DebugIT: Building a European distributed clinical data mining network to foster the fight against microbial diseases

Christian LOVIS[a,1], Teodoro DOUGLAS[a], Emilie PASCHE[a], Patrick RUCH[a], Dirk COLAERT[b], and Karl STROETMANN[c]
[a]*University Hospitals of Geneva, Switzerland*
[b]*Agfa HealthCare, Belgium*
[c]*empirica GmbH, Germany*
for the DebugIT consortium

Abstract. The expansion of clinical information systems and the reduction in computing costs have led to an explosion of patient data available for reuse. However, this data is rarely combined and analyzed in an integrated manner, The debugIT project is a large-scale integrating project funded within the 7th EU Framework Programme (FP7). The main objectives of the project are to build IT tools that should have significant impacts for the monitoring and the control of infectious diseases and antimicrobial resistances in Europe; this will be realized by building a technical and semantic infrastructure able to a) share heterogeneous clinical data sets from different hospitals in different countries, with different languages and legislations; b) analyze large amounts of this clinical data with advanced multi-modal data mining; c) apply the obtained knowledge for clinical decisions and outcome monitoring. The concepts and architecture underlying this project are discussed.

Keywords. Infectious disease, patient safety, semantic inter-operability, multi-modal data mining, decision support, clinical outcome monitoring

Introduction

Building a safer and more efficient care system has become the most shared goal of all actors involved in healthcare. From a historical perspective, there has been an impressive shift towards awareness of the impact of errors in medicine in the last 25 years. In the early nineties, research papers and reports about patient safety, incident reporting and initial order-entry systems were published, mostly originating from academic settings. At about the same time, the first reports of the US Institute of Medicine (IOM) on computerized patient record systems stressed the ability of ICT-based solutions to improve the quality of care [1]. Ten years later, by the end of the nineties, a famous report of the IOM called attention to the wide prevalence of errors in healthcare [2]. While medical errors are under the spotlight, (re-)emerging infectious diseases are becoming major challenges. Among them, the rapid development of anti-

[1] Corresponding Author: Service of Medical Informatics, University Hospitals of Geneva; E-mail: Christian.lovis@hcuge.ch

microbial resistances [3], the spread of nosocomial and other infections [4], the inadequate care and missing appropriate tools to lead the care system facing these new emergent problems [5] are major concerns. The issues around infectious diseases are strongly interrelated and have immediate and important effects on safety, quality of care and efficiency. In half a century of antibiotic use, new challenges have emerged: fast emergence of resistances among pathogens, misuse and overuse of antibiotics. Antimicrobial resistance results in escalating healthcare costs, increased morbidity and mortality and the (re-) emergence of potentially untreatable pathogens.

1. The project

Dedicated to infectious diseases, the DebugIT (Detecting and Eliminating Bacteria UsinG Information Technology) project aims at (1) detecting patient safety related patterns and trends, (2) acquiring new knowledge and (3) using this for better quality healthcare. A consortium of eleven partners has been built in order to gather scientific competencies in all domains involved, as well as to assure access to specific information of more than 2 million clinical records.

Table 1. Consortium

Belgium	Agfa HealthCare
Bulgaria	GAMA/SOFIA Ltd.
Czech Republic	IZIP - Internetový Přistup Ke Zdravotním Informacím Pacienta
France	Institut National de la Santé et de la Recherche Médicale (with European Hospital George Pompidou) (INSERM)
Germany	empirica GmbH
Germany	University Medical Center Freiburg (UKLFR)
Greece	Technological Educational Institute of Lamia
Sweden	Linköping University (LIU)
Switzerland	University Hospital of Geneva (HUG)
Switzerland	University of Geneva, Computer Science Department
United Kingdom	University College London

The project has a strong clinical lead guaranteed by a Clinical Advisory Board and a Scientific Advisory Board with European and American experts of the infectiology field and the scientific domains involved.

Outcomes and benefits, in clinical and socio-economic terms, will be measured. Clinical results will be integrated into Clinical Information Systems (CIS) of participating European hospitals, industry and their clients, and become available globally through a European or global Disease Control Centre/Public Authority, also as Open Source solution. Advanced ICT applications and innovations concern the virtualization of the Clinical Data Repository through ontology and terminology binding and mediation, advanced data mining techniques, the use of machine reasoning related to real, point-of-care patient data, as well as consolidation of all these

techniques in a comprehensive but open framework. Output will be applicable to other clinical fields.

The comprehensive concept, developed from such considerations as foundation for the DebugIT project, addresses all of these issues in an operational manner with the ultimate goal to develop a new, highly advanced and pre-eminent tool aiming at producing a new and efficient weapon for the war against infectious pathogens across all health system actors and levels.

The overall project outcome will not only be a theoretical work and proof of concept, but also a practical implementation of a highly improved and advanced computerised system in the field of infectious disease treatment and antibiotics usage. This application, which, due to its generic conceptual base, should be easily expandable and adaptable to other similar medical application fields, will initially be evaluated by participating project partners, but should be made publicly available to other healthcare organizations soon after.

2. The Conceptual Framework

The conceptual framework of this project is an ever continuing iterative cycle, implementing the principle of translational medicine and true Evidence Based Medicine (EBM). Translational medicine makes the connection between medical research and clinical care by providing to research clinical data and providing the results of the research – the medical knowledge – as input for clinical care. While medical research is often focusing on prospective and tightly controlled studies, retrospective studies with access to huge amounts of data, just waiting to be analyzed, are a welcome addition to clinical research.

The framework can be broken down into several distinct steps:

Collect Data. Clinical data will be collected and aggregated across different hospitals, countries, languages, information models and legislations, via advanced and commonly agreed data models (minimal data sets), standards and mapping algorithms.

Learn: Advanced data mining techniques on multimodal, multi-source, structured and unstructured data will detect patterns, relevant for patient safety and the treatment of infectious diseases such as: resistance of bacteria, adverse events and operational practices. This will result in new knowledge and new evidences for existing knowledge.

Store Knowledge: This knowledge will be stored, visualized, validated and aggregated together with pre-existing medical and biological knowledge (guidelines, regulations) to achieve a consolidated view on the needed knowledge, to be applied in the next step.

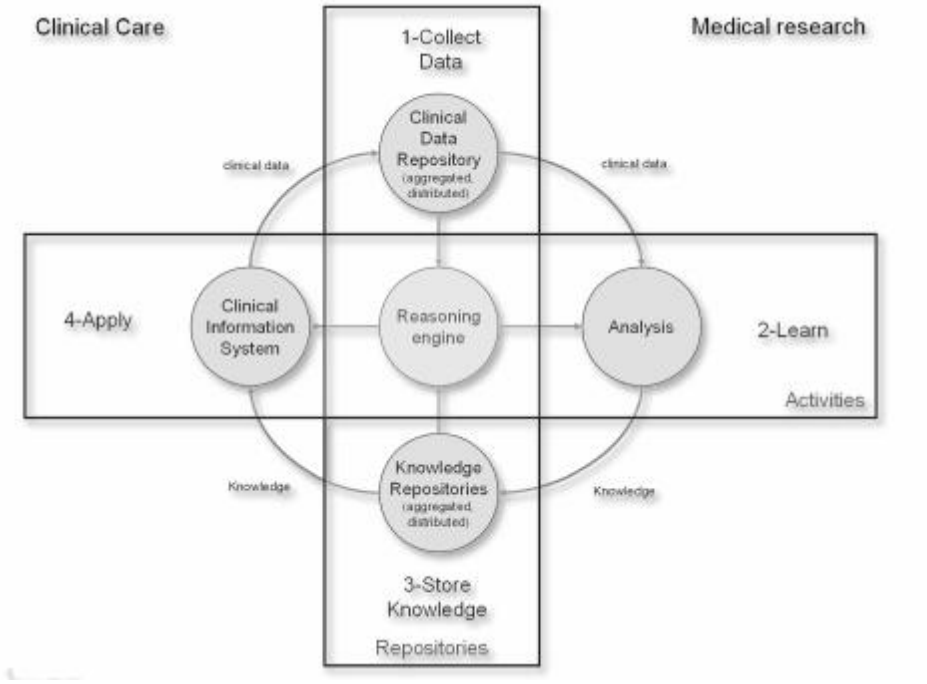

Figure 1. The Conceptual Framework

Apply: Software tools will be integrated into the available clinical and public health information systems. Decision support tools will apply the generated knowledge and help the clinician to provide clinical care (choice, dose and administration of antibiotics for example). The knowledge will also be used to monitor the ongoing care activities and even predict future outcome to give additional feedback, both on individual patient and cohort level. This will allow healthcare providers and decision makers to take appropriate actions at various levels of the healthcare system, including point-of-care, management or policy, and subsequently influence the future development of our health systems. Integration in existing CIS will enable to record activities and results and thus make sure the necessary data are generated for a next cycle.

3. Technology and Architecture

To achieve its goals, the DebugIT project will make extensive use of clinical and operational information originating from running Clinical Information Systems (CIS) across the EU, building a virtualized, fully integrated Clinical Data Repository (CDR).

The CDR will feature transparent access to the original CIS and provide data aggregations in local stores. The CDR is specifically tailored for knowledge discovery, featuring ethically sound, transparent access to data at or from the original CIS and/or collection and aggregation of data in a local data store.

Multimodal Data Mining (MDM) will have a strong focus on new fields of research doing mining on distributed storage, using highly advanced new text, image and structured data mining on individual patients as well as on populations.

New knowledge will be fed into a Medical Knowledge Repository (MKR) and mixed with domain knowledge coming from external sources (guidelines and scientific evidence). Innovative and user friendly knowledge representation paradigms will be developed in order to enable not only knowledge engineers but also clinicians to use the repository

After validating, this knowledge will be used by a decision support module (DSM) and monitoring tool in the clinical environment to prevent patient safety issues and report on them, both at the population and at the patient level for direct care.

Co-ordination and steering of both the analysis and the care process will be done by a performant and versatile reasoning engine.

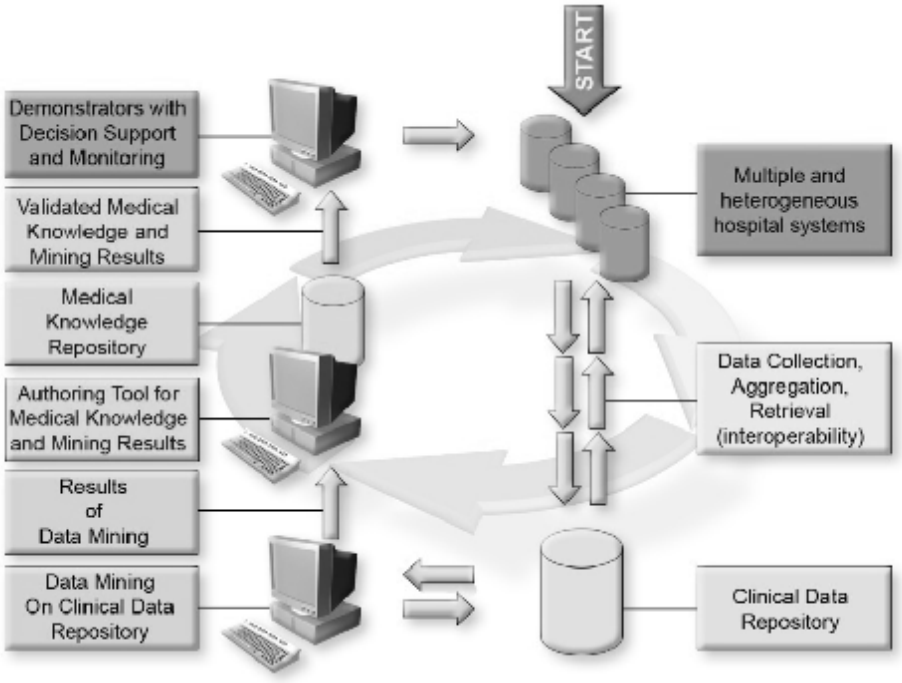

Figure 2: Overview of the project

4. Data integration approaches

There are currently a few different approaches proposed to integrate distributed heterogeneous data sources: link integration, data warehousing, view integration, workflows, and mashups.

According to their architectural model, these approaches can be classified either as centralised or decentralised. Another way to characterize them uses data overlapping. Where there is no data overlapping across the sources they can then be viewed as horizontal and, in the opposite case as vertical. Finally, for any integration approach, the access to the data can be via a push mechanism, where data is pushed into a periodically materialised database, or by a pull method, when a view on data is provided on demand.

a) Data warehousing

Data warehousing is a centralised approach that stores all the data of the sharing sites into one central database. In this approach, a unified data model is defined in order to accommodate all the information contained in the source sites. In addition, wrappers are created for each participant site in order to upload and manage the local data, so that it fits into the central database schema. Then, all the queries can be easily submitted centrally. Some projects using this approach are the Genome Database (IGD) project, ATLAS, BioWarehouse, and BioDWH.

High performance can be considered as one of the highest advantages of this system. In addition, it allows the user to modify and annotate the data since this is a replica, not the production database. There exist also other well-known techniques to build dynamic and online data navigation (OLAP) functionalities upon data warehouses that could serve the users for their data analyses.

This well-known industrial approach to data integration has proven to be efficient for relatively simple industrial contexts. In life science, with its fast evolving knowledge domain as well as the complexity of domain data models, this approach is less widely used due to the cost of data model updates in the central data stores, and to the additional risk of data exposure it leads to.

b) View Integration

View integration is a decentralised approach where source data is kept only in the source site. A single view around all the databases that share information within the system is created. Queries are converted into a common query language that is later handled by a mediator. The mediator identifies the sources that need to be accessed to retrieve the result. Then, it splits the query in many sub queries that are passed to wrappers. Those will transform their sub queries into the local language using mapping rules metadata, and access the data. After the data is fetched from the different sources, the result is globally integrated and returned to the user. HEMSYS [8] and TAMBIS [9] are examples of systems that follow this approach. Ontologies are usually used for view integration to help in aligning different data sources into a global concept of the data. The system can use either integration with a global ontology as in SIMS [10], where each source references the same domain ontology; or local ontology as in KRAFT [11].

When view integration is used, the data is always up to date and there is a relative decoupling from the data sources to the integration system. In the downside, usually the performance is not good since the loading of the data is on-demand and relies on the underlying network performance.

c) Others approaches

In the link integration approach, entries in a data source, usually a web page, are linked to other entries in another data sources via the integration system. SRS, one of the most used integration systems, follows this approach. Entrez and Integr8 are other examples.

Mashups are a Web 2.0 idea beginning to take hold in the life sciences. Mashups provide means to take data from more than one web-based resource and make a new web application. An example is the use of Google Earth to track the global spread of avian flu. Mashup development framework such as Google Mashup Editor, Microsoft's Popfly and Yahoo! Pipes speed up the development of mashups.

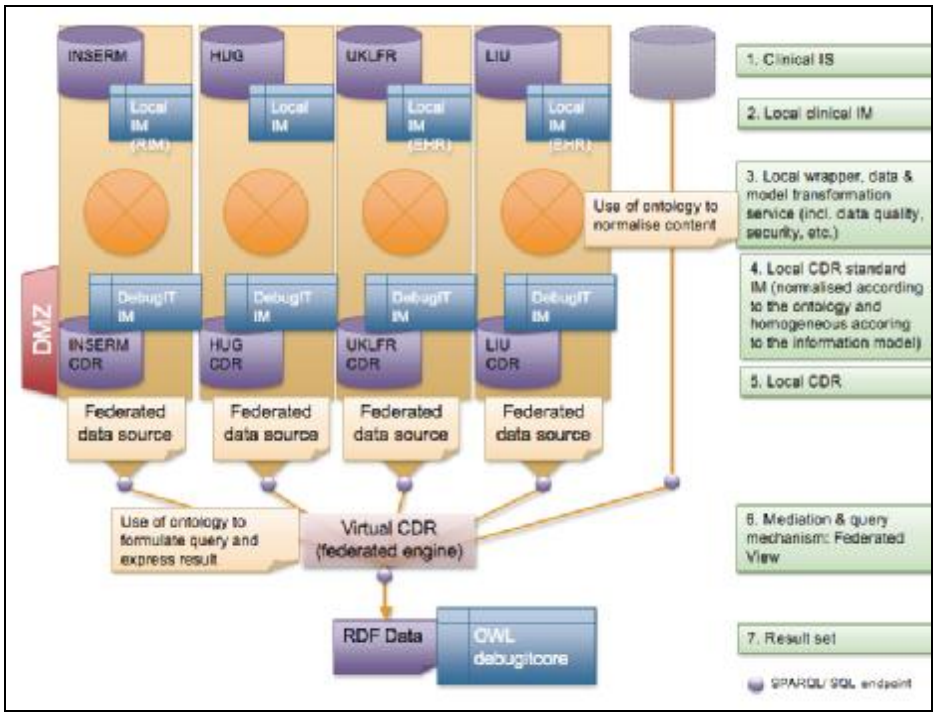

Figure 3. DebugIT CDR architecture.
(acronyms: see table 1)

5. Dealing with rules, information and knowledge

Another very important challenge is the ability to formalize existing rules, guidelines and other types of medical knowledge within a knowledge repository, which will hold new acquired knowledge. A very short example is given hereafter. Manually-generated rules are based on clinical guidelines provided by healthcare institutions. Typical recommendations are useful to prescribe the most appropriate antibiotic, according to different parameters, such as costs, benefits, adverse effects, as well as the risk of resistance development. As case study, we started investigating guidelines for the geriatrics (cf. Table 1) and the surgery services from the University Hospitals of Geneva (HUG).

Table 1. Slightly simplified sample of the guidelines

Pathologies	Pathogenic agents	Antibiotics	Alternatives	Treatment duration
Cholecystitis	Enterobacteriaceae Enterococcus Clostridium sp	ceftriaxone 1 g/24h iv + metronidazole 500 mg/8h iv	amoxi./clav. 1,2 g/8h iv	10-14 days
Gastroenteritis	Campylobacter Salmonella E. Coli Shigella (rare)	ciprofloxacine 500 mg/12h po	co-trimoxazole forte (160mg TMP/ 800mg SMX)/12h po & clarithromycine 500 mg/12h po	5-10 days

To transform such verbose documents into machine-readable data, the guidelines are transformed into databases tuples. The translation from French to English was performed manually, assisted by French-to-English translation tools (e.g. http://eagl.unige.ch/EAGLm/), and a SNOMED categorizer [6]. For some of the queries, several answers were possible – three on average – as shown in Table 2, where *diverticulitis* caused by *enterococcus* can be treated by three different antibiotics: *Ceftriaxone*, *Metronidazole* and *Piperacillin-Tazobactam*; each of them is unambiguously associated to a unique terminological identifier.

Automatically-generated rules are represented as triplets: 1) disease, 2) pathogen and 3) antibiotic. The objective of the automatic rule generation is to generate one of the items (so-called the *target*) of the triplet using the other two items (so-called the *sources*). The discovery of the third item relies on an advanced question-answering engine (EAGLi: Engine for Question Answering in Genomic Literature, http://eagl.unige.ch/EAGLi). Thus, the rule induction problem is reformulated as a question-answering problem, with a unique semantic solution. Three types of questions are designed:

1. Drug: Which *antibiotic* should be used against *this pathogen* causing *this disease*?
2. Pathogen: Which *pathogen* is responsible for *this disease* treated by *this antibiotic*?
3. Disease: Which *disease* is caused by *this pathogen* and treated by *this antibiotic*?

Further, two search engines, corresponding to different search models were tested: easyIR (a relevance-driven search engine well known for outperforming other search methods on MEDLINE search tasks [7] and PubMed (the NCBI's Boolean search instrument). All targets were normalized using standard terminologies: antibiotics and diseases in SNOMED CT or MeSH and bacteria in NEWT. To find the antibiotics, given disease and its pathogen, a list of 72 antibiotic targets was defined, corresponding mostly to the UMLS Semantic Type T195. To find the bacterial pathogens, given the disease and the antibiotic, we suggested a subset of the NEWT terminology corresponding to the bacteria taxonomy. To find pathological processes, knowing the pathogen and the antibiotic, we proposed a list of MeSH terms corresponding to disease, corresponding to the following UMLS Semantic Types T020|T190|T049|T019|T047|T050|T033|T037|T047|T191|T046|T184. For the antibiotics category, we tested the use of both SNOMED CT and the MeSH. Synonyms from these terminologies were also evaluated. Thus, "amoxycillin with clavulanate potassium" can also be mentioned as "amoxicillin-clavulanic acid" or "augmentin".

Furthermore, tuning the question-answering module includes terminology pruning. Indeed, several descriptors, in particular generic ones, need to be removed. Thus, *infectious diseases* or *cross-infection* were removed from the descriptor list for the disease type of target. Finally, specific keywords were used to refine the search equation in order to retrieve more accurate results. Thus, we added context specific descriptors such as *geriatrics*, *elderly*, for geriatric guidelines, etc. The impact of general keywords was also tested such as *recommended antibiotic*, *antibiotherapy*, etc.

6. Privacy

Strong attention is given to privacy concerns, taking into account the various legal and ethical frameworks to be met. Therefore, privacy is made a central part of the project *by design,* using a virtualized data repository without dealing directly with the original data. Identification elements provided by the clinical data repositories can be carried all along the process, blindly, in order to allow the original clinical information system to feed back decision support without need for patient identification.

7. Conclusion

The DebugIT project is focused on using large existing and heterogeneous clinical datasets covering hundreds of thousands of patients from several clinical information systems in different European countries. DebugIt proposes to build an interoperability platform to populate a pertinent dataset about the infectious domain to achieve a very large common shared virtualized clinical repository that enables knowledge-driven data mining. This "semantic mining" will be based on innovative methodologies to deal with the characteristics of real world clinical data. A knowledge repository will drive the data mining and serve as storehouse for the results. Finally, a decision support engine will exploit the aggregated knowledge to loop-back to the real world.

To achieve this system, several aspects will have to reach the frontiers of current state-of-the-art and beyond. Two strategies can be chosen for that. The first one is to invent something radically new. The second one consists of using all existing knowledge and methods, putting them together, and trying to build upon this base. For most of its research, the second strategy is the one chosen in this project, because operational results for clinical information systems must be available and sustain the DebugIT outcomes after the end of the project.

In order to meet these requirements, the project has been organized according to architectural component-based considerations:

- Interoperability Platform (IOP);
- Clinical Data Repository (CDR);
- Multimodal Data Mining (MDM);
- Medical Knowledge Repository and associated Knowledge Authoring Tool (MDR)
- Decision Support and Monitoring engine (DSM);
- Clinical applications.

This scientific and technical framework, associated with access to large amounts of clinical databases and led by experts in the medical field will lead to a serious advance in building a large IT infrastructure aiming at creating new knowledge in the field of monitoring, surveillance and efficient measures to fight infectious diseases.

8. References

[1] IOM. The computer-based Patient record: An essential technology for health care. Institute of Medicine report. 1991, revised 1997.
[2] Medicine Io, editor. To Err is Human. Building a safer Health System; 1999. National Academy Press.

[3] Frazee BW. Update on emerging infections: news from the Centers for Disease Control and Prevention. Severe methicillin-resistant Staphylococcus aureus community-acquired pneumonia associated with influenza--Louisiana and Georgia, December 2006-January 2007. Ann Emerg Med. 2007 Nov;50(5):612-6.
[4] O'Brien J A, Lahue BJ, Caro JJ, Davidson DM. The Emerging Infectious Challenge of Clostridium difficile-Associated Disease in Massachusetts Hospitals: Clinical and Economic Consequences. Infect Control Hosp Epidemiol. 2007 Nov;28(11):1219-27.
[5] McAlonan GM, Lee AM, Cheung V, Cheung C, Tsang KW, Sham PC, et al. Immediate and sustained psychological impact of an emerging infectious disease outbreak on health care workers. Canadian journal of psychiatry. 2007 Apr;52(4):241-7.
[6] Ruch P, Gobeill J, Lovis C, Geissbuhler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak. 2008;8 Suppl 1:S6.
[7] Aronson A DD-FD, Humphrey S, Lin J, Liu H, Ruch P, Ruiz M, Smith L, Tanabe L, Wilbur J. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents TREC National Institute of Standards and Technology. 2005.
[8] S. V. Pillai, R. Gudipati and L. Lilien, Design issues and an architecture for a heterogenous multidatabase system. *Proceedings of the 15th ACM Computer Science Conference (1987).*
[9] C. A. Goble et al., Transparent access to multiple bioinformatics information sources. *IBM SYSTEMS JOURNAL* **40** (2001), 532-551.
[10] The SIMS Corpus Project: *http://groups.ischool.berkeley.edu/SIMSCorpus/results.htm*
[11] A. D. Preece et al., The KRAFT Architecture for Knowledge Fusion and Transformation. *Proceedings of the 19th SGES Int. Conf. on Knowledge-based Systems and Applied Artificial Intelligence* (1999).