Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

Information Structures and Workflows in Health Care Informatics

Ph.D. Thesis
UMINF 10.08

Johan Karlsson
johank@cs.umu.se

Print & Media, Umeå University 2010

UMINF 10.08 ISBN 978-91-7459-026-5 ISSN 0348-0542

To my parents,
Birgitta and Christer

"The nice thing about standards is that there are so many of them to choose from."
(Andrew S. Tanenbaum)

Abstract

Patient data in health care have traditionally been used to support direct patient care. Although there is great potential in combining such data with genetic information from patients to improve diagnosis and therapy decisions (i.e. personalized medicine) and in secondary uses such as data mining, this is complex to realize due to technical, commercial and legal issues related with combining and refining patient data.

Clinical decision support systems (CDSS) are great catalysts for enabling evidence-based medicine in clinical practice. Although patient data can be the base for CDSS logic, it is often scattered among heterogenous data sources (even in different health care centers). Data integration and subsequent data mining must consider codification of patient data with terminology systems in addition to legal and ethical aspects of using such data. Although computerization of the patient record systems has been underway for a long time, some data is still unstructured. Investigation regarding the feasibility of using electronic patient records (EPR) as data sources for data mining is therefore important.

Association rules can be used as a base for CDSS development. Logic representation affect the usability of the systems and the possibility of providing explanations of the generated advice. Several properties of these rules are relatively easy to explain (such as support and confidence), which in itself can improve end-user confidence in advice from CDSS.

Information from information sources other than the EPR can also be important for diagnosis and/or treatment decisions. Drug prescription is a process that is particularly dependent on reliable information regarding, among other things, drug-drug

interactions which can have serious effects. CDSS and other information systems are not useful unless they are available at the time and location of patient care. This motivates using mobile devices for CDSS. Information structures of interactions affect representation in informatics systems. These structures can be represented using a category theory based implementation of rough sets (rough monads).

Development of guidelines and CDSS can be based on existing guidelines with connections to external information systems that validate advice given the particular patient situation (for example, previously prescribed drugs may interact with recommended drugs by CDSS). Rules for CDSS can also be generated directly from patient data but this assumes that such data is structured and representative.

Although there is great potential in CDSS to improve the quality and efficiency of health care, these systems must be properly integrated with existing processes in health care (workflows) and with other information systems. Health care workflows manage physical resources such as patients and doctors and can help to standardize care processes and support management decisions through workflow simulation. Such simulations allow information bottle-necks or insufficient resources (equipment, personnel) to be identified.

As personalized medicine using genetic information of patients become economically feasible, computational requirements increase. In this sense, distributing computations through web services and system-oriented workflows can complement human-oriented workflows. Issues related to dynamic service discovery, semantic annotations of data, service inputs/outputs affect the feasibility of system-oriented workflow construction and sharing. Additionally, sharing of system-oriented workflows increase the possibilities of peer-review and workflow re-usage.

Acknowledgments

First of all, I want to thank my thesis supervisor, Prof. Patrik Eklund, whose patience, guidance and encouragement has been very important. It has been a long road but we reached the goal. I hope this goal, rather than being an end, is the beginning of future collaborations!

I also want to give thanks to my colleges at the Department of Computing Science. In particular, Pedher has been a good friend and provided a lot of support. He has also helped to check the sample printing, thanks again! Inger and Anne-lie, of course, are doing a wonderful job. Thank you for helping me with so many things that I couldn't do myself from Spain.

Some of the papers in this thesis would not have been possible without my (unofficial) co-supervisor, Associate Professor Oswaldo Trelles at Malaga University in Spain. Thank you for introducing me into the world of bioinformatics and for encouraging me to finish my PhD studies.

Since moving to Spain, my colleagues in the BITLAB team at Malaga University have provided a great and friendly work environment. Thank you for putting up with my English (and I promise to practice Spanish more...).

I am grateful to all those who have read and commented drafts of this thesis. You have helped to find many strange sentences.

To my family-in-law, thanks for all the support and for making it easier for me to live in Spain.

My parents and brother have given me support throughout the work of this thesis. Although we are separated by many kilometers, you are always in my thoughts. Of

course, my parents were the ones that inspired me to study at the university.

Thanks to my wonderful wife María Ángeles. We met at the department in Sweden and now we have a life together in Spain! My dear daughter, Andrea, has (repeatedly) tried to contribute to this thesis by hitting the keys of the keyboard. Hopefully, your "input" was not included in the final draft... On the other hand, you have been the joy and light of my life ever since you were born!

Umeå, June 2010

Johan

Table of Contents

Preface

This is a compilation thesis that consists of a summary (Part I and II) and the following papers (in Part III):

Paper I P. Eklund, J. Karlsson, J. Rauch, and M. Simunek. *On the Logic of Medical Decision Support.* In Harrie C. M. de Swart, Ewa Orlowska, Gunther Schmidt, and Marc Roubens, editors, Theory and Applications of Relational Structures as Knowledge Instruments, volume 4342 of Lecture Notes in Computer Science, pages 50-59. Springer, 2006.

Paper II P. Eklund, J. Karlsson, and A. Näslund. *Mobile Pharmacology.* In Marcin S. Szczuka, Daniel Howard, Dominik Slezak, Haeng-Kon Kim, Tai-Hoon Kim, Il Seok Ko, Geuk Lee, and Peter M. A. Sloot, editors, ICHIT, volume 4413 of Lecture Notes in Computer Science, pages 522-533. Springer, 2006.

Paper III P. Eklund, S. Eriksson, J. Karlsson, H. Lindgren, and A. Näslund. *Software development and maintenance strategies for guideline implementation.* In Proc. EUNITE-Workshop "Intelligent Systems in Patient Care", pages 26-45. Austrian Computer Society, 2000.

Paper IV J. Ríos, J. Karlsson, and O. Trelles. *Magallanes: a web services discovery and automatic workflow composition tool.* BMC Bioinformatics, 10(1):334, 2009.

Paper V J. Karlsson, S. Ramirez, J.F. Aldana-Montes, and O. Trelles. *Workflow*

repositories for bioinformatics. Technical report, Department of Computer Architecture, Malaga University, 2008.

Paper VI P. Eklund, M.A. Galán, and J. Karlsson. *Rough Monadic Interpretations of Pharmacologic Information*. In ICCS 2007: Proceedings of the 15th International Workshops on Conceptual Structures, pages 108-113. Springer, 2007.

Paper VII P. Eklund, M.A. Galán, and J. Karlsson. Rough Set Theory: A True Landmark in Data Analysis, chapter *Categorical Innovations for Rough Sets*, pages 45-69. Springer, 2009.

In addition, the following publications have been made in topics related to the thesis. Some results from these publications are described in the thesis summary to provide background and context.

1. J. Karlsson. *A generic system for developing medical decision support*. Master's thesis, Department of Computing Science, Umeå University, 1998.

2. J. Karlsson, P. Eklund, C.-G. Hallgren, and J.-G. Sjödin. *Data warehousing as a basis for web-based documentation of data mining and analysis*. In P. Kokol, B. Zupan, J. Stare, M. Premik, and R. Engelbrecht, editors, Proc. Medical Informatics Europe '99, pages 423-427. IOS Press, 1999.

3. J. Karlsson and P. Eklund. *Data mining and structuring of executable data analysis reports: Guideline development in a narrow sense*. In A. Hasman, B. Blobel, J. Dudeck, R. Engelbrecht, G. Gell, and H.U. Prokosch, editors, Proc. Medical Informatics Europe '00, pages 790-794. IOS Press, 2000.

4. J. Karlsson and P. Eklund. *Workflow Design as a Basis for Component Interaction*. In Proc. Medinfo 2001, pages 1158-1160, 2001.

5. J. Karlsson and P. Eklund. *A one-step approach to data retrieval, analysis and documentation*. In G. Surján, R. Engelbrecht, and P. Mc-Nair, editors, Proc. Medical Informatics Europe '02, pages 320-324. IOS Press, 2002.

6. P. Eklund and J. Karlsson. *Simulations of workflows in radiation therapy.* In G. Surján, R. Engelbrecht, and P. McNair, editors, Proc. Medical Informatics Europe '03, 2003.

7. P. Eklund, J. Karlsson, J. Rauch, and M. Simunek. *Computational Coronary Artery Bypass Grafting.* In ICCIMA '05: Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications, pages 138-144, Washington, DC, USA, 2005. IEEE Computer Society.

8. S. Ramirez, J. Karlsson, JM. Rodríguez, R. Royo, and O. Trelles. *Mirroring BioMOBY services.* In VII Spanish Bioinformatics Symposium, 2006.

9. R. Fernández-Santa-Cruz, J. Karlsson, J. F. Aldana-Montes, and O. Trelles. *Workflow execution in the inb platform.* In VII Spanish Bioinformatics Symposium, 2006.

10. J. Karlsson. *Interface for Accessing Pharmacological Information.* In CBMS '08: Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, pages 176-178, 2008. IEEE Computer Society.

11. S. Ramírez, J. Karlsson, M. García, and O. Trelles. *Metadata repositories for web-services and workflows.* In VIII Jornadas de Bioinformática, pages 50-56, 2008.

12. S. Ramirez, J. Karlsson, M. García, J. Rios-Perez, and O. Trelles. *A flexible framework for the design of knowledge-discovery clients.* In Proceedings of the International Conference on Telecommunications and Multimedia (TEMU 2008), 2008.

13. The BioMoby Consortium (including J. Karlsson). *Interoperability with Moby 1.0- It's better than sharing your toothbrush!* Brief. in Bioinformatics, 9(3):220-231, 2008.

14. J. Rios-Perez, J. Karlsson, and O. Trelles. *Victoria: Navigating to a new style of searching for web-services and workflows*. In 17th Annual International Conference on intelligent systems for molecular biology (ISMB) and 8th European Conference on Computational Biology (ECCB), 2009.

15. M. García, J. Karlsson, and O. Trelles. *Web-services across an european biomedical grid infrastructure*. In Book of abstracts. IX Jornadas de Bioinformática. 9th symposium on Bioinformatics and Computational Biology, page 88, 2009.

16. P. Eklund, M. Johansson, J. Karlsson, and R. Åström. *BPMN and Its Semantics for Information Management in Emergency Care*. In Sungwon Sohn, Kae Dal Kwack, Kyhyun Um, Gye Young Lee, and Franz Ko, editors, Proceedings of Fourth International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2009), pages 273-279. IEEE Computer Society, November 2009.

17. M. Garcia, J. Karlsson, O. Trelles. *Web service catalogue for Biomedical GRID infrastructure*, HealthGrid Conference 2010, accepted.

Information Structures and Workflows in Health Care Informatics

Part I

The first part of this thesis summary contains an overview of the area and also some short case-studies.

Chapter 1

Introduction

Health care increasingly needs computerized support to enable practice of evidence-based medicine. Evidence-based medicine [27, 28] means that care givers should base their diagnosis and treatment decisions on (where available) empirical evidence. Electronic patient records (EPR) provide high availability of patient data and integration with clinical decision support systems (CDSS) which can be based on well-established clinical guidelines. Secondary uses of patient data, such as data mining, have the potential to increase medical knowledge. However, several aspects of health care data and processes should be considered as they contribute to the success or failure of introducing software systems in health care practice.

1.1 Health Care Informatics

The reason why the field of medicine needs a dedicated informatics science is based on aspects of medical data structures (Chapter 2) and health care processes (Chapter 3).

The scientific domain of health care informatics (or medical informatics) has been described as compromising "*...the theoretical and practical aspects of information processing and communication based on knowledge and experience derived from processes*

in medicine and health care" [13, 155]. The importance of ontologies in medical informatics was stressed in [91]: "*...formal ontologies are at the core of work in medical informatics– and have had that role for more than one hundred years*".

In this thesis, we adopt the view of [93], where medical informatics is looked upon as "*the field of information science concerned with the analysis and dissemination of medical data through the application of computers to various aspects of health care and medicine*".

Medical informatics is complex because many and diverse skills are needed. Development of CDSS requires both domain knowledge (i.e. medical training) and knowledge of software development (i.e computing science training). Additionally, human-computer issues and organizational concerns (workflow) must be taken into account. Workflow studies (Chapter 3) can influence the software design and result in a successful integration within the work processes of a clinic. Such assorted requirements in skills are in no way unique for clinical software development but particularly critical because of the complexity of the domain.

1.2 Bioinformatics

The field of (molecular) bioinformatics was described in 2001 as "*... conceptualising biology in terms of molecules (in the sense of Physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied maths, computer science and statistics) to understand and organise the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications*" [81].

Since this description was made, data sizes have grown considerably (see Section 2.2). Many of the well-established techniques of analysis in bioinformatics must accordingly be re-evaluated and improved to take advantage of high-performance computing techniques and centers.

1.3 Organization and limitation of scope

Work related to this thesis can be divided in several stages. The initial stage consisted of a study of issues in CDSS development, including human-oriented workflows and resulted in Papers I-III. Application of service-oriented architectures to bioinformatics as a base for system-oriented workflow discovery, composition and storage (Papers IV-V), represent the current stage. Initial (and on-going) efforts were also made towards the representation of drug-drug interactions (Papers VI-VII) using an abstract mathematical theory (category theory).

Thesis contributions can be found in medical informatics and bioinformatics by means of the application of various methods from computer science to health care data and processes. This thesis summary does *not* attempt to present a complete overview of all possible solutions and techniques. Instead, it attempts to introduce health care informatics as it relates to the Papers I-VII, describe selected methods from computer science (without going into details) and provide representative examples of similar work in health care informatics.

In some cases, Papers I-VII use methods which should not be considered part of their results. To clarify which are results from the point of view of this thesis, we describe these methods throughout Part I instead of in Chapter 4. The chapters in Part I also describe some case-studies which are based on the additional papers (see the preface). Although these papers are not part of the thesis, they provide important background to understand Papers I-VII.

1.4 Thesis overview

The papers included in this thesis are described in Part II. In this section, we describe the thesis on a high level and include some additional papers as references.

1.4.1 Data storage in electronic patient records

An initial study [71] of patient data retrieval was done in the urology clinic of Umeå university hospital. We collected patient data related to operations of patients with lower urinary tract symptoms (LUTS) to investigate which difference, if any, introducing an *electronic patient record* (EPR) in the clinic had for the possibilities of information retrieval which naturally requires support in the EPR. The requirements of an architecture for a data warehouse were discussed in the paper. In collaboration with domain experts at the clinic, we selected a set of patient data from the period when the EPR was introduced; data was collected from six months before and six months after. The data before the EPR introduction was collected manually from paper forms. The data after the EPR introduction was collected using the EPR software. The lessons learned was that, while having much higher availability, the EPR did not influence to a high degree the way the doctors documented their patient encounters which still appeared as narrative text. Later studies showed that data entry has improved and that it was possible, to a high extent, to automatically retrieve LUTS data.

The study in [68] discusses some requirements of data warehouses when used in data mining efforts and decision support development for health care (see also Section 2.4.2). EPRs are naturally designed for retrieval of patient-specific data and may not always structure data in a way that is sufficient for comparison. This raises concerns regarding the suitability of EPRs as data sources for DWs.

Issues related to decision support development in healthcare was discussed in [70]. Software acting as a layer between the EPR databases and the client applications could facilitate data extraction by adding metadata, such as semantic descriptions (for example, the value is a code from ICD-10 [149]) and context (the value was recorded as part of a specific workflow enactment).

1.4.2 Logic in clinical guidelines

Medical decision support can be based on rules from clinical guidelines (see Section 3.2.1). However, the logic used in guidelines is not always clearly defined in a way which enables it to be used when implementing CDSS. The study in Paper I discusses aspects of CDSS development from the point of view of logic. Selection of appropriate logic for particular guideline implementation is discussed together with other aspects of integrating decision support in clinical practice. CDSS logic and rules can also be inferred from patient data. Such rule inference is facilitated in the local care region of northern Sweden due to the use of a single EPR system.

Paper I describes a previously published method of data mining called GUHA [54, 55]. Using a data set related to coronary artery bypass grafting (CABG), we analyzed the data and obtained several rules (see also [42]). Rules produced by approaches like GUHA are suitable for developing CDSS since these rules are possible to motivate to doctors.

1.4.3 Human-oriented workflows in health care

Health care processes of a radiology clinic were studied [69] in order to better understand issues of health care information interchange. The clinic was considered suitable for such a study since it uses many and, to some extent, isolated information systems. An existing CASE-based workflow was used as a base during the study.

In a subsequent paper [40], a publicly available petri net design and simulation tool was used to represent radiology care processes as petri nets. The motivation was to study to which degree workflow simulation could be performed based on the available workflow. Although representation of the CASE-based processes as petri nets was time-consuming, such simulations could give important insight into the organizational and technical problems faced by the health care institution.

A regional guideline for emergency care was studied in [39]. Emergency and crisis response to large traffic accident management involves coordination of emergency rescue services and police. The management of information is a complex task where

decision support on site and distance management requires a well-founded understanding of the underlying processes. National guidelines for emergency care exist but are shallow and provide only overviews, typically with no specific information for on-site emergency care. Well-developed regional guidelines, such as [141] are fully adequate but are mostly insufficient for information and decision support system development.

The emergency care guidelines cover e.g. respiratory syndromes, circulatory failure, trauma and prenatal situations. The focus in [39] was on trauma which typically may involve skull fracture and spinal cord injury. It could also involve facial, thoracic and/or abdominal injuries, or bleeding, hypothermia and burn wounds. The guideline for trauma was described in detail and represented using the workflow standard Business Process Modeling Notation (BPMN). We also discussed syntactic and semantic aspects of BPMN.

1.4.4 System-oriented workflows in bioinformatics

Many analysis tools in bioinformatics are available as web-services. Efforts to standardize interfaces of these web-services have resulted in a trend from web-page based analysis to using web-services. The motivation for distribution of tools is that many tasks require substantial computational resources and/or access to large and frequently updated databases. It is quite common that the results from one web-service invocation must be further analyzed using other web-services. The standard BioMOBY [135] aims to simplify this task by defining a shared datatype ontology and a standardized web-service protocol. The web-service protocol was improved to increase availability of web-services by redundant service instances [113] and specification of a protocol for asynchronous web-service invocation [15] (see also Section 3.3.3) which enables long-running analysis. Software for sharing web-service metadata in a metadata repository was described in [114, 112, 49].

Combinations of web-services (web-service compositions or workflows) to solve high-level tasks are often re-usable by other researchers. Workflow metadata can

be shared through workflow repositories (Paper V). The paper describes a workflow repository which is integrated with the system from [94]. Workflows in bioinformatics can be used to solve complex problems. Bioinformatics processes were considered more suitable than health care processes since many bioinformatics workflows are purely system-oriented and because it is complicated for legal and ethical reasons to use external web-services to analyze patient-data.

Paper V discusses the workflow paradigm and requirements for sharing workflows in repositories. This enables not only re-use of typical service compositions but also peer-review of steps taken for data analysis. The paper exemplifies functionality with a demonstration of how to use web-services and the workflow repository to establish evolutionary relationships between species (phylogenetic trees). Several steps are described, for example uploading and documenting of workflows and enacting workflows. The paper is concluded with discussions of unsolved issues related to workflow discovery (see Section 3.3.1) and quality assurance.

Paper IV (and also [116]) describe a system (Magallanes) for service discovery and composition. It is able to use metadata from various repositories of service metadata, including the BioMOBY service repository [145]. Users specify a query (for example a word) which is matched against service and datatype descriptions. These matches can be exact or approximated using Levenshtein's text distance. If a search term does not exactly match any service, the discovery module automatically suggests other search terms to the user. Search results are ranked not only after their degree of matching with the search term but also based on previous user selections.

The system can also generate service compositions (workflows) based on user specifications of input and output data types. This is possible due to the shared data type ontology of BioMOBY. In the case that several compositions are possible, users may select the path which best fits the intended task. The compositions can be exported to Taverna [99] for further editing and enactment.

1.4.5 Pharmacology information systems

The need of interaction between clinical information systems was motivated by the study of a previously developed decision support system [106]. While this system was able to provide guideline-sanctioned pharmacological treatment advice for hypertension, the system could not interact with pharmacological databases. Such a connection was necessary to provide users with additional information regarding suggested drugs (thereby aiding the decision process) but also to allow potentially automatic checks for drug-drug interactions between the recommended drug and earlier prescribed drugs.

Decision support tools and information systems are not useful unless they are available at the time and point of patient care. Mobile devices have the potential to provide access to information regardless of where patient care takes place. The area of pharmacological decision support and information was selected and several prototypes of pharmacological information systems were developed for mobile phones (Paper III) and hand-held devices (Paper II). In a follow-up paper [67], an API for accessing a national database for pharmacological information was developed together with a prototype for mobile phones.

Various forms of CDSS are possible (Paper III). We exemplify and discuss three kinds of systems: a system for pharmacological hypertension treatment [106] based on a clinical guideline, a prototype for dementia diagnosis and finally a drug information system for mobile devices.

The structure of drug interaction information was studied from a rough-set point of view in Papers VI and VII. We showed that such interactions can be represented using rough sets and in an abstract mathematical area (category theory). However, extensions of these initial representations should be developed in future work.

Chapter 2

Information structures

> "I am fain to sum up with an urgent appeal for adopting ... some uniform system of publishing the statistical records of hospitals. There is a growing conviction that in all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life ... In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison ... If wisely used, these improved statistics would tell us more of the relative value of particular operations and modes of treatment than we have means of ascertaining at present"
>
> *Florence Nightingale*, Notes on Hospitals. Green, Roberts, Longman, and Green, 1863

What makes the information in medicine different from information in economics or physics? In other words, why is a dedicated science discipline (medical informatics) needed? The amount of information is certainly not particularly high in medicine. Databases in areas such as business and nuclear physics can reach sizes of several terabytes. Several aspects of medical data makes handling challenging [22]. Data in bioinformatics are also reaching extreme sizes (see Section 2.2). This chapter is dedicated to information structures in health care and bioinformatics.

2.1 Medical data

Information is collected using dissimilar collection and measurement methods (such as physical examinations, lab tests, radiology tests, patient interviews). Measurement equipment have different accuracy ranges and doctors have different education and experience levels. Values can be difficult to compare between patients and measurement opportunities. For example, ascertaining blood pressure measurements can be very difficult due to errors in measurement, variations in blood pressure, stress levels of patients etc. Information coming from patient interviews is subjective and its usability largely depends on the ability of physicians to ask correct questions, interpret the answers and later codify the information correctly.

Routines for entering patient data in EPRs might also affect the coverage and consistency of the data sources. For example, in [71], lab results and measurements were stored as free text notes in the electronic patient record (EPR), even though the system provided specific functionality to enter such data in a structured and codified way. This problem can only be addressed by better software which permits effective and suitable structured data entry and stricter quality control.

Nontheless, in general, the deployment of information systems in health care affects positively the quality of patient data, in particular for drug prescription data [136]. The motivations for EPRs are clear; faster and more efficient access to patient information compared to paper-based patient records. While advanced decision support systems can be used without an EPR, direct access to patient data makes the decision task more efficient and less error-prone [140, 80]. Other sources of medical data are radiology, ultra-sound and digital endoscopy (images/video) equipments.

2.2 Molecular biology data

The expression “information overflow” has historically referred to human incapability of integrating and interpreting increasingly huge amounts of information. In a special issue of the journal Science [31], we can see that this expression now also is

valid for automated processing by computers. In bioinformatics, technological breakthroughs are producing unprecedented "mountains" of data. Genomics sciences are experiencing a paradigm change where comparison of whole genomes is a real possibility. The medical applications are also important since these new technologies enable cost-effective personalized medicine as the cost of genome sequencing continue to decrease.

Sequencing of the human genome with Sanger technology was relatively slow and expensive [142]. Several breakthroughs have resulted in very cost-effective Next-Generation Sequencing (NGS) technologies [124], such as 454 sequencing from Roche, Solexa from Illumina and SOLiD from Applied Biosystems. Selection of NGS techniques is influenced by the intended goal, for example genome assembly, transcriptomic analysis, single nucleotide polymorphism (SNP) or copy number variations (CNV). In all those cases, the amount of data to process is considerable. Cost of data production is also decreasing. For example, the National Institutes of Health (NIH) has announced funding for projects with the aim of completely sequencing a human genome for less than 1000 US dollars.

Consequently, genome sequencing of patients, as a base for personalized medicine, will likely become economically viable and common in the near future. However, full genome sequencing is not always necessary. In specific cases, it is only necessary to study areas of the genome known to be associated to specific conditions. Nevertheless, genome-disease associations must be previously known [146] for such sequencing to influence determination of diagnosis or therapy. This is one of the motivations for the 1000 Genomes Project [130] which aims to obtain complete genome sequences for (at least) 1000 humans in only three years. The goal of the project is to discover genetic factors and variations related to diseases in humans.

Other areas in bioinformatics are also experiencing dramatically growing data sizes as a result of technology advances, for example in DNA microarrays [58].

Taking into account such data production, it is also essential to consider data quality aspects. Unfortunately, data curation lags behind the rapid generation of data

[59]. The field of biocuration aims to organize, represent and make available biological data to researchers. Tasks such as data representation, gene annotation in large databases and tagging research papers with concepts from controlled vocabularies are considered crucial for future research.

2.3 Dealing with sensitive patient data

Patient data is sensitive information and is, therefore, protected by privacy laws. For example, there exists a risk that medical insurance companies might deny coverage based on the medical history of patients. With the increased citizen mobility in the European Union, interchange of medical history of patients becomes a necessity. However, legislation concerning handling of medical data make it difficult, or even impossible, to attain necessary permissions to use data across country borders.

Many hospital information systems have traditionally been constructed as "islands of data" [66]. To some extent, this is a direct effect of health care organizations policies to build a secure "wall" around their information systems. Sharing of patient data with other care-givers outside of the organization is thereby limited although such sharing can be motivated both ethically and medically.

There are similar ethical concerns for molecular biology data since genetic data from patients can be used to identify a person. Additionally, genetic data can also be used to determine probability for diseases and even ethnical origins. Hence, any research involving genetic data must also consider ethical and legal issues. Ideally, even if legislation does not require it, informed and freely given consent by patients should be given prior to using their genetic data. Informed consent should also be possible to be retracted by patients at any time [23].

However, it is not always necessary to know the identity of the person from which data comes from. For example, statistical analysis can be made on the genetic data to find patterns without the need to associate data with specific patients. Various procedures of anonymization then become possible. In [8], legal aspects of genetic

data and different levels of anonymity are discussed; *truly anonymous* data which is impossible for anyone to associate to persons and *pseudonymized* data where information which can be used for personal identification has been replaced with a label. The association between labels and personal information is only possible if the researcher has access to a mapping file. If the researcher does not have such access, the genetic data can be considered anonymized from the point of view of the researcher. For example, in Austria researchers may use and transfer sensitive data if it has been pseudonymized and if the scientific research or statistical study does not aim to produce results directly for individual patients. If not, additional restrictions for data usage are applied. Other degrees of anonymization are also possible, including *de-facto anonymized data* which can only be turned into personal data with considerable time, expense and labor.

In [23], the architecture for data protection in the European Union co-founded project Advancing Clinico Genomic Trials (ACGT) [138] is described. The high-level objective of the ACGT project is the development of methods and software systems which can improve the understanding of cancer research data through integration and analysis of biomedical information. The ultimate goal of the project is to enable faster and more accurate diagnosis and treatments on an individual basis for cancer patients. This is accomplished by the development of an integrated and Grid-compatible software platform which can support post-genomic, multi-centre clinical trials. Management of patient data in clinical trials is organized with Virtual Organizations and data access is based on credentials.

The approach in ACGT is to pseudonymize sensitive data and to establish an authority for data protection which enters into legally binding agreements with project participants to ensure data protection. This authority is also responsible for taking into account country-specific variations in legislation. This is necessary since the current European legislation permits the member states to implement exceptions in patient data protection (which vary between states). The agency for data protection also establishes a trusted third party which holds software tools and cryptographic

keys, which can (when authorized) be used to de-anonymize data as required (for example, if a research result has clear relevance to an individual patient).

2.4 Integration of heterogenous medical data sources

One early effort to standardize patient record structures is the PRA document architecture [33] which suggested to use the information model in Health Level 7 (HL7) [56] as a reference to exchange patient data records. Despite such efforts, the multitude of sources and structures of data related to patient care remain great challenges for data integration and communication. However, integration and communication are fundamental to enable secondary usages of patient data, such as discovery of clinically interesting patterns in large data-sets or integration with biomolecular data (enabling personalized medicine).

This is the foundation for the development of clinical guidelines and CDSS (see Section 3.2.1). There are, of course, many approaches to data integration but this thesis summary only highlights two approaches which have been applied previously in medical informatics; data warehousing (see Section 2.4.2) and mediation-based approaches (see Section 2.4.3). For further details on the methods other than the brief summaries of this section, please see the provided references.

2.4.1 Terminology systems in medicine

Codification of medical data is important for efficient information interchange. Such codification assigns terms from terminology systems to patient data. A terminology system assigns terms to concepts or objects within a certain domain [29]. Patient data can thereby be understood with considerably less ambiguity, assuming that the terminology system accurately represents the domain both in scope and granularity. A multitude of terminology standards in medicine are available, see [20, 21] for

overviews.

Relations between concepts in a terminology system can imply additional information. Two types of hierarchical relations can be differentiated. Generic relations arrange terms in the sense that the species possesses all features of the genus with at least one other feature added, for example "`Hepatitis B`" (species) is a "`viral hepatitis''` (genus). Another kind of hierarchical relation is the partitive relation which is the relation between the whole and the parts, for example "`heart`" (part) and "`vascular system`" (whole). There are also non-hierarchical relations such as relations between concepts based on time, space and cause. Examples include "`HIV-virus`" causes "`AIDS`" and "`fever`" before "`spasms`". Especially time relations are important in medicine [90, 123, 12, 75] since they enable analysis of time-series and thereby trends.

The codification process is affected by the choice of terminology systems. Selection of standards is consequently important. Several aspects are important for terminology systems [140, 100]:

- Domain completeness
- Non-overlapping classes (clear criteria for class boundaries)
- Suitable for its purpose
- Homogeneous ordering (one principle per level)
- Usage guidelines
- Appropriate level of detail

Several types of terminology systems can be differentiated [140]. A *thesaurus* is a terminology with index and possibly synonyms and preferred terms. A *classification* is a terminology system in which concepts are related generically. A *vocabulary* is a terminology together with definitions of terms. A *nomenclature* is a terminology with rules to compose new concepts in a certain domain. A *coding system* is a terminology

with a coding scheme. An *ontology* is a set of concepts together with their attributes and relations to other concepts.

2.4.2 Medical data warehousing

One approach to data integration is collecting and storing data in a *data warehouse* (DW). A DW is a "*subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decisions*" [61]. DWs are suitable for data-mining since data in a DW is non-volatile and analysis is therefore repeatable. Data is loaded, transformed, cleaned and regularly updated from the data sources to the data warehouse [62]. Potential conflicts and inconsistencies between the data sources must also be resolved. Differences in naming, structures, semantics and implied knowledge in the data sources could complicate integration [133]. Online analytical processing (OLAP) tools are used for multidimensional analysis in DWs. OLAP permits analysts to successively focus on lower levels of detail.

DWs are also better suited than EPRs for analysis of patient data since the database structures of EPRs are designed for efficient access to data on a patient-specific basis. This causes efficiency problems when more general queries are needed. Typically, normalized data are stored in tables to save both space and ensure consistency. Since DWs are intended for data mining, they can also achieve better efficiency by re-organizing data and reducing the necessary table conjoinings during query processing. Depending on the level of structure and quality of patient data coming from data sources, data imported into the DW might need some "scrubbing" (for example by codifying existing data).

Data in a DW are collected from multiple data sources and transformed to use common data formats and terminology. This is a complicated task in medical DWs, as it involves collection and integration of data from various health care information systems (for example EPRs, nation-wide clinical registers and administrative databases containing patient billing information). The medical history of a patient can be of great importance in data analysis. If such patient data is stored in external

information systems (for example, when a patient has moved and changed hospital), it becomes necessary to identify the patient across data sources (using social security numbers or other strategies) and ensure that information is codified using the same terminology system.

In [134], a methodology is proposed for establishing DWs in medicine. A DW to support long-term care of elderly patients is used to exemplify the methodology. Decisions have to be made regarding intervals of data updates and protocols. It is important to establish trust among the owners of the data sources that patient data privacy is a high priority. Data is anonymized before being exported to the DW but each data source owner could, if needed, use a mapping between social security numbers of the patients and the internal identifiers in the DW system to identify patients. Reported problems were related to the quality of the data which resulted in the development of complex filtering procedures (in some cases even manual inspection was needed). The use and development of ontologies could be seen both as a requirement and as an important result from data integration since it motivates stake-holders (data source owners) to use common terminology for data.

The paper in [154] describes a DW established to support research in traditional Chinese medicine. Their process is simplified compared to [134] due to the use of a standardized and structured clinical record as a data source. Data is also anonymized before importing into the DW. The patient record used as data source is complex and to simplify analysis, the patient record is mapped to a clinical reference implementation model which is later used in the DW. Differences in terminology system usages were resolved during data transformation and pre-processing.

The significance of DWs for evidence-based medicine is discussed in [132]. Generation of logic for guidelines is suggested to be based directly on data mining of clinical data. Several case-studies are used as a motivation for important tasks that can be supported by DWs with regards to evidence-based medicine. Rules in guidelines can be continuously refined as additional data is included in the data warehouse. DWs can also support the development of clinical pathways (i.e. human-oriented processes,

see Section 3.1) through optimization of resources in a clinical environment based on past patient data, for example average hospital stay for recovery after operations.

Various aspects of data collection are discussed in [34], where a clinical data repository at the university of Virginia is reported. The data repository described supports similar tasks as the DWs in [154, 132, 134] but few details are given regarding data importing and scrubbing. Interestingly, this data repository includes (encrypted) patient identifiers; in effect, potentially identifying values were encrypted instead of being removed prior to importation. Such data can be decrypted but require permission from a committee. Few details about any schema transformations were given, suggesting that the data repository is a federated database [125] which is separated from the production systems and updated on regular basis. The main usage of the data repository is to generate administrative reports although some examples of medical queries are given.

The software infrastructure in caGrid [101] is designed to facilitate biomedical research. Its components support publishing, discovering and managing metadata for data sources and analytical tools. The infrastructure aims to facilitate, in particular, cancer research. caGrid extends existing Grid computing technologies and infrastructure with programmatic support for integrating structure and semantic aspects of biomedical data. Grid services can be discovered based on complex queries including input/output data semantics and other characteristics (statistics related to availability of services etc.). Service compositions can be constructed and enacted as workflows. Access to services is restricted according to grid credentials and trust levels. If a security breach is discovered in any service, the system will ensure that clients no longer communicate with that service.

While there are different levels of integration in caGrid, ideally services and data should be made interoperable by using common vocabularies to facilitate data integration, including annotation of data elements (meaning), value domains (which values an element can have), relations to other concepts and syntactic integration by sharing XML schemas. Services should also be annotated with metadata such as

input and output data types and under which authority (organization) the service is provided. Service metadata is registered and shared in a repository component.

Queries to caGrid data-sources are expressed in an object-oriented query language called CQL. The typical data source in caGrid is described using UML. These descriptions can be used with a development toolkit to produce a data service which is able to accept queries expressed in CQL. Data services translate CQL queries into database queries.

Combination of data sources was previously possible in caGrid but was considered a client task and there was no direct support in the architecture. In [85], extraction, transformation and loading (ETL) of data from heterogenous data sources in caGrid is facilitated with semantic information. The data sources initially used were caTissue with information about biospecimens and clinical data (patients, diagnosis etc.) and caArray with microarray and gene-expression data. Briefly, the ETL process was based on ontologies which were generated using descriptions of the data sources in UML and from a terminology with cancer concepts. The ontologies were loaded into a semantic triple store. Data was extracted from the two data sources and linked with a custom inference rule. The data was then loaded into a "semantic" data warehouse. Their approach is demonstrated by querying data from the datawarehouse and performing a Principal Components Analysis. Although the work reported in [85] is an early prototype, it shows the utility of using ontologies to improve data integration.

2.4.3 Mediator-based methods

The establishment of data warehouses is not always an optimal solution. For example, the enormous data sizes being produced from NGS projects (see Section 2.2) are not easily transferred. Accordingly, rather than collecting and storing data in a centralized repository, an alternative approach is to leave the data at the original data sources and instead dynamically combine data through queries resolved by a *mediator*.

A mediator-based system in data integration performs dynamic translation of user

queries between the schema of a mediator component and schemas of data sources. Data is not re-organized and expressed in common terminology system. Instead, mappings between a common mediator schema and data source schemas are established. Early bioinformatics systems for data integration consisted of federated databases where databases with different schemas can be queried as if they have a common schema. Mediator-based solutions can be seen as loosely coupled versions of federated databases [57].

In Global-As-View (GAV), a global schema is established with direct mappings to the local schemas. Adding new data sources requires a modification of the global schema and the mappings. The Local-As-View (LAV) approach is scalable and easier to maintain than GAV since the global schema is created independently of the local schemas. The schemas of new data sources are described in terms of the global schema (which thereby changes less frequently). LAV usually suffers from low query performance for very complex queries compared to GAV.

The ACGT project [138] (see also Section 2.3) highlights the complexity of integrating patient and genomics data from multiple centers and even countries. Data integration is mediator-based where a master ontology describes the domain concepts. Patient data is anonymized and stored in (potentially) heterogenous databases. Data-sources are represented by a source description consisting of a local ontology with mappings to the local data source schema together with references to concepts in the master ontology (a LAV approach). Additionally, source descriptions contain metadata specifying query capabilities and security information.

Queries are specified in terms of the master ontology concepts. Users (including ACGT services) do not need to be aware of data source schemas or access methods in the participating centers. Queries are sent to a mediator component which translates the queries to use terms from the local ontology of the data source. The translated queries are sent to wrappers which use the source description to form a query expressed in the local database schema. The result from the local database query is translated to use concepts of the local ontology and sent to the mediator

which translates the data to use terms from the master ontology and integrates results from potentially several sources. Results from queries or data mining software are shared in a common storage system.

2.4.4 Discussion

When designing any system which deals with patient data (such as CDSS), many factors have to be considered. The usage of consistent standards must to be a priority in order to promote secondary uses of patient data. Because of semantic problems when dealing with codes in different terminologies, heterogenous data from many sources are difficult to correctly integrate and interpret. Identification of similar terms (aliases) among various terminology systems is error-prone since the systems can differ in structural organization, semantics and in granularity of concepts.

The semantics of patient data is important. The string "C61" does not convey any content by itself. By adding the knowledge that the code is an ICD-10 [149] code we can understand that it represents a diagnosis code for malignant neoplasm of prostate. By adding additional information, such as temporal information (when was the value entered in the system), the source of the information (measuring equipment model, doctor, nurse, CDSS advice etc.), patient history (previous examinations and diagnoses), what ranges of values are normal (or possible), we can describe the background of the data (i.e. metadata). Even references to the actual workflow at the source clinic can be specified, allowing data analysts to determine if data are trustworthy.

Another important aspect is the coverage of medical terms. It is not sure that a single standard can be used for all fields in medicine. This is, in part, the motivation for the multitude of standards in medicine. In some cases, existing standards have been adapted for use in speciality fields, see for example the Diagnostic and Statistical Manual for Mental Disorders (DSM) standard [5] which is a more detailed version of ICD-10 but focused on mental disorders. Secondary uses, such as statistics and knowledge extraction, make it difficult to predict who will use the data.

Mappings of medical terminology systems are likely to be necessary when integration of heterogenous data sources is pursued. The Unified Medical Language System (UMLS) [16] attempts to unify medical terminologies by using a metathesaurus. It enables code translations between the terminology systems (nearly one hundred, counting national variants) which is an unavoidable requirement when faced with heterogenous patient data sources. The metathesaurus itself already provides relations between clinical concepts. This is further specified in the semantic network of UMLS. Biomedical terms are specified in a specialist lexicon. Being able to communicate data between systems using different terminologies is not only recognized as an important feature by UMLS but also by more traditional terminology systems such as Read codes (also called the Read Clinical Classification).

Additionally, genetic information and, in particular, patient data are protected by data-privacy legislation. Transferring such data is complex from a legal point of view, in particular over country borders (see Section 2.3).

2.5 Rule generation from patient data

In Paper I, we studied generation of association rules from clinical data (see also Section 4.1.1). In this section, the method used in that study is summarized.

2.5.1 Association rules

Similar to rough sets (see Section 2.6.1), *association rules* [3] can be used to find associations between objects in a data set. Association rules describe frequent attribute-value combinations.

Let $I = \{I_1, I_2, \ldots, I_n\}$ be a set of items and $T = \{t_1, t_2, \ldots, t_n\}$ a set of transactions, where each transaction is a subset of I. An association rule is usually written as $A \Rightarrow B$, where *antecedent* A and *succedent* B are disjoint subsets of I. Such a rule should be interpreted to mean that transactions containing A tend to contain B.

The *support* of $A \Rightarrow B$ is given by the frequency of transactions containing items

from A and B. The *confidence* is the number of transactions containing A and B divided by the number of transactions containing A, i.e. it is the conditional probability of B given A.

In medical applications, useful rules could associate patients with certain lab test results and symptoms to a specific diagnosis. Even without a specific outcome defined as a target, it is possible to find clinically interesting association rules by testing variables. In [17], association rules were generated based on one year of *Pseudomonas aeruginosa* infection control data. The purpose was to find low confidence but high support rules, since high confidence, high support rules were assumed to be already known by the physicians. This was also a way to filter the large amount of rules. The data was divided into various time-slices and almost 20000 rules were discovered, illustrating the need to have a filtering strategy which shows only the most relevant rules to the end-user.

Another study of how to apply data mining techniques to medical data is presented in [128]. A healthcare DW was established in order to have a stable data source as a base for investigations. The data warehouse was an example of the cross-domain collaboration needed, including medical staff, computer scientists and patient record experts. The aim of the data mining was to support business decisions for the health care organization. Several data mining techniques were combined, including association rule discovery. Some results from the data mining tools were not easy to interpret by users and therefore an OLAP tool was used to visualize results. Besides providing support for management decisions, the system has had a positive effect on the codification of patient data and services. The analysis can also be semi-automated and thereby more easily repeated on-demand by users.

2.5.2 General Unary Hypothesis Automation

General Unary Hypothesis Automation (GUHA) generates hypotheses based on data. The formulas generated are of the form $A \Rightarrow_{p,t} S$ where A is the antecedent and S is the succedent, meaning that $100 \cdot p$ percent of the objects (patients) in the data

satisfying A also satisfy S. The antecedent A is t-good if at least t objects satisfy it. In many ways, this is very similar to the notion of association rules with support and confidence.

Several procedures are available but, in particular, the *4ft-quantifier*, $\approx$, is important since it directly provides association rules $\varphi \approx \psi$, where antecedents φ and *succedents* ψ are conjunctions of Boolean attributes, or *literals*, such as $\texttt{Age}(70; 80)$ and $\texttt{AnginaPectoris}(STABLE)$.

Given a data matrix M, the association rule $\varphi \approx \psi$ is verified using the four-fold table (4ft), see Table 2.1.

M	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Table 2.1: The four-fold table.

The table should be interpreted to mean that "a is the number of objects satisfying both φ and ψ, b is the number of objects satisfying φ but not ψ, $a + b$ is the number of objects satisfying φ", and so on. Note that an association rule can be either true or false given a data matrix.

Various 4ft-quantifiers are defined in [54]. A *founded implication* $\Rightarrow_{p;Base}$, with $0 < p \leq 1$ and $Base > 0$, is subject to conditions:

$$\frac{a}{a+b} \geq p \text{ and } a \geq Base$$

The association rule

$$\varphi \Rightarrow_{p;Base} \psi$$

is interpreted as "$100 \cdot p$ percent of objects satisfying φ also satisfy ψ" or "φ implies ψ on the level of $100 \cdot p$ percent". The "*above average relation*" $\approx^{+}_{p;Base}$, again with $0 < p \leq 1$ and $Base > 0$ is subject to conditions

$$\frac{a}{a+b} \geq (1+p) \cdot \frac{a+c}{a+b+c+d} \text{ and } a \geq Base$$

The 4ft-Miner approach can be embedded into a logical calculus for a first-order language. The deduction rule

$$\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$$

is applicable if we have that $\varphi' \approx \psi'$ is true in M whenever $\varphi \approx \psi$ is true in M. Further details are available in Paper I.

2.6 Categorical approaches to rough sets

In Papers VI and VII, we use category theory, in particular the structure of *rough monads* (see Section 2.6.2), to represent rough sets in order to deal with drug-drug interactions.

2.6.1 Rough sets

The theory of *rough sets* [103] is based on the idea that objects (in this case patient cases) containing a subset of attribute values in common are *indiscernible* (considered identical) *based on the available knowledge*. Using analysis of medical data as an example, an outcome is chosen together with a subset of attributes to be studied and algorithms create the rough set representing patient cases with that outcome. Patient cases with the outcome are considered to be in the *lower approximation* of the rough set unless they have identical attribute values with a patient case with a different outcome (i.e. they are counter-examples). The latter patient cases are considered undecidable (possible members in the rough sets) and are in the *boundary region* of the rough set. The *upper approximation* of the rough set contains both the lower approximation and the boundary region. Other patient cases are not considered as part of the rough set. Please see section 2.6.2 for details.

As mentioned before, medical data often has many attributes and therefore the process of finding a *reduct* is important. A reduct is the minimal set of attributes that preserves the *indiscernible* relation. In other words, a reduct maintains the ratio of

properly classified objects. Redundant attributes are in this aspect no longer regarded as they do not contribute to the classification.

In [76, 77], rough sets were applied to data on solitary pulmonary nodule patients in order to find a smaller subset of patient data, test results, etc. than the set currently used in medical diagnosis practice. This is important because this might mean that patients do not have to undergo biopsy. The results showed perfect (100%) prediction accuracy for 91.3% of the patients. On the other hand, the patient data set was small and it is difficult to draw any conclusions from this result.

A good survey of applications of rough sets in medicine and other areas can be found in [73].

2.6.2 Rough monads

In Papers VI and VII, we use an implementation of rough sets using an abstract mathematical theory called *category theory* to represent drug-drug interactions. One main advantage of category theory is the ability to abstract in the sense that many structures can be described in terms of relatively few categorical ones. Category theory has been applied to computer science: for instance, term rewriting systems, game semantics and concurrency.

Briefly, a *category* consists of a classe of *objects* and *morphisms* (mappings between objects). Each morphism has one object as the domain and another object as the co-domain. Identity morphisms are mappings between the same domains while composite morphisms are mappings between the domain of one morphism and the co-domain of another morphism.

If the objects in a category are themselves categories and the morphisms are mappings between the categories, those morphisms are called *functors*. *Natural transformations* provide mappings between functors. A *monad* consists of a functor together with two natural transformations fulfilling certain conditions. In Paper VII, we include formal descriptions of these structures.

When a functor is extendable to a monad, it is especially interesting, in particular

with partial order integrated. To enrich the structure of a powerset, we can consider the partial order given by set inclusion. The powerset functor extended to a monad can thus be further extended to a so called *partially ordered monad*. These partially ordered monads are powerful tools for providing rough set operations, and in general for working with rough set like general structures.

Background

By using category theory, we can represent rough sets based on monads. Monads have been used in different areas, such as unification where they were used to extend the classical concept of a term to a many-valued set of terms [48].

Rough sets build upon relations, and categorically, a relation can be seen connected to a powerset functor that can be extended to a monad. Further, whenever general powerset monads can be extended to partially ordered monads, this structure is sufficient for the provision of rough set operations in a category theory setting.

Rough monads

Applications in health care reveal a need for extension of traditional rough sets to handle vague information [73]. In this context, *rough monads* [38] provide a promising setting for further developments. It is possible to represent uncertainty with both rough sets and fuzzy sets. These two concepts can be connected by using partially ordered monads. The partially ordered powerset monad P, can be used to provide rough set operations to complement the many-valued situation. This section briefly summarizes the main concepts.

Rough sets are traditionally defined on ordinary set based relations, i.e. relations on X as subsets of $X \times X$. More specifically, we start by defining a relation on a set X with a mapping $\rho_X : X \to PX$, where $\rho_X(x) = \{y \in X \mid xRy\}$ and the inverse relation R^{-1} is represented as $\rho_X^{-1}(x) = \{y \in X | xR^{-1}y\}$. To be more formal, given a subset A of X, the lower approximation of A corresponds to the objects that surely (with respect to the considered relation) are in A.

The lower and upper approximations of A are, respectively, obtained by:

$$A^{\downarrow} = \{x \in X | \rho_X(x) \subseteq A\} \text{ and } A^{\uparrow} = \{x \in X | \rho_X(x) \cap A \neq \emptyset\}$$

These basic definitions of rough sets are reformulated using the partially ordered powerset monad, $(P, \leq, \eta, \mu)$ (where η and μ are the natural transformations providing the monad). This allows us to represent the upper and lower approximations of a subset A of X as:

$$A^{\uparrow} = \bigvee_{\rho_X(x) \wedge A > 0} \eta_X(x) = \mu_X \circ P\rho_X^{-1}(A)$$

and

$$A^{\downarrow} = \bigvee_{\rho_X(x) \leq A} \eta_X(x),$$

respectively.

These upper and lower approximations can be further generalized, as shown in [38].

2.7 Case study: Data extraction from patient journals

In [71], we performed a case study to determine the possibilities and difficulties of using an EPR as a data source for analysis. Data warehousing approaches can be useful in quality assurance and data analysis (see Section 2.4.2). Data retrieval from record systems has been previously attempted, see for example the approach in [37], where a system for data retrieval from MUMPS based record systems was developed for use within a laboratory environment. Search criteria given by users were converted to querying code, and resulting data files delivered to the end-user, thus minimizing retrieval time from the end-user point of view. A similar approach was reported in [11], which describes graphical query generation based on object-oriented user views.

We studied patient data related to Lower Urinary Tract Symptoms (LUTS). These symptoms affect the quality of life of patients. In most cases, the cause is benign but could be more serious, including prostate cancer.

LUTS symptoms can be organized in three categories [2]:

- storage symptoms, including high frequency of urination, nighttime urgency to urinate, urgency to urinate and leakage.
- voiding symptoms, including slow stream, hesitance to initiate urination and muscular straining.
- post-urination symptoms, including feeling of incomplete emptying of bladder and urine leakage directly after urination.

Particularly for women, age is an important factor which influences the occurrence of LUTS [88]. Among men, these symptoms are often associated with benign prostatic hypertrophy (BPH) [137], sometimes called enlarged prostate. Other causes among men for these symptoms are prostate cancer, neurological diseases and diabetes. Treatment of LUTS depends on the diagnosis. If the patient suffers from prostate cancer, options include surgery, radiation therapy, hormone therapy or biological therapy. The treatments also vary depending on the individual patient. One example of such variations is the treatment of non-benign BPH. This condition could be treated surgically, but the patient may choose to avoid surgery and instead be treated conservatively.

The data collected included information about male patients operated during a one-year period. These patients had been diagnosed with LUTS (including enlarged prostate) and their conditions required prostate surgery.

The data collection was performed manually with the assistance of a domain expert. Data was gathered from the year when the EPR of the hospital was introduced at the urology department of Norrlands Universitetssjukhus (NUS). The record system BMS Cross was used in clinics and health centers within the County Council of

Västerbotten in Northern Sweden. About half of the data were collected from paper-based patient records and the rest from the EPR.

Several laboratory tests were selected after discussions with physicians at the department (see Table 2.2). Data in the EPR were not always stored in the intended database tables, thereby further complicating data collection. For example, laboratory values were entered as part of the clinician's free text note even though the EPR provided functionality to enter such values in a structured way. The reason for this could partly be that the EPR was recently introduced at the time of the study.

Data for several patients were incomplete. Identified reasons were:

- The physician decided that a particular test was not necessary to make a diagnosis (although this test might be considered interesting in subsequent data analysis).
- Department-specific routines suggested to record test values elsewhere than in the EPR.
- Cumbersome user interfaces had the effect that values of tests were recorded as a non-structured note rather than at the proper and codified location in the EPR.
- Some values were not possible to measure since the patient was using a catheter.

Encouraging structured data entry (SDE) from the physicians enables also linkage to medical terminology systems. Changes in routines at the clinic have since then improved data entry. The conclusion from this study is that it is indeed important to design and maintain carefully structured databases encoded with information encoded with useful and internationally recognized terminology. Several values in the EPR for LUTS data were not standardized according to a terminology system, for example, bacterias found in urine cultures were inconsistently named.

Similar problems were reported in [108] which describes establishment of a medical DW. Data regarding obstetrical patients were collected to identify factors that contribute to perinatal outcomes. Some values, although valid, were not usable in data

analysis, since they were stored in free text notes. Analysis of medical data should therefore be preceded by the application of techniques that can handle missing values, see for example [43] for some solutions.

Variable name	*Description*
OP_DAT	Date of operation
FÖDELSE	Year of birth
PSA	Level of Protein Specific Antigen
U-ODL	Date of urine culture
NEG	Result of urine culture (boolean)
BAKT	Name of bacteria found in urine culture (if any)
RESECERARA	Number of grams of prostate removed during operation
TURP	Operation method (boolean)
TUIP	Operation method (boolean)
STEN	Bladder stone present? (boolean)
FÖRE	Date of measurements before operation
KV1	Urine (ml)
MF1	Average flow of urine (ml/s)
RU1	Remaining urine in bladder (ml)
IPSS1	International Prostate Symptom Score
QULIFE1	Quality of Life score
3_MÅN	Date of first follow-up measurements (around 3 months)
KV3	Urine (ml)
MF3	Average flow of urine (ml/s)
RU3	Remaining urine in bladder (ml)
IPSS3	International Prostate Symptom Score
QULIFE3	Quality of Life score
1_ÅR	Date of first follow-up measurements (around 1 year)
KV1_ÅR	Urine (ml)
MF1_ÅR	Average flow of urine (ml/s)
RU1_ÅR	Remaining urine in bladder (ml)
IPSS1_ÅR	International Prostate Symptom Score
QULIFE1_ÅR	Quality of Life score
ÖVRIGT_1_ÅR	Comments from physician

Table 2.2: The data collected

Chapter 3

Processes

> "Now, listen carefully: path: follow path; gate: open gate, through gate, close gate." Stephen Falken, War Games (movie)

3.1 Introduction

Standardization of processes in health care and bioinformatics is a key issue for promoting evidence-based medicine. A *process* represents an act of doing something, for example some type of data processing or patient care. Process descriptions in health care (see Section 3.2) specify information flow among staff, patients and software systems. Focus in these *human-oriented* processes is on the management of human or physical resources. Diagnostic and therapeutic processes in health care can be represented as clinical guidelines (see Section 3.2.1). Process specifications in bioinformatics (see Section 3.3) typically specify information flow between software components. Such processes are called *system-oriented processes*.

A *workflow* describes a series of activities (or sub-processes) as part of a process. Activities can be linked and represent transitions between activities according to predefined events or conditions determined on-the-fly. The links can carry information from one activity to the next. In these cases, the information that flows into an activity is the input of the activity and the information that flows out is the output

of the activity. If activities are linked, the links represent dependencies (for example, one type of dependency states that one activity must occur before another). *Workflow modeling* is the act of creating a model based on existing or planned processes. A *workflow enactor* controls, monitors and coordinates activities within processes as defined by a workflow.

3.2 Processes in healthcare

Medical practice is continuously improved as the knowledge about the human body, diseases, genetics and drugs increase. It is almost impossible for a doctor to keep up with the latest scientific advances, unless the doctor is a specialist and is allowed time for continued education/study in the area of expertise. Doctors should also be able to judge the validity of both methodology and findings in research papers and decide if their patients are similar enough with the subjects in the study to apply new results in their patient care. Clearly, this is an overwhelming task.

For this reason, health care can benefit from the structure and standardization achieved by process modeling. Typically, standardization can be achieved through clinical guidelines and decision support systems (which can be based on guidelines). These guidelines encourage using a scientific basis for decisions in medical practice and are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [44].

Through a process of consensus among a group of experts, such guidelines can be said to represent, if not the only correct advice at least based on the current state of research in the field, good enough advice [104]. Although workflow studies are made in health care (see Section 3.2.2 for an example), it is common to structure decision processes in healthcare through clinical guidelines. However, the outcome of guideline introductions are not always positive, in fact an early study [53] showed limited effect of clinical guidelines on the medical practice. They conclude that the effect of guidelines largely depends on subsequent evaluations.

Computerized clinical decision support systems (CDSS) are active knowledge-based systems that use patient data to generate case-specific advice [140]. Clinical guidelines sometimes provide the basis for logic in CDSS (see for example [106]). Such logic can also be inferred directly from patient data (see for example [6, 42]). Successful application of CDSS largely depend on the use of terminology systems to enable data exchange with other clinical systems, see Section 2.4.1. Information systems building on medical knowledge databases (see for example the systems with pharmacological information described in [67, 41]) do not necessarily use patient data and, as such, are not considered to be CDSS. Nevertheless, they can contribute with important background knowledge to clinical decisions.

Compliance with clinical guidelines is a step towards evidence-based medicine. For effective dissemination and use in clinical practice, guideline logic needs to be synthesized into rules that can be implemented in CDSS [36]. While some clinical guidelines describe different steps and information flow, they often lack a formal description that facilitates development of CDSS. The reason for this lack of formalism is that such guidelines are often primarily intended for usage by care staff and only to a lesser degree for direct use as a base for software developments. Another difference compared to system-oriented workflows is that clinical guidelines are typically developed by a consensus among a group of domain experts. They involve the synthesis of the current state-of-the-art in the given domain and may take considerable time and resources to produce. This type of peer-review is typically not made for system-oriented workflows in bioinformatics (see Section 3.3).

Many clinical practice guidelines are published as text documents, see for example the JNC guideline for hypertension treatment [92]. While the JNC guideline is evidence-based and widely accepted, it does not contain any clear logic (rules) to use for reasoning in software. In [106], a decision support system for pharmacological treatment of hypertension was developed based on the JNC hypertension guideline. Logic for the software was derived by manual review of the guideline. To avoid such time-consuming and potentially erroneous reviews, a more formal way of representing

guidelines is needed.

3.2.1 Clinical guideline representation

The representation of clinical knowledge (such as rules) can cause problems regarding re-use and maintenance of knowledge. In [18], approaches to representation of clinical guidelines are divided in three categories; procedural approaches, rule-based systems and task-based systems. In early systems such as Arden [109], a procedural approach was used where the clinical knowledge is integrated in the source code itself as Medical Logic Modules.

Alternatively, in rule-based systems, the logic and medical knowledge can be represented in a declarative form. This allows developers to clearly separate algorithms that are used to apply the rules from the rules themselves. While this rule-based approach certainly simplifies knowledge re-use, it is still hard to formulate rules without connecting them closely to the specific context. This problem was addressed in task-based systems such as PROforma [18] where the interpretation of the rules varies between different contexts.

Asbru [122], in particular, comments the problems of using other formalisms, such as traditional Workflow Management Systems (WfMS), to represent healthcare processes. In many cases, WfMSs are not able to cover enough information to accurately represent the sometimes ad-hoc processes in health care. However, WfMS can be useful if the process can be clearly structured.

Guideline models are compared in [105] to identify similarities and differences and, if possible, identify areas of similarity which could contribute to a consensus for representing clinical guidelines. The models included in the study are Asbru [122], EON [139], GLIF [98], GUIDE [111], PRODIGY [63] and ProForma [18]. Aspects such as decision models, temporal representation, connections to terminology systems and EPRs are compared. Several common mechanisms were identified and efforts are under way to standardization within the context of HL7 [56]. It should be noted that the goal was not the identification of and agreement on a universal guideline model.

Such a model could become overly complex and, additionally, it is better to let developers choose whatever model best fits the guideline complexity and integrates well with other software systems in the health institution. Indeed, the guideline models were developed with different philosophies and, naturally, could not be incorporated under a universal guideline model.

3.2.2 Case study: Information flow of radiology treatment

While it is clear that informatics support of processes in health care is useful [53], it should be noted that those processes are often complex and not easy to model. New information system in health care need to be integrated in the care process and with other software systems. The human-oriented processes at the radiology clinic of the Norrland Universitetssjukhus (NUS) were studied in [69] to better understand the information flow. Radiology treatment is an information-intensive activity and the clinic was using several software systems. The process was previously modeled [47] using a CASE tool and included administrative information (appointments), determinations of radiation dosages and angles to isolate effects in tumors, simulations of radiation, care plans etc. Approximate time requirements and frequencies of the activities were also included. Based on this information, the paper discusses requirements of workflow modeling tools for health care applications (robustness of workflow enactments, workflow sharing and effects on software development).

The model from [47] was further studied to determine the suitability of using it for simulation purposes [40]. Simulation is useful to determine bottlenecks in the work processes and, specifically, could pinpoint places where improvements in informatics support is needed.

Petri nets provide a method for discrete simulation [115]. We chose time Petri nets, a high level Petri net augmented with time and container places, for simulating the radiation therapy CASE models. One problem discussed in [40] was how to represent the CASE models as time Petri nets. It became apparent that the CASE model, although very specific, was not enough by itself as a basis for representing the

processes as Petri nets. Additional study of information systems and processes at the clinic as well as radiation therapy in general was necessary. Compared to the CASE model, the time Petri nets could also represent resources such as doctors and nurses explicitly in the model.

A similar approach was taken in [110] where Petri nets are used for simulation purposes of a workflow for management of Acute Ischemic Stroke. Although petri nets could represent the workflows accurately, non-expert users had problems to understand the notations. An intermediate representation in Workflow Process Definition Language (WPDL) is therefore suggested together with a software that translates the WPDL model into Petri nets for simulation purposes. Resources (human and technological) are represented in an organizational model. Through workflow simulation, new workflows can be tested and potential bottlenecks or problems be discovered prior to the actual implementation in the care environment. Establishments of organizational model and study of information systems, already in place, thereby increase the chance of successful introduction of computerized support.

A model-driven approach is described in [7], where HL7 and DICOM [87] standards are used as a base to create workflows in health care. The workflows involve web-services and security requirements are handled with WS-Security. In this case, there was no prior analysis of specific work processes but instead the workflows were based on sequence diagrams from the IHE initiative [126]. The paper focuses on establishing the approach for a semi-automatic generation of workflows based on IHE artifacts.

While the approach in [7] is interesting and could be useful in settings where HL7 messaging is used, applying it to the systems of the radiology clinic of NUS would almost certainly require extensive changes in those systems. At the time of the workflow study in [69], the electronic patient record used in most other clinics at NUS had not been introduced yet due to the specific requirements of the clinic. Several different systems were used for the different tasks and many parameters had to be manually re-introduced as the patient care process advanced. Support of HL7 for message interchange should be a requirement for future systems. Normal activities

in the clinic would not require messaging through web-services but since the clinic is also supporting other hospitals, this could be a future requirement.

Workflow modeling and simulation of radiation therapy is important as a preparation for various organizational changes, such as distributed decision making for radiation therapy. The clinical competence at the clinic is very high and other care institutions wish to take advantage of that competence through tele-medicine. As smaller radiation therapy units are conjoined into larger units, difficulties arise e.g. in information logistics, which in turn places heavy requirements on respective patient record systems to provide necessary clinical information about the patient to define suitable treatment plans. Enhancing record systems with necessary information structures in turn requires a thorough workflow analysis of the considered clinics so that information logistics and resource management can be well understood and jointly organized.

3.3 Processes in bioinformatics

Significant opportunities for personalized health care are possible because of advances in post-genomic research [84]. Personalized health care can be applied to diagnosis through genetic profiling of patients, if clear disease-genome associations are known. In some areas, diagnosis is well-established and straight-forward. For example, allergy diagnosis is often easier than allergy treatment. Treatment decisions can be improved by studying treatment response based on genetic information. However, clinical trials, which constitute the scientific base of such health care [121], need increased informatics support [120]. Such informatics support includes administrative tasks, trial monitoring, data management and analysis.

Approaches to data analysis in bioinformatics are typically based on web-services and system-oriented workflows. We adopt the view of [150] where a web service "...is a software system designed to support interoperable machine-to-machine interaction

over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards."

Web services are intended to be used by other software components. Consequently, their interfaces must be described in a machine-friendly way. Service consumers that use these distributed tools need to be able to dynamically discover (see Section 3.3.1) and execute tools and is thereby able to use new algorithms to process data. This discovery process is supported by tool metadata and this metadata is therefore shared in a service registry (public metadata repository).

In service-oriented architectures (SOA), several roles can be defined:

- Service provider: This role provides access to an analysis software wrapped in a web-service. This web-service is accessed via the internet and described with an interface in a machine-interpretable format. WSDL [144] is the de-facto standard for describing the machine-interpretable interface, including details regarding how to interact with the service, syntax used for input and output etc. There are various formats for describing service compositions (workflows), a typical example is the Business Process Execution Language (BPEL) [64].

- Service consumer: This role is requesting a service to solve a specific problem. The service consumer (or *client*) already has some input data and is looking for a service that a) solves the problem and b) can consume the data.

The following activities are typical for SOA (see Figure 3.1):

- Register: This activity is performed by the service provider who registers (publishes) the service by sending a description along with additional metadata to the service registry.

- Discover: This activity is performed by the service consumer which sends metadata describing the requested service to the registry. If successful, the service registry returns matching service descriptions.

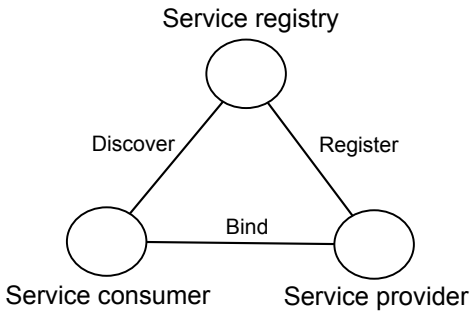

Figure 3.1: Important concepts in a typical service-oriented architecture.

- Bind: This activity is performed by the service consumer by using the service description from the discover activity to communicate with one of the places where the service is available.

Considering technological advances in high-throughput sequencing, gene expression monitoring and proteomics data acquisition, research in bioinformatics could potentially require processing and analysis of massive data sizes (see Section 2.2). On-line software for data analysis is in bioinformatics, where large computational resources and access to large and constantly updating databases are requirements for solving many typical problems. This is the main motivation for delegation of tool invocation (execution) to remote servers with required computational resources. High performance computing has been suggested as an effective solution [65] to process such data.

Such a distribution of analysis steps presents several challenges that should be addressed, among others:

- *Limitations of data transfer.* Since data in bioinformatics (see Section 2.2) can be very large, data transfers must be minimized.

- *Long-running service invocations.* Analysis may require substantial time to complete (not necessarily related to the size of the inputs). There is a risk that connections to web-services are terminated, making it impossible to retrieve results. This problem is discussed in Section 3.3.3.

- *Service data compatibility.* Typically, several service invocations are necessary to produce interesting results. Data formats used in services are typically defined using XML Schema but since these definitions usually are not shared, sending the output from one service as input to another service not developed by the same service provider can be problematic.

It has, however, become clear that current techniques need to be complemented with semantic descriptions of services, ideally using standards developed for the semantic web [14, 78]. Web-service (or simply service) metadata can be used in the sense of supporting discovery and definition of data-mining services through a catalogue or registry with service metadata. Metadata can be described as "data about data". The approach of publishing service metadata in repositories is not new in bioinformatics, see for example Grimoires [148], MyExperiment/BioCatalogue [52] or the ACGT metadata repository [114].

BioMOBY [145] aims to facilitate integration of services. The current approach is based on a metadata repository (MobyCentral) where WSDL descriptions for SOAP services and other metadata are shared. Discovery of services is performed through a well-defined API. Metadata is defined using a service ontology which has been developed in collaboration with MyGrid [131]. Discovery mechanisms in BioMOBY are limited since the MobyCentral API only provides fixed but well-specified queries in the form of web services. Therefore, in a separate branch of the project, S-BioMOBY, the developers show how the same metadata can be exploited using standard approaches adopted from the semantic web. Clients supply a graph of concepts to a semantic reasoning mechanism which tries to fill the graph. In this branch of the project, services follow the REST-approach to service invocations (based directly on XML and HTTP).

In bioinformatics, processes are often system-oriented (in-silico). Specification of workflows based on these processes makes it possible to repeat sequences of tool invocations, varying input values and thus allowing a researcher to compare results. This

is a distinguishing characteristic of in-silico bioinformatics as compared to human-oriented processes in health care where patients are affected during the process (administration of drugs, operations etc). System-oriented workflows facilitate and standardize combination of resources to solve high-level tasks, for example complex web-service compositions (see Section 3.3.2).

Bioinformatics workflows typically operate on a high level of abstraction in the sense that they are using and connecting existing components. Workflow design is aimed at domain experts and therefore many systems have customized graphical user interfaces which attempt to hide complex details. On the other hand, there are cases when complex data manipulation is required (for example converting data from one format to another). One system with such capabilities is Taverna [99] where scripts can be integrated in the workflow to process data locally before sending data to additional web-services for further processing.

Several requirements for workflow systems are discussed in [30]. Workflow composition (or service composition) is sometimes automated to various degrees (see Paper V) based on semantic annotations (see Section 3.3.1). Depending on the system, the activities in a workflow are only abstract and systems map the abstract activities to specific resources (for example computational resources) during workflow enactments. This mapping involves scheduling and optimization of workflow enactments (if several instances of a resource are available). However, some systems, such as Taverna, directly specify which resource to use during the design-phase (i.e. no scheduling). Provenance metadata is often available from the workflow enactor. This is important since it documents *how* results from workflow enactments were obtained, and can be used as one aspect in workflow recommendation (for example, recommending workflows that have been successfully enacted in the past).

3.3.1 Semantic annotation of web-services

Basic requirements for service discovery are well-defined ontologies and use of such ontologies in service annotation [147]. A trend in automatic workflow generation is to

use standards developed for the semantic web. There are several levels of complexity and ambition for such approaches.

Clearly, specification of only syntax is not enough to enable efficient web-service discovery since a web-service that declares a string as input (i.e. syntax) could be expecting a biological sequence, name of a person, etc. In [50], a system is described which uses an algebraic representation of analytical processes and syntactic compatibility as a way to ensure compatibility between web-services. The experiences from that system shows the importance of semantic annotation with concepts from ontologies to enable efficient and scalable service composition.

A storage and sharing utility for XML-based representations of workflows and metadata is presented in [9]. The functionality of this system is basic but it could be useful in a setting where the number of services is expected to be low and the possible overhead of (manual) semantic annotation could be avoided.

In [60], authors show how MyGrid technologies can be used for sharing research results and experiment setups. Historically, research groups have used their own proprietary tools for analysis and data preparation. In such settings, it is very difficult for external groups to reproduce results. By using a workflow metadata repository and standardized technologies (for example, make tools available as web-services), other groups could take advantage of prepared workflows by enacting them with their own data with or without modifications to the workflow. The conclusion is that workflow technology could significantly improve research practices through sharing and comparing of experimental setups.

Several requirements for workflow metadata repositories are discussed in [95]. The base for their system is a shared workflow metadata model. Several perspectives of workflow metadata are specified, such as information specifying what a workflow does, how data flows between components of a workflow and provenance metadata (when, by whom etc). The technology used is CORBA but the information model could be implemented using other technologies such as SOAP.

Authors in [79] describe an object-oriented design of a workflow repository. They

outline a set of requirements for workflow repositories. Workflow components are suggested to be annotated with properties independent of workflow instances such as the function of the task. These properties allow components to be re-used.

In [143], metadata is specified for workflow components. To facilitate workflow discovery, each workflow component is associated with a set of simple keywords. This represents a limited approach since the keywords, as far as the paper describes, are not standardized or from an ontology describing functionalities of workflows. Keywords can be useful depending on the scale of the workflow repository but there should exist a mechanism for more sophisticated annotation when required.

The architecture in [1] supports design, testing and enactment of workflows. Metadata can be shared through a workflow repository where simple text description or XML elements can be associated with workflow instances. Authors recommend using XPATH or XQUERY queries for selecting workflows. There is no reported support for semantics or more advanced workflow discovery. However, the approach is interesting as it takes a simple but standards-based approach to workflow representation, discovery and storage.

Although metadata and service annotation can assist in selection of services, decisions regarding suitability of service for solving a specific task will ultimately be decided by individual researchers by inspecting additional human-oriented information, such as software documentation and authorship.

3.3.2 Web-service composition

Paper IV includes a brief overview of web-service composition in bioinformatics. This section extends that overview and discusses the use of semantic annotation of web services as a base for (semi)-automatic combination of web services.

Automatic service composition, also called automatic workflow generation, aims to automate to various degrees the task of finding optimal combinations of atomic services to solve higher-level tasks. The problem addressed in service composition can be described as follows; *given two sets of data types (representing inputs and*

outputs respectively), find the optimal set of inter-connected services where the input data types of the initial service(s) are of the specified input data types and the output data types of the last services are of the specified output data types. Two services can be inter-connected if the output of one is compatible with the input of the other.

The optimal set above can be determined by a combination of semantic closeness (in terms of input/output data) and non-functional service metadata (QoS). For example, the algorithms usually favor services whose input/output parameters consume/produce messages that are semantically similar to the requested semantics. Examples of non-functional metadata include level of availability, error rates, cost (expected CPU time or data transfer time).

Assuming that the services are well documented and that the number of possible services and/or data types is low, service discovery during workflow construction can be performed manually. However, when the number of services reach a certain level, support for connecting and selecting web services becomes necessary.

Different levels of computerized support for making such connections are possible, ranging from semi-automatic where the user is interactively given advice regarding suitable services during each step of workflow generation, to fully automatic where the user only provides input and output data sets and the algorithm directly generates possible service compositions.

An ambitious framework for service composition is presented in [19]. The paper gives an overview of the state-of-the-art and makes a strong argument for automation in service composition. The framework adopts DAML-S [26] for semantic descriptions of web services. While recognizing the importance of WSDL service descriptions, they also note that WSDL is not sufficient for efficient service discovery (and thus service composition) when the number of services grows large. Service descriptions using DAML-S [26] complements the syntactic descriptions from WSDL with semantic annotations. The descriptions consist of a service profile that describes the capabilities of the web service, a process model that describes the internal activity of the service (to support clients in their service selection) and grounding specification where the

abstract messages (described in the process model) are mapped to the messages described in WSDL (the actual syntax and details of client/service message exchanges). This approach has several advantages; it recognizes the role of WSDL as a syntactical description while allowing semantic annotations of the services to be made. Of course, there is a certain amount of duplication of information between the semantic and syntactic parts of the DAML-S description and efforts like the one described in [102] aim to enable semi-automatic generation of DAML-S semantic descriptions based on the information found in WSDL. The approach of [19] is promising and should enable large scale use of semantic annotations of services.

DAML-S is also used in [129, 51, 10] to describe services semantically. Other approaches [153, 89, 127] use OWL-S [83] with similar descriptions of service semantics as in DAML-S.

In [151], SAWSDL [74] is used to simplify automatic composition of web-services using SAWSDL. Workflows are generated based on semantic descriptions of pre-conditions, effects and semantics of input/output messages. SAWSDL directly annotates parts of WSDL service descriptions with concepts from ontologies. The authors extend the SAWSDL standard by including pre-conditions and effects, which are important for describing stateful services. A stateful service is not fully defined by the input/output but also its behavior also depends on previous actions; for example, the state of a computation (job) in an asynchronous service for a given job identifier cannot be determined unless a job with that job identifier was previously started.

In [82], authors describe an ambitious framework to facilitate service composition. Fully automated service composition is suggested for some domains where the number of available services is very high. While a semantically well-described service is both good practice and quite possible in some domains, it is, in general, difficult to encourage service developers to annotate the services to the extent suggested. Nevertheless, the approach is interesting and is a good example of how service composition can be aided by careful and detailed annotation of web-services.

Developments [147] within the MyGrid project present a more limited expectation

for the feasibility of automated service composition compared to [19, 82]. Authors claim that in the domain of bioinformatics, researches are reluctant to give up control over their experimental setups. Automated service composition requires very fine-grained annotations and many services in bioinformatics are poorly documented. The expense of annotating these services makes the feasibility of the vision described in [82] questionable. Authors do, however, recognize the importance of service annotation to guide workflow designers in their work. In fact, using ontologies to annotate shims (simple data transformation services that enable two other higher level services to communicate) are considered especially useful.

The Artemis project [32] claims to be one of the first projects to consider semantics for health care web-services. The use of web-services has only recently been employed in considerable scale [126] but without considering semantics. Artemis considers a combination of semantic annotation of web service functionality and of messages as both aspects are considered essential for integration (similar observations are made in bioinformatics oriented efforts such as [145, 60]). Quite extensive efforts in the health care domain attempt to establish semantics of patient data through standards development (SNOMED-CT [119]). Different health care institutes may not use the same ontology for describing messages and some mediation is thus necessary. Several examples are given of how to map between messages describe using HL7 and ENV-13606 standards. Some messages may not be directly mappable and it is therefore necessary to perform complex compositions of web-services to produce results.

3.3.3 Case study: Management of long-running web-service invocations

Many typical analysis tasks in bioinformatics can take considerable time to finish. One such example is the structural analysis of molecular data (3D protein folding, etc.) using molecular dynamics. In fact, the analysis of complete biological processes can be extremely demanding both in terms of CPU and memory. Web-services can be used as a way to provide access to distribute computations but service protocols

must support long-running analysis.

Work in addressing service protocol issues resulted in modifications to a service integration architecture called BioMOBY [145]. A specification was developed for asynchronous calls to BioMOBY services [15], including how to submit data and request processing status and results, essentially defining stateful web-services.

There are two possible approaches to supporting asynchronous web-services; *notification* and *polling*. Notification is the case when a client not only submits the input for a web service but also a contact address where the service can notify the client that the processing has finished (and possibly also return the result). Polling is when a client submits the input for a web service and gets a ticket from the service. The ticket can be used to contact the service and inquire about the progress of analysis and later also to retrieve the results (see Figure 3.2).

We evaluated several external standards for the implementation. In particular, WSRF and ASAP were interesting since they are based on SOAP (as is BioMOBY).

WSRF [97] is an OASIS standard for inter-operable and stateful web-services that has been applied to grid environments (Globus [46]). WSRF defines how to represent resources and messages. These resources are represented by property documents with zero or more properties. Such a document represents the state of the resource and is transferred in the SOAP body. WSRF also provides a standardized way to request cleanup, the destroy operation. This operation is a request from the client that the resource should be removed. It is also possible for the clients to request a "TerminationTime" (i.e. scheduled cleanup)

ASAP [96] defines a standard of how to invoke asynchronous services and communicate status. Input to service instances can (in some cases) be altered during the lifetime of the call. Partial results of a service invocation can be returned even before the entire result is finished. Naturally, it depends on the particular service whether altering inputs and/or retrieving partial results are possible.

WSRF is general in its approach and can be used to implement stateful services in general. ASAP is conceptually clear since it is specifically designed for asynchronous

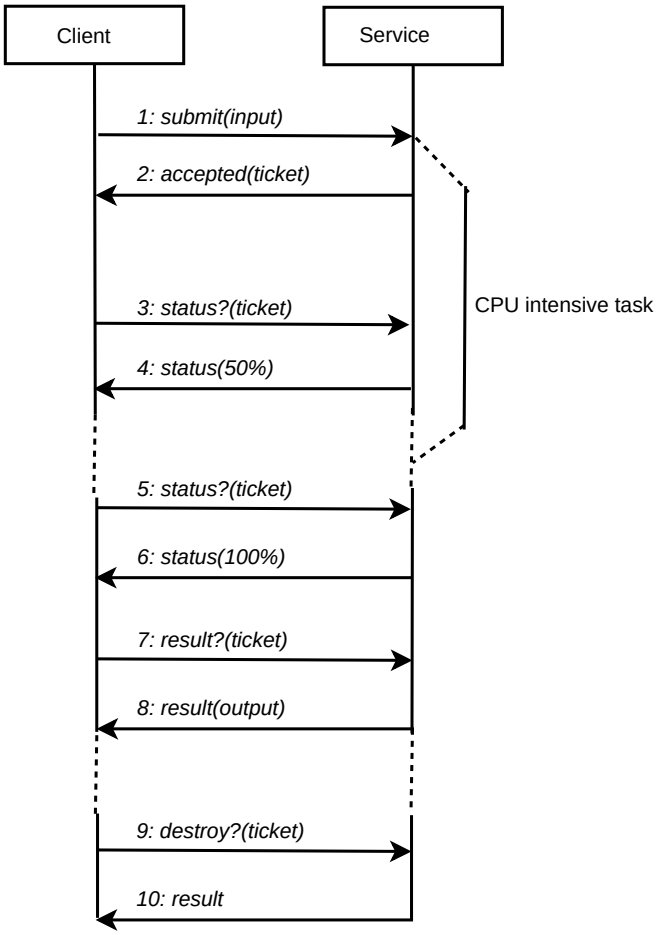

Figure 3.2: Typical steps in polling for asynchronous services

services. This standard also supports multiple invocations of a webservice from a single call (sending several inputs at once), something important since BioMOBY services can perform such batch calls.

Both standards can be used to implement polling and notification (in fact, ASAP directly supports this). However, we selected WSRF as a base for implementing asynchronism in BioMOBY. The main reasons were that WSRF could potentially be useful for other purposes that require states in web services and also had better support for software development with libraries in many languages, for example in C (Globus toolkit), Java (Apache WSRF) and Perl (WSRF::Lite).

We adopted a polling approach (see Figure 3.2) where clients start by submitting the data to the service which replies with an identifier when the initial data transfer is done. This identifier is later used by clients to poll for the state of the data processing. When the service indicates that the results are finished (or that an error has occurred), clients can retrieve the results (or error description) using the identifier.

Because a call to a BioMOBY service can contain several unrelated inputs; i.e. a batch-call which creates several analysis instances (or jobs), it is probable that some jobs finish before other jobs or that some jobs fail while others do not. This was supported in the specification for asynchronous BioMOBY services.

Services are also required to return status of jobs. Depending on the service implementation, different levels of status reporting are possible. Some services may only be able to say that the job is not finished, others can give detailed progress reports. Service implementors can choose between the event types defined in LSAE: Heartbeat event (the job is still running), Percent progress event (percentage finished of total job), Job State changed event (states include "created", "running" and "finished"), Step progress event ("step 2 of 3"), Time progress event (total time spent on the job).

The modifications were designed to not interfere with the previous synchronous specification in BioMOBY to avoid modifications of existing clients. The modifications are now part of the abstract programming interface (API) of the BioMOBY architecture.

Part II

This part of the thesis summary consists of an overview of the included papers and the bibliography.

Chapter 4

Summary of papers

This chapter summarizes Papers I-VII and is organized in two parts:

- Design aspects of CDSS and clinical guideline logic representation (see Section 4.1)
- Process modeling and sharing in bioinformatics (see Section 4.2)

Consequently, the order of the paper summaries differs from that in Part III.

4.1 Decision support system design and guideline logic

Issues related to clinical guidelines and clinical decision support systems are discussed in Chapters 2 and 3. This is the focus of the initial and future stages of the research (Papers I-III, VI-VII).

4.1.1 Paper I

Medical decision support can be based on rules from clinical guidelines (see Section 3.2.1). However, the logic in such guidelines is not always clearly defined so as to

enable re-usage in CDSS development. The study of Paper I discusses aspects of decision support system development from the point of view of logic support. The need for selecting an appropriate logic for particular guideline implementation tasks is discussed as well as other aspects of integrating decision support in clinical practice. Guideline logic and rules can also be inferred from patient data. Such rule inference is a possibility in the northern care region of Sweden, due to the use of a single electronic patient record.

The paper describes a previously published method of data mining called GUHA [54, 55], see also Section 2.5.2. We used a data set containing information about coronary artery bypass grafting. The data included pre- and post-operative information and were used in a software called 4ft-miner. Several rules were discovered during data mining, including some preoperative to postoperative predictions. The paper discusses the approach of GUHA and concludes that the rules are indeed useful as a base for development of CDSS. Several properties of the rules closely relate to concepts known by doctors such as support and confidence, GUHA rules are therefore possible to motivate for domain experts which contributes to confidence among end-users of CDSS.

4.1.2 Paper II

Pharmacological therapy is an increasingly complex activity that benefits from informatics support. Databases with information on drug-drug interactions and side-effects can complement information from EPRs. Drug terminology systems are essential for efficient communication of prescription data between information systems. Software interfaces encourage and simplify usage of drug databases. Remote access to drug information is also crucial to promote end-user usage. Paper II describes software developments related to a database with drug information to support mobile scenarios. A software prototype is presented which runs on a hand-held device with an integrated bar-code scanner. This device can be used to quickly collect information regarding drugs currently used by a patient. Additionally, a programming interface

to access pharmacological information was described and also further reported in [67].

One scenario that motivates the developments in Paper II is the discovery of cheaper but still therapeutically equivalent alternatives for an already prescribed drug. This is a complex task that requires more than just a general-purpose pharmacological database. To formalize when drugs can be safely exchanged, governmental agencies maintain lists of substitutable medicinal products. In Sweden, this task is performed by the Medical Products Agency. Only if substitution is allowed according to the Medical Products Agency's list [86], may a pharmacy exchange a prescribed drug for another. However, there are exceptions to the list. For example, drugs should not be exchanged when the container of the prescribed drug is specially adapted for use by rheumatic patients. Additional complexity comes from varying responses to drug therapy by individual patients. Patients may have tried a cheaper replacement drug in the past with low therapeutic effect and therefore would be suited to attempt a more expensive drug.

There are strong economical motivations to exchange an expensive drug for a cheaper one, if the therapeutic effect on a given patient can safely be assumed to be equivalent. In a study of medical profiles of hypertension patients [107], authors found that by strictly following a guideline for pharmaceutical hypertension treatment, a reduction of cost up to 40% was possible. Although the reduction in cost in practice would probably be less since doctors cannot be expected to always follow the guideline recommendations, economic (and medical) reasons clearly motivate increased guideline adherence.

The need for integration with clinical systems, such as EPRs and pharmacological databases, became apparent by the study of a CDSS for hypertension treatment [106]. While this system was able to provide a guideline sanctioned pharmacological treatment advice, the system could not interact with pharmacological information systems. Such a connection would have been needed both to contribute information on suggested drugs to the decision process but also to allow potentially automatic checks for drug-drug interactions between the recommended drug and earlier prescribed drugs

for a given patient.

To be helpful during drug prescription, such information must be available at the point of care [24, 25, 118]. Clinical workstations can provide CDSS but require bedside computers for every patient to be available at the point of care. Home visits do need some form of mobile device to facilitate access to patient information and decision support.

Early mobile devices used in medicine decision support were quite limited in hardware and screen size. Additionally, specialized software development environments were necessary but in comparison to those used in personal computer systems, these environments were rather poor. Devices were also large and cumbersome to bring. Naturally, hardware and also software development environments have continued to improve. The same can be said for data transfer speeds. Specialized protocols for mobiles such as the Wireless Application Protocol (WAP) are no longer necessary (although still used).

Paper II also discusses issues related to mobile decision support systems. Related work shows that integration with EPR is quite useful, in particular when several patient attributes are used as input to the CDSS (eliminating the risk of data entry errors). However, EPR integration is not uncomplicated and often requires specially adapted software. Patient information is also sensitive information (see Section 2.3). Simpler information look-up systems can therefore still be useful.

Additionally, it must be possible to identify drugs without ambiguity. The Anatomic Therapeutic Chemical (ATC) classification system is a terminology system for drugs. Codes in this system are constructed according to the organ system for which the drug is intended, the main indication of use (therapeutic use) and chemical class.

We used a Swedish database for drug information as a base for developing several prototypes for drug information. In particular, we looked at drug-drug interaction information from the database. The first prototype is an application for mobile devices which lets users search for interactions, barcodes and prices. The database is installed on a server and the mobile device connects to this database to get information.

The second prototype is a server application and development kit (PharmaJ) [67]. The application encodes information from the database in a chosen markup-language. Installation is not required on the client-side since a normal web-browser can be used. Information is transformed to WAP or to HTML, depending on the client.

4.1.3 Paper III

Given proper educational measures and integration with clinical information systems, decision support systems can have an important impact on clinical practice (for example on drug subscription [107]) and guideline adherence [80]. Paper III discusses implementation and maintenance aspects of three types of decision support systems.

The system in [106] was directly based on a clinical guideline for pharmacological hypertension treatment. However, rules in the decision support system had to be manually extracted by domain experts based on the guideline text. Any updates to the clinical guideline would mean that the process has to be repeated for the parts that changed.

Early diagnosis of dementia is important as it enables more possibilities in the treatment options (including drugs which can slow down the progression of the disorder). Dementia is complicated to diagnose as many and diverse symptoms can be important. In fact, only around 25% of dementia cases are detected during the first appointment with a doctor. The rule base for the prototype system was organized in frames and forms with attributes based, in part, on guidelines complemented with additional rules for treatment adapted to the local clinic.

The third system provides drug information lookup via mobile devices (phones). To cope with limited transmissions speed, the system tries to minimize the amount of information communicated. Although this system does not use patient specific information, it can nevertheless support drug prescription during home visits by doctors.

4.1.4 Paper VI and VII

In Paper VI, we describe a typical classification system for pharmacological information, the ATC classification. This classification organizes drugs according to their anatomic, therapeutic, pharmacological and chemical group. We used a formal description [38] of rough sets using category theory to represent drug-drug interactions. The need to extend the traditional view of rough sets is not only interesting from a theoretical point of view but is an advantage in many application domains, including medicine. Information representation based on medical ontologies is usually rather narrow and oriented towards crisp specifications of data. At the same time, many applications in health care call for representation of vagueness and uncertainties. In the paper, we discuss various fields of health care and possible uses of generalized rough sets in the area of decision support and, more specifically, treatment decisions. Pharmacological treatment is a good example where generalized rough sets can be applied to drug-drug interactions.

In Paper VII, we provide additional background descriptions of the categorical framework, such as categories, functors, natural transformations and monads. Additionally, we include a short discussion of possible application to differential diagnosis of dementia.

Drug-drug interaction

Prescriptions of drugs is not an easy task and should take into account previously prescribed drugs (stored in the EPR). This can be illustrated with a simplified example where we assume the following situation: "*An 80 year old patient with reduced kidney function and asthmatic is diagnosed with hypertension. The CDSS for hypertension treatment from [106] recommends, based on those parameters, a treatment based on a furosemide drug frequently used for treating this condition. A second, alternative, treatment to consider is an enalapril drug. Which treatment option is the most convenient for this patient?*"

E	furosemide	enalapril	gentamicin
furosemide	1	α	β
enalapril	α	1	γ
gentamicin	β	γ	1

Table 4.1: Similarities representing drugs compatibilities

By studying the patient journal we can see that the patient is also taking antibiotics, a gentamicin drug which has known interactions with furosemide. The gentamicin drug must be maintained. Based on this knowledge, the doctor will take the decision to prescribe an enalapril drug.

This situation can be illustrated using a similarity relation E, on sets of drugs rather than just drugs, based on compatibilities (non-interactions) between different drugs, as shown in Table 4.1.

The current medication of a patient is represented by a set of drugs. In this case, the situation for our 80 year old patient is $P = \{gentamicin\}$. The treatments are represented by the sets of drugs considered. The two possible treatments are denoted by

$$T_1 = \{gentamicin, enalapril\}$$

and

$$T_2 = \{gentamicin, furosemide\}.$$

To see the compatibility of the two possible treatments, we study the similarity of the drug from the EPR and each of the treatments. With $\gamma > \beta = \alpha$ and

$$E(P, T_1) = \gamma$$

$$E(P, T_2) = \beta$$

we should select T_1 as a basis for pharmacologic treatment of the patient.

Naturally, the situation in this example is simplified. The quantification of degree of interaction is unclear. Qualifications have, however, been suggested. For drugs showing therapeutically significant interactions, it is important to distinguish between different types of interactions:

- recommended combination
- neutral combination (no harmful interactions)
- risky combination (should be monitored)
- dangerous combination (should be avoided)
- possible interaction (not tested)

In this qualification it is clear that a corresponding linear quantification is not straightforward. Furthermore, drugs are affected in different ways:

- no change in effect
- increases effect
- reduces effect
- other (e.g. a new type of side effect)

Interaction type and effect need to be considered in the guideline for respective treatments.

Future extensions

In [35], it is shown how *signatures* of a logic can be applied to assessment of service levels for dementia patient care. The backbone of most decisions regarding dementia care support is provided by assessment scales (such as MMSE [45] or GDS [152]). The paper argues that not only should uncertainty in values (from assessment scales) be taken into account but also the uncertainty in the process that decided the values. For example, an assigned value in GDS-4 has a certain degree of uncertainty in itself but the uncertainty of the process is different if the value was assessed by a domain expert rather than by a novice.

A quantification of degree of drug-drug interaction is suggested in Paper VI but it does not take into account frequencies of the interactions (i.e. uncertainties). Ideas from [35] provide inspiration for future extensions in this aspect.

4.2 Processes in health care and bioinformatics

Papers IV-V represent the second stage of the thesis work and are related to service discovery and workflows in bioinformatics.

4.2.1 Paper IV

Magallanes is a software which can discover web services using service descriptions and a scoring system based on the number of occurrences and relative word positions of hits. Currently it has AND, OR and regular expression operators. The default search space is BioMOBY services and data types but others are possible.

The software initially searches for perfect matches between the keywords and the metadata descriptors. When no hits are found, it uses an approximate regular expression matching. This done with the *Levenshtein distance* defined between two strings (not necessarily having the same length) as the minimum number of insertions, deletions, and substitutions of characters required to transform a string into another. When a solution is not reached, Magallanes returns a "did you mean" suggestion with the closest combination of keywords similar to the query.

The results are also ranked according to "feed back"-values which are really historical records of user selections (or the "popularity" of the keywords/resource combination). It is possible to increment the feedback value associated to the keyword-resource combination. This value can be decreased when the user selects another resource using the same keyword.

It is also possible to configure a learning rate which adjusts the rate of changes in feedback values. For example in a shared environment with several users, changes in the feedback values should be smaller than for single-user application scenarios.

The algorithm (Levenshtein distance) used for the approximated matching against the service descriptions is certainly not new but the application to service discovery in bioinformatics is, however, novel. In general, the textual descriptions of the services are of high quality and thus lend themselves to the usage of such algorithms. An

alternative approach is to base service discovery on semantic metadata (see Section 3.3.1). However, for these techniques to be successful, services have to be properly annotated and the ontologies have to be useful and well defined. It is possible to use Magallanes for different service catalogues and the feasibility of using semantics as a base for service discovery and composition varies between catalogues.

Initial efforts have been focused on BioMOBY services. These services are annotated with semantics in the sense that services are associated with a *service type* which, in essence, is a functional description and service parameters are associated with a data type ontology which both has a semantic meaning and is also used to derive syntax. These semantics could be used to restrict the results from service discovery but it is important to note that the ontologies were created as a result of a community effort. There is no curation or oversight that ensures that the ontologies are meaningful (although this is, of course, in the interest of all developers).

Some examples of the current problems of the ontologies can be found in the data type ontology. There are sections which are well-defined (for example, sequences). This is not always the case. Several data formats in bioinformatics are previously defined (for example, formats used in sequence databases, FASTA etc.) and the object ontology does not describe those formats well. In fact, FASTA is simply described as plain text. The reason is that the data type ontology is also used in BioMOBY to define the syntax (i.e. the XML structure). Since FASTA is not XML based, the only possible way to describe FASTA is to annotate it as plain text and not specify the syntax further.

Another issue is the use of the base class (called *Object*) in BioMOBY. Sequence retrieval services from databases typically require an identifier (of the sequence in the database) and what the identifier represents. The consensus is that such information is transferred via data formatted as an *Object* (which has suitable fields). Consequently, the input parameter for sequence retrieval services are specified to be of the data type *Object*. The problem is that the BioMOBY protocol for discovery of services specifies that compatible services are those services whose parameters data types match the

data type of the data directly or by inheritance. For example, if the user already has a sequence (i.e. the data type inherits from the sequence base class), the sequence *retrieval* services will also be considered to be compatible with this data (since the data type of the sequence is a *Object* by inheritance). Strictly, the sequence retrieval services are able to parse the input data but sending a complete sequence to retrieve a sequence when only the identifier is needed makes no sense.

The motivation to *initially* use the textual descriptions as a base for BioMOBY service discovery is the current state of the ontologies in BioMOBY. If the ontologies were improved, or complemented with other semantic annotations, such an approach would be a suitable extension of Magallanes.

Magallanes uses the data type ontology for service composition (workflow generation), since that ontology is the base for service integration in BioMOBY. Magallanes considers possible service compositions based on user selection of input and output data types. In some cases, several alternative compositions are possible. In the standalone application, users can select one path. Magallanes can export the generated workflow as a Taverna workflow, thereby making it compatible with MOWServ where generated workflows can be shared and enacted (see Paper V).

4.2.2 Paper V

In [72], we report functionality to share and enact workflows in a generic platform (MOWServ [94]) for integration of bioinformatics tools. MOWServ is based on the BioMOBY standard and adds client-type features such as persistence of results (data) in separate user accounts, service invocation and monitoring through automatically generated user interfaces. However, many typical tasks required combining several services. One such simple but repetitive task is to retrieve a sequence from an external database and compare against databases of known sequences (typically using a variation of the BLAST algorithm [4]) and finally to select the best hits (the most similar sequences to the input sequence). This requires invocation of three different web-services. By sharing previously defined workflows, such repetitive steps could be

automated.

Taverna [99] is a software for creating and executing data intensive, in-silico experiments in molecular biology. The workflows are stored in a format called the Simplified Conceptual workflow language (Scufl). The application, Taverna, provides a graphical user interface where workflow designers can compose web services (and other executable components) by explicitly designing the data flow.

Taverna workflows are the de-facto standard in the bioinformatics community for workflows. Workflows defined using the Scufl format consist of several types of concepts. *Processors* are the main elements of the workflows because they represent the executable components. Processors can represent SOAP web-services described in WSDL, sub-workflows (nested workflows), constant values, local functions/scripts and specialized services such as the BioMOBY service standard. *Links* are the elements that connect outputs and inputs of processor executions. *Sources* are the inputs of workflows and *sinks* are the outputs. Coordination *constraints* allow users to define conditions that must be carried out to execute specific processors.

BioMOBY web services are very suitable for service composition (combination) into workflows since data formats are shared among services, making it possible to send output data from a service to any service that has been declared as compatible with the data format of the output data.

Workflows constructed with Taverna can be uploaded to the workflow repository of MOWServ where the workflow is validated (this process involves checking that the services exist and that input/outputs are indeed compatible). The automatically generated workflows (in Scufl format) from Paper IV are possible to upload and enact in MOWServ. Additional documentation of the workflow is encouraged before the workflow is publicly available. Users may later browse the public workflows and view documentation or even download the workflow definition. Moreover, the workflows can be enacted (executed) directly in the web browser without the need for installing any additional software. Results can be visualized, downloaded or used as input to other services/workflows.

The paper also specifies requirements for sharing of workflows, such as workflow discovery and annotation (see also Section 3.3.1), quality assurance policies, workflow enactment and monitoring of enactments, workflow documentation and (automatic) validation.

The paper concludes with some unsolved but important problems which makes it difficult to share workflows efficiently. The first problem is related to workflow discovery and annotation. The use of ontologies is crucial but, on the other hand, also requires substantial work to develop and ensure that all workflows are properly annotated. The approach in [117] is a step towards a solution if workflows are considered as abstract and composite services. However, that approach employs text-mining over service descriptions instead of further annotating services.

The second problem is related to the fact that services are external components and may temporarily or permanently become unavailable (service hardware problems, network problems, etc.). Service availability rates (and thereby, in extension, workflow functionality) can be measured and taken into account when presenting available workflows to users. Since been, MOWServ automatically tests services for availability and functionality using, if specified, input and/or output examples. This has not been integrated in the workflow repository but is available in the service tree of MOWServ where non-functioning services are marked.

Bibliography

[1] A-WARE: 2010, 'http://www.a-ware-project.eu/'.

[2] Abrams, P., L. Cardozo, M. Fall, D. Griffiths, P. Rosier, U. Ulmsten, P. Van Kerrebroeck, A. Victor, and A. Wein: 2002, 'Standardisation Sub-committee of the International Continence Society. The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society'. *Neurourol Urodyn* **21**(2), 167–78.

[3] Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, et al.: 1996, 'Fast discovery of association rules'. *Advances in knowledge discovery and data mining* **12**, 307–328.

[4] Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman: 1990, 'Basic local alignment search tool'. *Journal of molecular biology* **215**(3), 403–410.

[5] American Psychiatric Association. Task Force on DSM-IV: 1994, *DSM-IV: diagnostic and statistical manual of mental disorders*. American Psychiatric Association Washington, DC.

[6] Anagnostou, T., M. Remzi, M. Lykourinas, and B. Djavan: 2003, 'Artificial Neural Networks for Decision-Making in Urologic Oncology'. *European Urology* **43**(6), 596–603.

[7] Anzbock, R. and S. Dustdar: 2005, 'Semi-automatic generation of Web services

and BPEL processes-A Model-Driven approach'. *Lecture Notes in Computer Science* **3649**, 64.

[8] Arning, M., N. Forgó, and T. Krügel: 2007, 'Data protection issues with regard to research in genetic data'. In: *2nd Workshop on Personalization for e-Health*. pp. 11–19.

[9] Arpinar, I., J. Miller, and A. Sheth: 2001, 'An efficient data extraction and storage utility for XML documents'. In: *Proceedings of 39th Annual ACM Southeast Conference, Athens, GA*. pp. 293–295.

[10] Arpinar, I., R. Zhang, B. Aleman-Meza, and A. Maduko: 2005, 'Ontology-driven web services composition platform'. *Information Systems and E-Business Management* **3**(2), 175–199.

[11] Banhart, F. and H. Klaeren: 1995, 'A graphical query generator for clinical research databases.'. *Methods of information in medicine* **34**(4), 328–339.

[12] Bellazzi, R., C. Larizza, P. Magni, S. Montani, and G. D. Nicolao: 1999, 'Intelligent Analysis of Clinical Time Series by Combining Structural Filtering and Temporal Abstractions'. *Lecture Notes in Computer Science* **1620**, 261–270.

[13] Bemmel, J. V.: 1984, 'The Structure of Medical Informatics'. *Medical Informatics* **9**, 175–180.

[14] Berners-Lee, T., J. Hendler, and O. Lassila: 2001, 'The Semantic Web: Scientific American'. *Scientific American*.

[15] BioMOBY community: 2008, 'BioMOBY Asynchronous Services'. `http://biomoby.open-bio.org/CVS_CONTENT/moby-live/Docs/asyncDocs/BioMOBY%20Asynchronous%20Service%20Specification%20v2.4.2.pdf`.

[16] Bodenreider, O.: 2004, 'The unified medical language system (UMLS): integrating biomedical terminology'. *Nucleic Acids Research* **32**(Database Issue), D267–270.

[17] Brossette, S., A. Sprague, J. Hardin, K. Waites, W. Jones, and S. Moser: 1998, 'Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance'. *Journal of the American Medical Informatics Association* **5**(4), 373–381.

[18] Bury, J., J. Fox, and D. Sutton: 2001, 'The PROforma guideline specification language: progress and prospects'. *Computer-based support for clinical guidelines and protocols: proceedings of EWGLP 2000* p. 13.

[19] Cardoso, J. and A. Sheth: 2003, 'Semantic e-workflow composition'. *Journal of Intelligent Information Systems* **21**(3), 191–225.

[20] Chute, C.: 2000, 'Clinical Classification and Terminology'. *Journal of the American Medical Informatics Association* **7**(3), 298–303.

[21] Chute, C. and D. Koo: 2002, 'Public Health, Data Standards, and Vocabulary: Crucial Infrastructure for Reliable Public Health Surveillance'. *Journal of Public Health Management Practice* **8**(3), 11–17.

[22] Cios, K. and G. Moore: 2002, 'Uniqueness of Medical Data Mining'. *Artificial Intelligence in Medicine* **26**(1-2), 1–24.

[23] Claerhout, B., N. Forgo, T. Krügel, M. Arning, and G. De Moor: 2008, 'A Data Protection Framework for Transeuropean genetic research projects'. *Studies in health technology and informatics* **141**, 67.

[24] Clark, I., B. McCauley, I. Young, P. Nightingale, M. Peters, N. Richards, and D. Abu: 1999, 'Electronic Drug Prescribing and Administration - Bedside Medical Decision Making'. *Lecture Notes in Articificial Intelligence - Artificial Intelligence in Medicine* **1620**, 143–147.

[25] Clauson, K., M. Seamon, A. Clauson, and T. Van: 2004, 'Evaluation of drug information databases for personal digital assistants'. *American Journal of Health-System Pharmacy* **61**, 1015–1024.

[26] Coalition, D., A. Ankolekar, M. Burstein, and J. Hobbs: 2002, 'DAML-S: Web service description for the semantic Web'. In: *First International Semantic Web Conference, Sardinia, Italy.*

[27] Cohen, A., P. Stavri, and W. Hersh: 2004, 'A categorization and analysis of the criticisms of Evidence-Based Medicine'. *International Journal of Medical Informatics* **73**, 35–43.

[28] Davidoff, F., B. Haynes, D. Sackett, and R. Smith: 1995, 'Evidence based medicine'. *British Medical Journal* **310**, 1085–1086.

[29] de Keizer, N., A. Abu-Hanna, and J. Zwetsloot-Schonk: 2000, 'Understanding Terminological Systems I: Terminology and Typology'. *Methods of Information in Medicine* **39**, 16–21.

[30] Deelman, E., D. Gannon, M. Shields, and I. Taylor: 2009, 'Workflows and e-Science: An overview of workflow system features and capabilities'. *Future Generation Computer Systems* **25**(5), 528 – 540.

[31] Doctorow, C.: 2008, 'Big data: Welcome to the petacentre'. *Nature* **455**, 16–21.

[32] Dogac, A., G. Laleci, S. Kirbas, Y. Kabak, S. Sinir, A. Yildiz, and Y. Gurcan: 2006, 'Artemis: deploying semantically enriched web services in the healthcare domain'. *Information Systems* **31**(4-5), 321–339.

[33] Dolin, R., L. Alschuler, F. Behlen, P. Biron, S. Boyer, L. Harding, T. Lincoln, J. Mattison, W. Rishel, R. Sokolowski, J. Spinosa, and J. Williams: 1999, 'HL7 Document Patient Record Architecture: An XML Document Architecture Based on a Shared Information Model'. In: *Proceedings AMIA Symposium.* pp. 52–56.

[34] Einbinder, J., K. Scully, R. Pates, J. Schubart, and R. Reynolds: 2001, 'Case study: a data warehouse for an academic medical center'. *Journal of Healthcare Information Management* **15**(2), 165–176.

[35] Eklund, P.: 2010, 'Signatures for assessment, diagnosis and decision-making in ageing'. In: *IPMU 2010 (accepted)*.

[36] Eklund, P., S. Eriksson, J. Karlsson, H. Lindgren, and A. Näslund: 2000, 'Software Development and Maintenance Strategies for Guideline Implementation'. In: *Proc. EUNITE-Workshop "Intelligent Systems in Patient Care"*. pp. 26–45.

[37] Eklund, P., J. Forsström, A. Holm, M. Nyström, and G. Selén: 1994, 'Rule generation as an alternative to knowledge acquisition: a systems architecture for medical informatics'. *Fuzzy Sets and Systems* **66**(2), 195–205.

[38] Eklund, P. and M. Galán: 2006, 'Monads can be rough'. *Lecture Notes in Computer Science* **4259**, 77.

[39] Eklund, P., M. Johansson, J. Karlsson, and R. Åström: 2009, 'BPMN and its Semantics for Information Management in Emergency Care'. In: S. Sohn, K. D. Kwack, K. Um, G. Y. Lee, and F. Ko (eds.): *Proceedings of Fourth International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2009)*. pp. 273–279.

[40] Eklund, P. and J. Karlsson: 2003, 'Simulations of workflows in radiation therapy'. In: G. Surján, R. Engelbrecht, and P. McNair (eds.): *Proc. Medical Informatics Europe '03 (CD version)*.

[41] Eklund, P., J. Karlsson, and A. Näslund: 2006, 'Mobile Pharmacology'. In: M. S. Szczuka, D. Howard, D. Slezak, H.-K. Kim, T.-H. Kim, I. S. Ko, G. Lee, and P. M. A. Sloot (eds.): *ICHIT*, Vol. 4413 of *Lecture Notes in Computer Science*. pp. 522–533.

[42] Eklund, P., J. Karlsson, J. Rauch, and M. Simunek: 2005, 'Computational Coronary Artery Bypass Grafting'. In: *ICCIMA '05: Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications*. Washington, DC, USA, pp. 138–144.

[43] Ennett, C., M. Frize, and C. Walker: 2001, 'Influence of Missing Values on Artificial Neural Network Performance'. In: V. P. et al (ed.): *Proc. Medinfo 2001*. pp. 449–453.

[44] Field, M. and K. Lohr: 1990, *Clinical practice guidelines: directions for a new program*, Chapt. Attributes of good practice guidelines, pp. 53–77. National Academy Press.

[45] Folstein, M. F., S. E. Folstein, and P. R. McHugh: 1975, 'Mini-mental state : A practical method for grading the cognitive state of patients for the clinician'. *Journal of Psychiatric Research* **12**(3), 189 – 198.

[46] Foster, I. T.: 2005, 'Globus Toolkit Version 4: Software for Service-Oriented Systems.'. In: H. Jin, D. A. Reed, and W. Jiang (eds.): *NPC*, Vol. 3779 of *Lecture Notes in Computer Science*. pp. 2–13.

[47] Franzén, L., M. Karlsson, G. Duvefjäll, P. Eklund, I. Lax, K. Nordlinder, I. Näslund, C. Svedberg, and B. Zackrisson: 1999, 'Information Technology in Radiation Therapy (abstract)'. ESTRO Meeting in Physics for Clinical Radiotherapy, Germany.

[48] Galán García, M.: 2004, 'Categorical Unification'. PhD Dissertation, Department of Computing Science, University of Umeå.

[49] García, M., J. Karlsson, and O. Trelles: 2009, 'Web-services across an European Biomedical Grid Infrastructure'. In: *Book of abstracts. IX Jornadas de Bioinformática. The 9th symposium on Bioinformatics and Computational Biology.* p. 88.

[50] Garcia Castro, A., S. Thoraval, L. Garcia, and M. Ragan: 2005, 'Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator'. *BMC Bioinformatics* **6**(1), 87.

[51] Gekas, J. and M. Fasli: 2005, 'Automatic web service composition based on graph network analysis metrics'. *Lecture notes in computer science* **3761**, 1571.

[52] Goble, C. and D. De Roure: 2008, 'Curating Scientific Web Services and Workflows'. *Educause Review* **43**(5).

[53] Grimshaw, J. and I. Russell: 1993, 'Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations'. *The Lancet* **342**(8883), 1317–1322.

[54] Hájek, P. and T. Havránek: 1978, *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory*. Springer-Verlag.

[55] Hájek, P., M. Holeňa, and J. Rauch: 2010, 'The GUHA method and its meaning for data mining'. *J. Comput. Syst. Sci.* **76**(1), 34–48.

[56] Hammond, W.: 1991, 'Health Level 7: an application standard for electronic medical data exchange.'. *Topics in health record management* **11**(4), 59.

[57] Hernandez, T. and S. Kambhampati: 2004, 'Integration of biological sources: current systems and challenges ahead'. *ACM SIgmod Record* **33**(3), 51–60.

[58] Hoheisel, J.: 2006, 'Microarray technology: beyond transcript profiling and genotype analysis'. *Nature reviews genetics* **7**(3), 200–210.

[59] Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. Hill, R. Kania, M. Schaeffer, S. St Pierre, et al.: 2008, 'Big data: the future of biocuration'. *Nature* **455**, 47–50.

[60] Howison, J., A. Wiggins, and K. Crowston: 2008, 'eResearch Workflows for Studying Free and Open Source Software Development'. In: *Open Source Development, Communities and Quality: IFIP 20th World Computer Congress, Working Group 2.3 on Open Source Software, September 7-10, 2008, Milano, Italy*. p. 405.

[61] Inmon, W.: 1995, 'Tech topic: What is a data warehouse?'. `http://www.cait.wustl.edu/cait/papers/prism/vol1_no1`.

[62] Jarke, M., M. Lenzerini, Y. Vassiliou, and P. Vassiliadis: 2003, *Fundamentals of data warehouses*. Springer Verlag.

[63] Johnson, P., S. Tu, and N. Jones: 2001, 'Achieving reuse of computable guideline systems'. *Stud Health Technol Inform* **84**, 99–103.

[64] Jordan, D., J. Evdemon, A. Alves, A. Arkin, S. Askary, C. Barreto, B. Bloch, F. Curbera, M. Ford, Y. Goland, et al.: 2007, 'Web services business process execution language version 2.0'. *OASIS Standard* **11**.

[65] Kalyanaraman, A., S. Emrich, P. Schnable, and S. Aluru: 2006, 'Assembling genomes on large-scale parallel computers'. In: *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*. pp. 10 pp.–.

[66] Kamel, M. and M. Zviran: 1991, 'Heterogeneous databases integration in a hospital information systems environment: a bottom-up approach'. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. p. 363.

[67] Karlsson, J.: 2008, 'Interface for Accessing Pharmacological Information'. In: *CBMS '08: Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*. Washington, DC, USA, pp. 176–178.

[68] Karlsson, J. and P. Eklund: 2000, 'Data mining and structuring of executable data analysis reports: Guideline development in a narrow sense'. In: A. Hasman, B. Blobel, J. Dudeck, R. Engelbrecht, G. Gell, and H. Prokosch (eds.): *Proc. Medical Informatics Europe '00*. pp. 790–794.

[69] Karlsson, J. and P. Eklund: 2001, 'Workflow Design as a Basis for Component Interaction'. In: *Proc. Medinfo 2001*. pp. 1158–1160.

[70] Karlsson, J. and P. Eklund: 2002, 'A one-step approach to data retrieval, analysis and documentation'. In: G. Surján, R. Engelbrecht, and P. McNair (eds.): *Proc. Medical Informatics Europe '02*. pp. 320–324.

[71] Karlsson, J., P. Eklund, C.-G. Hallgren, and J.-G. Sjödin: 1999, 'Data warehousing as a basis for web-based documentation of data mining and analysis'. In: P. Kokol, B. Zupan, J. Stare, M. Premik, and R. Engelbrecht (eds.): *Proc. Medical Informatics Europe '99*. pp. 423–427.

[72] Karlsson, J., S. Ramirez, J. Aldana-Montes, and O. Trelles: 2008, 'Workflow repositories for bioinformatics'. Technical report, Department of Computer Architecture, Malaga University.

[73] Komorowski, J., Z. Pawlak, L. Polkowski, and A. Skowron: 1999, 'Rough sets: A tutorial'. *Rough fuzzy hybridization: A new trend in decision-making* pp. 3–98.

[74] Kopeckỳ, J., T. Vitvar, C. Bournez, and J. Farrell: 2007, 'SAWSDL: Semantic Annotations for WSDL and XML Schema'. *IEEE Internet Computing* pp. 60–67.

[75] Kosara, R. and S. Miksch: 2001, 'Visualizing Complex Notions of Time'. In: *Proc. Medinfo 2001*. pp. 211–215.

[76] Kusiak, A., J. Kern, K. Kerstine, K. McLaughlin, and T. Tseng: 2000a, 'Autonomous Decision-Making: A Data Mining Approach'. *IEEE Transactions on Information Technology in Biomedicine* **4**(4), 274–284.

[77] Kusiak, A., K. Kerstine, J. Kern, K. McLaughlin, and T. Tseng: 2000b, 'Data Mining: Medical and Engineering Case Studies'. In: *Proceedings of the Industrial Engineering Research 2000 Conference*. pp. 1–7.

[78] Lee, S., T. D. Wang, N. Hashmi, and M. P. Cummings: 2007, 'Bio-STEER: A Semantic Web workflow tool for Grid computing in the life sciences'. *Future Gener. Comput. Syst.* **23**(3), 497–509.

[79] Liu, C., X. Lin, X. Zhou, and M. Orlowska: 2000, 'Repository Support for Workflow Management Systems'. *Special Issue on Cooperative Systems of Journal of Applied Systems Studies* **1**(3).

[80] Lobach, D. and W. Hammond: 1994, 'Development and evaluation of a Computer-Assisted Management Protocol (CAMP): improved compliance with care guidelines for diabetes mellitus.'. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. p. 787.

[81] Luscombe, N., D. Greenbaum, and M. Gerstein: 2001, 'What is bioinformatics? A proposed definition and overview of the field'. *Methods of information in medicine* **40**(4), 346–358.

[82] Majithia, S., D. Walker, and W. Gray: 2004, 'A framework for automated service composition in service-oriented architectures'. *Lecture Notes in Computer Science* pp. 269–283.

[83] Martin, D., M. Paolucci, S. McIlraith, M. Burstein, D. McDermott, D. McGuinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, et al.: 2005, 'Bringing semantics to web services: The OWL-S approach'. *Lecture Notes in Computer Science* **3387**, 26–42.

[84] Martin-Sanchez, F., I. Iakovidis, S. Nørager, V. Maojo, P. de Groen, J. V. der Lei, T. Jones, K. Abraham-Fuchs, R. Apweiler, A. Babic, R. Baud, V. Breton, P. Cinquin, P. Doupi, M. Dugas, R. Eils, R. Engelbrecht, P. Ghazal, P. Jehenson, C. Kulikowski, K. Lampe, G. D. Moor, S. Orphanoudakis, N. Rossing, B. Sarachan, A. Sousa, G. Spekowius, G. Thireos, G. Zahlmann, J. Zvárová, I. Hermosilla, and F. J. Vicente: 2004, 'Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care'. *Journal of Biomedical Informatics* **37**(1), 30 – 42.

[85] McCusker, J., J. Phillips, A. Beltran, A. Finkelstein, and M. Krauthammer: 2009, 'Semantic web data warehousing for caGrid'. *BMC Bioinformatics* **10**(Suppl 10), S2.

[86] Medical Products Agency: 2005, 'Utbytbara läkemedel 2005-01-20'.

[87] Mildenberger, P., M. Eichelberg, and E. Martin: 2002, 'Introduction to the DICOM standard.'. *European radiology* **12**(4), 920.

[88] Møller, L., G. Lose, and T. Jørgensen: 2001, 'The prevalence and bothersomeness of lower urinary tract symptoms in women 40-60 years of age'. *Acta Obstetricia et Gynecologica Scandinavica* **79**(4), 298–305.

[89] Moulin, C. and M. Sbodio: 2005, 'Using ontological concepts for Web service composition'. In: *Proceedings of The*. pp. 487–490.

[90] Müller, R., O. Thews, C. Rohrbach, M. Sergl, and K. Pommerening: 1996, 'A Graph-Grammar Approach To Represent Casual, Temporal And Other Contexts In An Oncological Patient Record'. *Methods Of Information In Medicine* **35**(2), 127–144.

[91] Musen, M.: 2001, 'Creating and Using Ontologies: What Informatics is All About'. In: *Proc. Medinfo 2001*. p. 1514. Keynote adress.

[92] National Institutes of Health: 1997, 'The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure'. *Arch Intern Med* **157**(21), 2413–2446.

[93] National Library of Medicine: 2005, 'Medical Subject Headings'. Technical report, U.S. Government Printing Office.

[94] Navas-Delgado, I., M. d. M. Rojano-Munoz, S. Ramirez, A. J. Perez, E. Andres Leon, J. F. Aldana-Montes, and O. Trelles: 2006, 'Intelligent client for integrating bioinformatics services'. *Bioinformatics* **22**(1), 106–111.

[95] Neeb, J., M. Schlundt, and H. Wedekind: 2000, 'Repositories for workflowmanagement-systems in a middleware environment'. In: *Proceedings of the 33rd International Conference on System Sciences*.

[96] OASIS: 2005, 'Asynchronous Service Access Protocol'. `http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=asap`.

[97] OASIS: 2006, 'Web Services Resource Framework'. `http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf`.

[98] Ohno-Machado, L. et al.: 1998, 'The GuideLine Interchange Format - A Model for Representing Guidelines'. *J Am Med Inform Assoc* **5**(4), 357–372.

[99] Oinn, T., M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li: 2004, 'Taverna: a tool for the composition and enactment of bioinformatics workflows.'. *Bioinformatics* **20**(17), 3045–3054.

[100] Olhede, T.: 1996, 'Can concepts and terms in existing classifications satisfy the clinical demands on electronic health care records?'. In: *Proc. Medical Informatics Europe '96*. pp. 484–487.

[101] Oster, S., S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. Covitz, K. Shanbhag, et al.: 2008, 'caGrid 1.0: an enterprise Grid infrastructure for biomedical research'. *Journal of the American Medical Informatics Association* **15**(2), 138–149.

[102] Paolucci, M., N. Srinivasan, K. Sycara, and T. Nishimura: 2003, 'Towards a semantic choreography of web services: From WSDL to DAML-S'. In: *Proceedings of ICWS03*. pp. 22–26.

[103] Pawlak, Z.: 1982, 'Rough Sets'. *International Journal of Information and Computer Sciences* **11**(5), 341–356.

[104] Pearson, S., C. Margolis, S. Davis, L. Schreier, H. Sokol, and L. Gottlieb: 1995, 'Is consensus reproducible? A study of an algorithmic guidelines development process'. *Medical Care* **33**(6), 643–660.

[105] Peleg, M., S. Tu, J. Bury, P. Ciccarese, J. Fox, R. Greenes, R. Hall, P. Johnson, N. Jones, A. Kumar, et al.: 2003, 'Comparing computer-interpretable guideline models: a case-study approach'. *Journal of the American Medical Informatics Association* **10**(1), 52.

[106] Persson, M., J. Bohlin, and P. Eklund: 2000a, 'Development and maintenance of guideline based decision support for pharmacological treatment of hypertension'. *Computer Methods and Programs in Medicine* **61**, 209–219.

[107] Persson, M., T. Mjörndal, B. Carlberg, J. Bohlin, and L. Lindholm: 2000b, 'Evaluation of a computer-based decision support system for treatment of hypertension with drugs: retrospective, nonintervention testing of cost and guideline adherence'. *Journal of Internal Medicine* **247**, 87–93.

[108] Prather, J., D. Lobach, L. Goodwin, J. Hales, M. Hage, and W. Hammond: 1997, 'Medical data mining: knowledge discovery in a clinical data warehouse.'. In: *Proceedings of the AMIA Annual Fall Symposium*. p. 101.

[109] Pryor, A. and G. Hripcsak: 1993, 'The arden syntax for medical logic modules'. *International Journal of Clinical Monitoring and Computing* **10**(4), 215–224.

[110] Quaglini, S., C. Mossa, C. Fassino, M. Stefanelli, A. Cavallini, and G. Micieli: 1999, 'Guidelines-Based Workflow Systems'. *Lecture Notes In Computer Science; Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making* **1620**, 65–75.

[111] Quaglini, S., M. Stefanelli, G. Lanzola, V. Caporusso, and S. Panzarasa: 2001, 'Flexible guideline-based patient careflow systems'. *Artificial Intelligence in Medicine* **22**(1), 65–80.

[112] Ramirez, S., J. Karlsson, M. García, J. Rios-Perez, and O. Trelles: 2008, 'A flexible framework for the design of knowledge-discovery clients'. In: *Proceedings of the International Conference on Telecommunications and Multimedia (TEMU 2008)*.

[113] Ramirez, S., J. Karlsson, J. Rodríguez, R. Royo, and O. Trelles.: 2006, 'Mirroring biomoby services'. In: *VII Spanish Bioinformatics Symposium*.

[114] Ramírez, S., J. Karlsson, M. García, and O. Trelles: 2008, 'Metadata repositories for web-services and workflows'. In: *VIII Jornadas de Bioinformática)*. pp. 50–56.

[115] Reisig, W.: 1985, *Petri Nets: An Introduction*. Springer-Verlag.

[116] Rios-Perez, J., J. Karlsson, and O. Trelles: 2009, 'Victoria: Navigating to a new style of searching for web-services and workflows'. In: *17th Annual International Conference on intelligent systems for molecular biology (ISMB) and 8th European Conference on Computational Biology*.

[117] Ríos, J., J. Karlsson, and O. Trelles: 2009, 'Magallanes: a web services discovery and automatic workflow composition tool'. *BMC Bioinformatics* **10**(1), 334.

[118] Rothschild, J., T. Lee, T. Bae, and D. Bates: 2002, 'Clinician Use of a Palmtop Drug Reference Guide'. *Journal of the American Medical Informatics Association* **9**(3), 223–229.

[119] Rothwell, D., R. Cote, J. Cordeau, and M. Boisvert: 1993, 'Developing a standard data structure for medical language - the SNOMED proposal'. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care. Symposium on Computer Applications in Medical Care*. p. 695.

[120] Salgado, N. C. and A. Gouveia-Oliveira: 2000, 'Towards a common framework for clinical trials information systems'. In: *Proc AMIA Symp*. pp. 754–758.

[121] Sander, C.: 2000, 'Genomic Medicine and the Future of Health Care'. *Science* **287**(5460), 1977–1978.

[122] Shahar, Y., S. Miksch, and P. Johnson: 1998, 'The ASGAARD Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines'. In: *Artificial Intelligence in Medicine*. pp. 29–51.

[123] Shahar, Y. and M. Musen: 1996, 'Knowledge-Based Temporal Abstraction in Clinical Domains'. *Artificial Intelligence in Medicine* **8**(3), 267–298.

[124] Shendure, J. and H. Ji: 2008, 'Next-generation DNA sequencing'. *Nature Biotechnology* **26**(10), 1135–1145.

[125] Sheth, A. and J. Larson: 1990, 'Federated database systems for managing distributed, heterogeneous, and autonomous databases'. *ACM Computing Surveys (CSUR)* **22**(3), 183–236.

[126] Siegel, E. and D. Channin: 2001, 'Integrating the healthcare enterprise: a primer'. *Radiographics* **21**(5), 1339.

[127] Silva, E., L. Pires, and M. van Sinderen: 2007, 'An algorithm for automatic service composition'. In: *1st International Workshop on Architectures, Concepts and Technologies for Service Oriented Computing, ICSOFT*. pp. 65–74.

[128] Silver, M., H. Su, and S. Dolins: 2001, 'Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse'. *Journal of Healthcare Information Management* **15**(2), 155–164.

[129] Sirin, E., J. Hendler, and B. Parsia: 2003, 'Semi-automatic composition of web services using semantic descriptions'. In: *Web Services: Modeling, Architecture and Infrastructure workshop in conjunction with ICEIS2003*.

[130] Siva, N.: 2008, '1000 Genomes project'. *Nature biotechnology* **26**(3), 256.

[131] Stevens, R., A. Robinson, and C. Goble: 2003, 'myGrid: personalised bioinformatics on the information grid'. *Bioinformatics-Oxford* **19**(1), 302–304.

[132] Stolba, N. and A. M. Tjoa: 2006, 'The Relevance of Data Warehousing and Data Mining in the Field of Evidence-Based Medicine to Support Healthcare Decision Making'. In: *Proceedings of International Conference on Computer Science (ICCS 2006)*. pp. 12–17.

[133] Sujansky, W.: 2001, 'Heterogeneous database integration in biomedicine'. *Journal of biomedical informatics* **34**(4), 285–298.

[134] Szirbik, N., C. Pelletier, and T. Chaussalet: 2006, 'Six methodological steps to build medical data warehouses for research'. *International Journal of Medical Informatics* **75**(9), 683–691.

[135] The BioMoby Consortium (including J. Karlsson): 2008, 'Interoperability with Moby 1.0–It's better than sharing your toothbrush!'. *Brief Bioinform* **9**(3), 220–231.

[136] Thiru, K., A. Hassey, and F. Sullivan: 2003, 'Systematic review of scope and quality of electronic patient record data in primary care'. *British Medical Journal* **326**(7398), 1070.

[137] Trueman, P., S. Hood, U. Nayak, and M. Mrazek: 1999, 'Prevalence of lower urinary tract symptoms and self-reported diagnosed benign prostatic hyperplasia, and their eVect on quality of life in a community-based survey of men in the UK'. *BJU international* **83**, 410–415.

[138] Tsiknakis, M., D. Kafetzopoulos, G. Potamias, A. Analyti, K. Marias, and A. Manganas: 2006, 'Building a European biomedical grid on cancer: the ACGT Integrated Project'. *Studies in health technology and informatics* **120**, 247.

[139] Tu, S. W. and M. A. Musen: 2001, 'Modeling Data and Knowledge in the EON Guideline Architecture'. In: *Stud Health Technol Inform.*, Vol. 84. pp. 280–284.

[140] van Bemmel, J. and M. Musen: 1997, *Handbook of Medical Informatics*. Springer.

[141] Västerbottens läns landsting: 2004, 'Akutsjukvården i Västerbotten, Behandlingsriktlinjer'.

[142] von Bubnoff, A.: 2008, 'Next-Generation Sequencing: The Race Is On'. *Cell* **132**(5), 721 – 723.

[143] von Laszewski, G. and D. Kodeboyina: 2005, 'A repository service for grid workflow components'. In: *International Conference on Autonomic and Autonomous Systems International Conference on Networking and Services. IEEE*. pp. 23–28.

[144] W3C: 2007, 'Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language'. http://www.w3.org/TR/2007/REC-wsdl20-20070626/.

[145] Wilkinson, M. D. and M. Links: 2002, 'BioMOBY: An open source biological web services proposal'. *Brief Bioinform* **3**(4), 331–341.

[146] Wolinsky, H.: 2007, 'The thousand-dollar genome. Genetic brinkmanship or personalized medicine?'. *EMBO reports* **8**(10), 900.

[147] Wolstencroft, K., P. Alper, D. Hull, C. Wroe, P. Lord, R. Stevens, and C. Goble: 2007, 'The my Grid ontology: bioinformatics service discovery'. *International journal of bioinformatics research and applications* **3**(3), 303–325.

[148] Wong, S., V. Tan, W. Fang, S. Miles, and L. Moreau: 2005, 'Grimoires: Grid registry with metadata oriented interface: Robustness, efficiency, security'. *IEEE Distributed Systems Online* **6**(10).

[149] World Health Organization: 1992, 'The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines'. *Geneva, Switzerland: World Health Organisation*.

[150] World Wide Web Consortium: 2006, 'Web Services Architecture Working Group'. `http://www.w3.org/2002/ws/arch/`.

[151] Wu, Z., K. Gomadam, A. Ranabahu, A. Sheth, and J. Miller: 2007, 'Automatic composition of semantic web services using process mediation'. In: *Proceedings of the 9th Intl. Conf. on Enterprise Information Systems ICES*.

[152] Yesavage, J. et al.: 1982, 'Development and validation of a geriatric depression screening scale: a preliminary report'. *J Psychiatr Res* **17**(1), 37–49.

[153] Zahreddine, W. and Q. H. Mahmoud: 2005, 'A Framework for Automatic and Dynamic Composition of Personalized Web Services'. In: *AINA '05: Proceedings of the 19th International Conference on Advanced Information Networking and Applications*. Washington, DC, USA, pp. 513–518.

[154] Zhou, X., S. Chen, B. Liu, R. Zhang, Y. Wang, P. Li, Y. Guo, H. Zhang, Z. Gao, and X. Yan: 2010, 'Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support'. *Artificial Intelligence In Medicine* **48**(2-3), 139–152.

[155] Zimmerman, J., J. van Bemmel, and O. Rienhoff: 1999, 'Medical informatics education'. *Journal of the American Society for Information Science* **39**(2), 138–141.

Part III

In this last part of the thesis, the following papers[1] can be found:

Paper I P. Eklund, J. Karlsson, J. Rauch, and M. Simunek. *On the Logic of Medical Decision Support.* In Harrie C. M. de Swart, Ewa Orlowska, Gunther Schmidt, and Marc Roubens, editors, Theory and Applications of Relational Structures as Knowledge Instruments, volume 4342 of Lecture Notes in Computer Science, pages 50-59. Springer, 2006.

Paper II P. Eklund, J. Karlsson, and A. Näslund. *Mobile Pharmacology.* In Marcin S. Szczuka, Daniel Howard, Dominik Slezak, Haeng-Kon Kim, Tai-Hoon Kim, Il Seok Ko, Geuk Lee, and Peter M. A. Sloot, editors, ICHIT, volume 4413 of Lecture Notes in Computer Science, pages 522-533. Springer, 2006.

Paper III P. Eklund, S. Eriksson, J. Karlsson, H. Lindgren, and A. Näslund. *Software development and maintenance strategies for guideline implementation.* In Proc. EUNITE-Workshop "Intelligent Systems in Patient Care", pages 26-45. Austrian Computer Society, 2000.

Paper IV J. Ríos, J. Karlsson, and O. Trelles. *Magallanes: a web services discovery and automatic workflow composition tool.* BMC Bioinformatics, 10(1):334, 2009.

Paper V J. Karlsson, S. Ramirez, J.F. Aldana-Montes, and O. Trelles. *Workflow*

[1]Papers I, II, VI, VII are reprinted with permission from Springer

repositories for bioinformatics. Technical report, Department of Computer Architecture, Malaga University, 2008.

Paper VI P. Eklund, M.A. Galán, and J. Karlsson. *Rough Monadic Interpretations of Pharmacologic Information*. In ICCS 2007: Proceedings of the 15th International Workshops on Conceptual Structures, pages 108-113. Springer, 2007.

Paper VII P. Eklund, M.A. Galán, and J. Karlsson. Rough Set Theory: A True Landmark in Data Analysis, chapter *Categorical Innovations for Rough Sets*, pages 45-69. Springer, 2009.