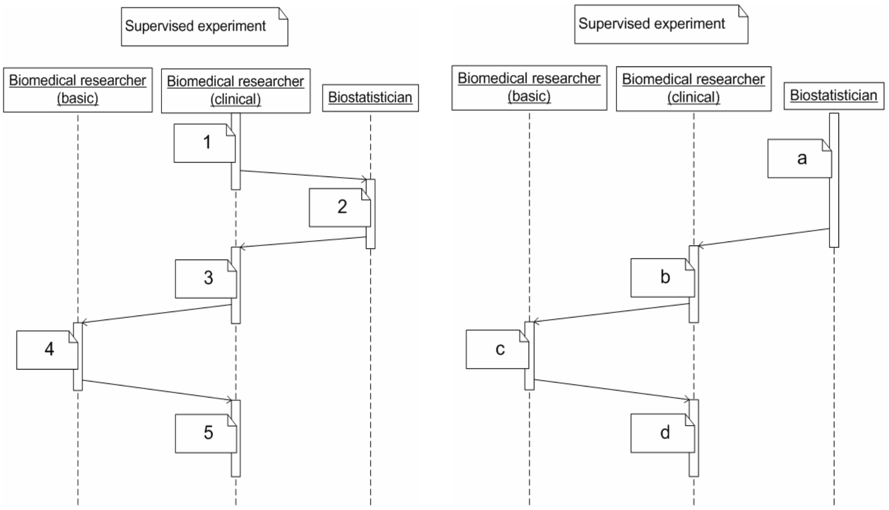
Supervised experiment
Biomedical researcher (basic)
Biomedical researcher (clinical)
Biostatistician
1
2
3
4
5
Supervised experiment
Biomedical researcher (basic)
Biomedical researcher (clinical)
Biostatistician
a
b
c
d

Fig. 3. Interaction among researchers: supervised and unsupervised experiments.

In the supervised experiment, the biomedical researcher – clinical approaches the data sample from a clinical point of view. She/he may perform some data analysis steps relying on the class of tools normally used and understood by clinicians. These may include, for instance, clinical data quality check, preliminary study of clinical data distribution among patients in the sample, etc. These activities are marked as <1> in the above interaction diagram. Based on the knowledge she/he has extracted from these initial activities, the biomedical researcher – clinical may ask the intervention of a biostatistician to carry out genomic related analysis, for instance to correlate an interesting clinical variable to differentially expressed genes. From this point on, it is up to the biostatistician to perform data analysis steps, relying on the class of genomic/proteomic statistical analysis tools that she/he normally use (activity marked as <2> in the diagram). When the biostatistician obtains the results of her/his activity, she/he passes the experiment thread back to the biomedical researcher – clinical, who studies the additional knowledge coming from the biostatistician work (activity marked as <3>). Then the biomedical researcher – clinical may continue with other analysis cycles, possibly by requiring again the intervention of colleagues. The sequence diagram shows, for example, that she/he asks intervention of a biomedical researcher – basic to obtain more biological insight in some part of the experiment (activities marked as <4> and <5> in the diagram).

However, it is not always that case that the experiment conduction starts in the hands of the biomedical researcher – clinical. As shown in the Unsupervised experiment part of the sequence diagram above, the biostatistician may start analysis of data, for instance by looking for interesting clustering patterns among genes (activity marked as <a>). It is important to note that this is a purely statistical activity, that has not yet any strong connection with clinical/biological semantics. If the biostatistician identifies something that stimulates curiosity, she/he may communicate this to a colleague, for instance to a biomedical researcher – clinical, who can try to understand if the special statistical behaviour may reveal a biological explanation, generating in this process new clinical knowledge (activity marked as <b> in the diagram). As in the previous case, to complete her/his work the biomedical researcher – clinical may ask support from other colleagues, for instance from a biomedical researcher – basic (activities marked as <c> and <d> in the diagram).

The **Workflow Designer and Execution Engine (MK-WF)** allows principal investigators to define experiment steps to be conducted asynchronously or according to declared workflow patterns, passing control back and forth from different researchers. The experiments consist of dynamical cycles of data collection and analysis that aim at progressively achieving the scientific goal initially stated for the experiment.

When a team member, possibly after receiving a suggestion sent by another team member or by the principal investigator, decides to execute an experiment step, he/she:

- revises the proposed experiment step definition (a XML file) and possibly improves it based on her/his specific knowledge;
- executes the experiment step;
- generates a report, presenting the motivations for the experiment step as well as comments on step's execution and outcome.

The workflow engine reacts by logging the experiment step that has been executed, in terms of task identifier, ask parameters and used data set, and by recording the report produced by the team member that executed the step.

4.3 Data collection and normalization

Different kinds of biological data can be collected by means of the **Data collection and normalization (MK-DCNS)** module. Referring to RNA expression arrays and protein arrays, *microarray measurements* are given as a set of feature extraction (FE) files and an indication of which columns from them to use. Each FE file represents one experiment and contains all the data derived from that microarray. Each expression FE file contains data on about 40000 genes and each protein FE file contains data on about 100 proteins. Furthermore, metabolomics data are given as tab delimited text files with two columns. The first column contains the metabolite description and the second column contains the corresponding numerical values and units. Finally, *IMT and FMD data* from each patient are entered to the system either manually, or through the MK-DCNS GUI.

4.4 Data analysis, visualization, and reporting

Starting from a patients' data sample, usually defined and collected in the first work phases, the experiment is set to conduct successive data analysis cycles, aimed at extracting new knowledge through the exploitation of full integration among heterogeneous data (clinical, demographical, genomic and proteomic) managed by a diverse set of researchers. Data analysis steps form the core of the experiment's analysis cycle. Through them, the data sample is successively analysed by different classes of researchers (having different "scientific cultures" and backgrounds) that use different analysis tools, work in different environments, at geographically dispersed sites.

The **Data Analysis Tool (MK-DA)** supports many data mining processes, such as *GO (Gene Ontology) analysis*, *classification*, *clustering*, *class discovery*, and *sequence motifs finding*. The MK-DA is integrated with the **Visualization Tool (MK-VIZ)** and **Report Generator and Manager (MK-REP)**. The latter includes a ReportGenerator which builds PDF reports using the experiment description produced by the MK-WF module, together with data analysis results and images produced by the MK-VIZ tool. The MK-REP module also includes a ReportManager service, implemented with Web Service technologies, which can be accessed by the Mk-DA tool to store or retrieve reports in a specific database.

5 Pilot Experiments

In a first instance of Multi-Knowldge pilot study, clinical, laboratory, instrumental and genomic information has been collected from 50 subjects by the Department of Internal Medicine of the University of Parma. The sample has been used to validate the first Multi-Knowledge platform prototype, in particular the system modules related to data collection and normalization. Presenting medical results is out of the scope of this paper, but the interested reader can refer *e.g.* to [19].

The second instance of the pilot experiment has required further recruitment, up to a sample of about 150-200 subjects. This study is being performed to test a full-featured Multi-Knowledge system prototype, involving Internal Medicine Department at University of Parma, King's College London, Stanford University Medical Center, Technion Tel Aviv, University of Milan , and in conjunction with two related projects (the european Pocemon and the italian Sympar). These partners are testing in particular the MK-WF and the MK-DA to exchange information about analysis workflow and to perform incremental data analysis as the pilot data set is modified.

6 Conclusions

In this paper we presented the Multi-Knowledge platform, supporting geographically dispersed groups of researchers, dealing with different data sources as well as technological and organizational contexts, to create, exchange and manipulate new knowledge in a seamless way. We started from the analysis of the socio-organizational context in which users are operating, with the listing of use cases and roles which may be assumed by participating actors. Then we presented relevant CSCW literature, in particular that related to knowledge sharing and workflow management systems. We also compared the Multi-Knowledge project to other initiatives funded by the EU, with similar purposes (*i.e.* creating and sharing integrated biomedical information for better health). In the second part of the paper, we described the modules which compose the MK platform, with particular emphasys on the workflow engine. We described in details the process models of two kinds of experiment, supervised and unsupervised. Finally we summarized the activities carried out in the first pilot experiment, and those that are being conducted in the second (more complex) pilot experiment.

Acknowledgements

This work has been supported by the european project CEE IST MULTI-KNOWLEDGE (FP6-IST-2004-027106). We would like to thank Dr. Franco Mercalli (Centro di Cultura Scientifica "A. Volta"), Prof. Ivana Zavaroni and Dr. Diego Ardigò (Department of Internal Medicine of the University of Parma) for their valuable contribution to the finalization of this work.

References

1. MULTI-KNOWLEDGE Consortium: project homepage. http://www.multiknowledge.eu

2. Ruppel, C., Konecny, J.: The role of IS Personnel in Web-based Systems Development: The Case of Health Care Organization. In: ACM SIGCPR Conference on Computer Personnel Research, pp. 130--135. ACM Press, New York (2000)

3. Kindberg, T., Bryann-Kinns, N., Makwana, R.: Suppoting the shared care of diabetic patients. In: International ACM SIGGROUP Conference on Supporting Group Work, pp. 91--100. ACM Press, NewYork (1999)

4. Bringay, S., Barry, C., Charlet, J.: Annotations: A Functionalty to support Cooperation, Coordination and Awareness in the Electronic Medical Record. In: Cooperative Systems Design (COOP '06), pp. 39--54. IOS Press, Amsterdam (2006)

5. Bardram, J. E.: Collaboration, Coordination, and Computer Support, An Activity Theoretical Approach to the Design of Computer Supported Cooperative Work. PhD Thesis, University of Aarhus, Denmark (1998)

6. Edwards, W.K.: Policies and Roles in Collaborative Applications. In: ACM Conference on Computer-Supported Cooperative Work (CSCW'96), pp. 11--20. Cambridge, USA (1996)

7. Guzdial, M., Rick, J., and Kerimbaev, B.: Recognizing and Supporting Roles in CSCW. In: ACM Conference on Computer-Supported Cooperative Work (CSCW'00), pp. 261--268. Philadelphia, Pennsylvania, USA (2000)

8. Smith, R. B., Hixon, R. and Horan, B.: Supporting Flexible Roles in a Shared Space. In: ACM Conference on Computer-Supported Cooperative Work (CSCW'98), pp. 197--206. Seattle, Washington, USA (1998)

9. LeMoine, D.: Going with the Flow: Interaction Design for Healthcare. In: Journal of Design, Cooper Consulting (2003)

10. Stein, L.: Creating a bioinformatics nation. In: Nature, No. 417, pp. 119--120 (2002)

11. Romano, P., Rasi, C. and Marra, D.: The automation of bioinformatics processes through workflow management systems. In: Seventh Spanish Symposium on Bioinformatics and Computational Biology (JdB06), Zaragoza, Spain (2006)

12. Oinn, T., Addis, M. et al.: Taverna: a tool for the composition and enactment of bioinformatics workflows. In: Bioinformatics, Vol. 20, No. 17, pp. 3045--3054 (2004)

13. Stevens, R., Robinson, A. and Goble, C.: myGrid: personalised bioinformatics on the information grid. In: Bioinformatics, Vol. 19, No.1, pp. 302--304 (2003)

14. COCOON consortium: project homepage. http://www.cocoon-health.com

15. ARTEMIS consortium: project homepage. http://www.srdc.metu.edu.tr/webpage/projects/artemis/index.html

16. Hull, D.: The Biological Web Services page. http://taverna.sourceforge.net/index.php?doc=services.html

17. Liefeld, T., Reich, M., Gould, J., Zhang, P., Tamayo, P. and Mesirov, J. P.: GeneCruiser: a Web Service for the annotation of microarray data. In: Bioinformatics, Vol. 21, No. 18, pp. 3681--3682 (2005)

18. Krishnan, S., Baldridge, K., Greenberg, J., Stearn, B. and Bhatia, K.: An End-to-end Web Services-based Infrastructure for Biomedical Applications. In: 6th IEEE/ACM International Workshop on Grid Computing, Seattle, Washington, USA (2005)

19. Steinfeld, I., Navon, R., Ardigò, D., Zavaroni, I. and Yakhini, Z.: Semi-supervised class discovery using quantitative phenotypes – CVD as a case study. In*:* BMC Bioinformatics, 8(Suppl 8)**:**S6 (2007)