Linköping Studies in Science and Technology

Dissertation No. 1035

Integration of Biological Data

by

Vaida Jakonienė

Department of Computer and Information Science
Linköpings universitet
SE-581 83 Linköping, Sweden

Linköping 2006

Abstract

Data integration is an important procedure underlying many research tasks in the life sciences, as often multiple data sources have to be accessed to collect the relevant data. The data sources vary in content, data format, and access methods, which often vastly complicates the data retrieval process. As a result, the task of retrieving data requires a great deal of effort and expertise on the part of the user. To alleviate these difficulties, various information integration systems have been proposed in the area. However, a number of issues remain unsolved and new integration solutions are needed.

The work presented in this thesis considers data integration at three different levels. 1) Integration of biological data sources deals with integrating multiple data sources from an information integration system point of view. We study properties of biological data sources and existing integration systems. Based on the study, we formulate requirements for systems integrating biological data sources. Then, we define a query language that supports queries commonly used by biologists. Also, we propose a high-level architecture for an information integration system that meets a selected set of requirements and that supports the specified query language. 2) Integration of ontologies deals with finding overlapping information between ontologies. We develop and evaluate algorithms that use life science literature and take the structure of the ontologies into account. 3) Grouping of biological data entries deals with organizing data entries into groups based on the computation of similarity values between the data entries. We propose a method that covers the main steps and components involved in similarity-based grouping procedures. The applicability of the method is illustrated by a number of test cases. Further, we develop an environment that supports comparison and evaluation of different grouping strategies.

The work is supported by the implementation of: 1) a prototype for a system integrating biological data sources, called BioTRIFU, 2) algorithms for ontology alignment, and 3) an environment for evaluating strategies for similarity-based grouping of biological data, called KitEGA.

Acknowledgements

Many people have supported my graduate work and made this PhD thesis possible.

I am grateful to my supervisor, Associate Professor Patrick Lambrix, for his support and guidance during this work. His constructive comments and our many conversations brought insight and helped to shape this thesis. His encouragement, patience, and devotion as a teacher helped me to grow as a researcher. I am glad that I had the opportunity to work with him.

I would like to express my appreciation to Professor Nahid Shahmehri for providing valuable comments, pointing out important aspects of the research world and giving support during this work.

The members of IISLAB (Laboratory for Intelligent Information Systems) created a stimulating and supportive working environment. I am thankful for their friendship over all these years. Specially, I want to mention my student colleagues: Shanai Ardi, Ioan Chisalita, Claudiu Duma, Almut Herzog, Dennis Maciuszek, He Tan, Eduard Turcan and Cécile Åberg.

This work would have been much harder without the support of my family, relatives and friends. I would especially like to express my gratitude to my Mum and Dad for caring so much about me, and for welcoming me with such warmth when I returned home to Lithuania. I would also like to thank my sister for being such a generous and joyful person. I am very lucky to have her. The friends I made in Linköping made my stay in Sweden much more enjoyable. In particular, I want to thank my dearest friends Akvilė, Aleksandra and Joe. I greatly value their company and our conversations.

This research work was funded by CUGS (the national graduate school in computer science). I also acknowledge the financial support of the EU Network of Excellence REWERSE (Sixth Framework Programme project 506779).

Vaida Jakonienė

Linköping, September 2006

Enclosed Papers

This thesis contains revised versions of the following papers.

1. Lambrix P, Jakonienė V. Towards transparent access to multiple biological databanks. *Proceedings of the First Asia-Pacific Bioinformatics Conference*, pp 53-60, Adelaide, Australia, 2003.

2. Jakonienė V, Lambrix P. Information integration systems for biological data sources: requirements and opportunities. Submitted.

3. Jakonienė V, Lambrix P. Ontology-based integration for bioinformatics. *Proceedings of the VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems - ODBIS 2005*, pp 55-58, Trondheim, Norway, 2005.

4. Tan H, Jakonienė V, Lambrix P, Aberg J, Shahmehri N. Alignment of Biomedical Ontologies using Life Science Literature. *Proceedings of the International Workshop on Knowledge Discovery in Life Science Literature*, pp 1-17, Singapore, 2006. LNBI 3886.

5. Jakonienė V, Rundqvist D, Lambrix P. A method for similarity-based grouping of biological data. *Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences - DILS06*, pp 136-151, Hinxton, UK, 2006. LNBI 4047.

6. Jakonienė V, Lambrix P. A Tool for Evaluating Strategies for Grouping of Biological Data. Submitted.

List of Additional Papers

Related Papers

The following are related research articles not included in the thesis.

1. Doms A, Jakonienė V, Lambrix P, Schroeder M, Wächter T. Ontologies and Text Mining as a Basis for a Semantic Web for the Life Sciences. *Reasoning Web, Second International Summer School*, Springer-Verlag, pp 164-183, 2006. LNCS 4126.

2. Jakonienė V. *A Study in Integrating Multiple Biological Data Sources*. Licentiate thesis No 1149, Linköpings universitet, Sweden, 2005.

3. Lambrix P, Tan H, Jakonienė V, Strömbäck L. Biological Ontologies. Chapter in Baker CJO, Cheung KH (eds) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, Springer, 2006. To appear.

4. Strömbäck L, Jakonienė V, Tan H, Lambrix P. Representing, storing and accessing molecular interaction data: a review of models and tools. *Briefings in Bioinformatics*, 2006. Invited contribution. To appear.

Other

1. Backofen R, Badea M, Barahona P, Berndtsson M, Burger A, Dawelbait G, Doms A, Fages F, Hotaran A, Jakonienė V, Krippahl L, Lambrix P, McLeod K, Nutt W, Olsson B, Schroeder M, Schroiff A, Soliman S, Tan H, Tilivea D, Will S. Requirements and specification of use cases. REWERSE Deliverable A2-D3, 2005.

2. Backofen R, Badea M, Barahona P, Burger A, Dawelbait G, Doms A, Fages F, Hotaran A, Jakonienė V, Krippahl L, Lambrix P, McLeod K, Möller S, Nutt W, Olsson B, Schroeder M, Soliman S, Tan H, Tilivea D, Will S. Usage of bioinformatics tools and identification of information sources. REWERSE Deliverable A2-D2, 2005.

3. Jakonienė V, Nilsson R. *Abstract Book of the Fourth Swedish Bioinformatics Workshop for PhD students and PostDocs*, Linköping, Sweden, 2003.

Table of Contents

Introduction **1**

Motivation 3

Problem Statement 5

Contributions 7

Paper Summaries 9

Related Work 11

Future Work 14

References 16

Introduction

1 Motivation

Researchers in areas, such as, medicine, agriculture and environmental sciences, intensively use the available biological data to answer different research questions or to solve various tasks [CGG03]. One of the main goals is to understand how various organisms function as biological systems. To achieve this goal, it is important to explore functions and interactions of genome-encoded components. This type of knowledge may be used for different purposes. For instance, it is used to identify genes responsible for a disease, to develop drugs enabling treatment of diseases and to predict organisms' responses to a drug.

The significance of these areas, the worldwide interest and the available tools and techniques caused the generation of an enormous amount of biological data, such as DNA and protein sequences, gene regulatory and protein interaction networks, and secondary and tertiary structures of molecules. This data is spread over a large number of autonomous data sources that are often publicly available on the Web. For instance, 858 data sources are listed in the 2006 Database Issue of the Nucleic Acids Research [NAR] journal. As the data sources are developed and supported independently by different groups and organizations, they are highly heterogeneous in various aspects. For example, the data sources vary in the type of the stored data, the data format, and access methods. Further, there is a terminology discrepancy at the schema and data levels. In addition to data sources, a large number of bio-ontologies describing domain knowledge are publicly available in the area [LTJ06]. For instance, OBO [OBO], an umbrella web address for ontologies covering the genomics and proteomics domains, lists 29 orthogonal ontologies. Some of the ontologies have reached the status of de facto standard and are used extensively to annotate the data sources.

Data integration is an important procedure underlying many research tasks in the life sciences, as often multiple data sources have to be accessed to collect the relevant data. For instance, to find publications describing a given disease that relates to a certain type of sequences may require analysis of data sources for publications, diseases and sequences together with some other data sources combining these types of information [LMN04]. To support health care applications by using results in functional genomics, the integration of clinical data and genomic data is important [MIN04].

To successfully accomplish tasks that require data retrieved from multiple data sources, a lot of effort and knowledge is required from the user. Several

steps are performed to acquire the data: data sources that contain relevant data are selected, queries over each data source are formulated and decisions are made on how to combine the results. To find relevant data sources, the user has to be acquainted with the content of different data sources. To formulate a query and decide on how to execute the query, the user has to be familiar with the ways the data sources support data retrieval and how data at different sources relate to each other. To execute the query, the user has to know the location of the data sources that are spread over the Web, the different query languages and data formats. During query execution the user may need to translate the data between different formats and combine the results. A mistake in any of these steps may either result in inefficient query execution or not finding results. The process is also time consuming since a large amount of data is usually processed. Data retrieval may take a long time, e.g. when tools are used to acquire the results. As biological data sources change often and data sources appear and disappear, the user has to be aware of these changes.

To alleviate these difficulties various information integration solutions have been proposed. Specialized integration solutions focus on solving a single task based on a set of relevant data sources. In contrast, general purpose information integration systems aim to support a broad range of tasks and integration of various data sources. Such systems may provide a common interface through which a user accesses multiple data sources. In this case the location and different query languages of the data sources are hidden from the user. Other types of information integration systems even hide the integrated data sources from the user. During query processing, these systems handle also the selection of data sources that are relevant to the query. However, new integration solutions are needed to better support life science researchers in their tasks. A number of open issues remain in the available integration solutions. For instance, it may be difficult to integrate new data sources into the existing systems or to reuse the systems for new tasks. Furthermore, solutions are lacking for managing incomplete and incorrect data, and for handling semantic heterogeneity. For solving some of the problems specialized solutions have to be developed while in other cases developments in other areas could be adapted.

This thesis focuses on data integration at three different levels. This includes integration of biological data sources, integration of ontologies and integration or grouping of biological data entries. **Integration of biological data sources** deals with the fact that biologists face a number of problems

when they want to use biological data sources to find relevant information for their research and analyzes ways of dealing with these problems in combination from an information integration system point of view. Further, two specific tasks in integrating biological data are dealt with. **Integration of ontologies** deals with finding overlapping information between ontologies. This includes finding relationships, called alignments, between the related terms in the ontologies. **Grouping of biological data entries** deals with organizing data entries into groups based on the computation of similarity values between the data entries. Grouping of data entries is an abstraction of the problem of finding entries that represent the same entity in different data sources that is a basic operation for integrating the data entries.

2 Problem Statement

The work presented in this thesis aims to develop approaches and techniques that alleviate the challenges met when using and integrating biological data, and in particular, the heterogeneity present at different levels in the data and data sources. The thesis focuses on the identification and analysis of the available knowledge about data and data sources, and the development of mechanisms that use the available knowledge for integration of biological data. To achieve these goals, we focus on the following tasks in the thesis.

2.1 Integration of biological data sources

In this thesis we deal with a few aspects in the context of integrating biological data sources: requirements and query languages for information integration systems, and the use of ontologies for integrating the data sources. Despite the fact that a number of information integration solutions are proposed in the life sciences, not so much research has been performed on the requirements for such systems. Such a study of requirements is needed as the complexity of the life sciences, the tasks to be solved, the style of the scientific research and the properties of the available data sources pose special requirements for information integration systems in the area. Further, the difference in focus of the existing information integration systems together with different design and development choices led to the fact that often systems support a unique query language. The variety of the available query languages makes it difficult to select between the query languages and to judge their suitability for expressing the tasks in the life sciences. It is im-

portant to know a subset of query language operators that should be present in any query language for integrating biological data sources, for instance, to support the development of new integration solutions. In addition, during the recent years some solutions were proposed for using ontologies in information integration systems. However, this is still done in a limited way and only a small part of the possible ontology-based knowledge is currently used.

In this thesis we focus on:

- Study of requirements for systems providing integrated access to biological data sources with focus on systems providing virtual integration of data sources, i.e. preserving autonomy of data sources.
- Specification of a query language that allows formulation of different types of queries commonly used by biologists.
- Specification of a high-level architecture for an information integration system that meets a selected set of requirements and that supports the specified query language.
- Design and development of a prototype for the information integration system. The system should conform to the high-level architecture and enable deeper exploration of issues related to query processing over multiple biological data sources.
- Identify types of ontological knowledge publicly available in the area of life sciences and study how this knowledge could be used to enhance current integration approaches.

2.2 Integration of ontologies

The task of aligning ontologies is not well explored and is considered to be one of the major issues in the life sciences [CGG03]. A number of alignment strategies are proposed, but further research and development of new strategies are needed [LT06a, LT06b]. For instance, not much work has been done on ontology alignment using life science literature as a resource for finding alignments. Also not many strategies use information about the structure of the ontologies.

In this thesis we focus on:

- Study how ontology alignment could be performed based on life science literature.

- Study how the structure of ontologies could be used in ontology alignment.

2.3 Grouping of biological data entries

Many tools for analyzing biological data use some form of grouping and are used in, for instance, data integration, data cleaning, prediction of protein functionality, and correlation of genes based on microarray data. A number of aspects influence the quality of the grouping results: the data sources, the grouping attributes and the algorithms implementing the grouping procedure. Many methods exist, but it is often not clear which methods perform best for which grouping tasks. The study of the properties, and the evaluation and the comparison of the different aspects that influence the quality of the grouping results, would give us valuable insight in how the grouping procedures could be used in the best way. It would also lead to recommendations on how to improve the current procedures and develop new procedures. To be able to perform such studies and evaluations we need environments that allow us to compare and evaluate different grouping strategies.

In this thesis we focus on:

- Specification of a method that covers the main steps and components that should be included in environments.

- Design and development of a prototype for an environment supporting the evaluation of similarity-based grouping procedures. The environment should be based on the defined method.

3 Contributions

The main contributions of the thesis are the following:

Integration of biological data sources

- Study of biological data sources. The results are presented in paper 1 and 2. Paper 2 extends the work done in paper 1.

- Identification of requirements for information integration systems for biological data sources. Paper 2 presents and discusses the requirements.

- Study of current information integration systems for biological data sources with respect to the identified requirements. The work is included in paper 2.

- Proposal for a query language and architecture for the BioTRIFU[1] system. The contributions appear in paper 1.
  As a feasibility study and to get an overview of issues related to query processing over multiple biological data sources, a subset of the defined query language and the ideas included in the architecture definition were implemented in a prototype. The prototype supports the main steps and components needed to integrate two data sources that can be accessed at different locations. For details we refer to [Jak05].

- Identification of ontological knowledge and its use in information integration systems for biological data sources. Paper 3 discusses the results.

- Proposal of an ontology-based approach for information integration systems for biological data sources. The approach is presented in paper 3.

Integration of ontologies

- Development and evaluations of algorithms for ontology alignment. The algorithms use life science literature and take the structure of the ontologies into account. The contributions are described in paper 4.
  The ontology alignment algorithms were implemented and incorporated into the SAMBO system [LT06a].

Grouping of biological data entries

- Proposal of a method for similarity-based grouping of biological data. The method is introduced in paper 5.
  As a feasibility study, two grouping tasks were implemented and analyzed through a number of test cases.

- Development and implementation of KitEGA[2], an environment for evaluating strategies for similarity-based grouping of biological data. The environment is based on the proposed method. The tool and its

[1] *T*he *R*ight *I*nformation *F*or yo*U* in *Bio*informatics
[2] Tool*Kit* for *E*valuation of *G*rouping *A*lgorithms

use are presented in paper 6.
The current implementation of KitEGA supports the specification of test cases through the use of plug-ins and user interfaces, and provides a number of user interfaces supporting analysis of the grouping results.

4 Paper Summaries

In this section we give short summaries of the six papers included in this thesis. Papers 1, 2 and 3 deal with integration of biological data sources, with paper 3 focusing on ontology-based integration. Paper 4 deals with integration of ontologies. Papers 5 and 6 deal with grouping of biological data entries.

Paper 1: Towards transparent access to multiple biological databanks

In paper 1 we discuss common problems met by the users of biological data sources. The discussion includes a study of current biological data sources. Based on the observations, the paper proposes a base query language that contains operators that should be present in any query language for biological data sources. Further, the paper presents an architecture for a system supporting such a language and enabling transparent and integrated access to biological data sources.

Paper 2: Information integration systems for biological data sources: requirements and opportunities

In paper 2 requirements for information integration systems in the area of bioinformatics are identified. This paper extends the study of problems and requirements identified in paper 1. First, we study biological data sources and identify their properties that make querying multiple biological data sources a difficult task. Then, we formulate requirements for information integration systems for biological data sources. We also discuss how well current information integration systems satisfy these requirements and identify opportunities for future research.

Paper 3: Ontology-based integration for bioinformatics

In paper 3 we argue that the current approaches for integrating biological data sources should be enhanced by ontological knowledge. We identify the different types of ontological knowledge that are available on the Web

(ontologies, ontology alignments, annotations, mappings between data values and ontological terms) and propose an approach to use this knowledge to support integrated access to multiple biological data sources. We also show that current ontology-based integration approaches only cover parts of our approach.

Paper 4: Alignment of biomedical ontologies using life science literature

In paper 4 we propose strategies for aligning ontologies based on life science literature. We propose a basic algorithm as well as extensions that take the structure of the ontologies into account. We evaluate the strategies and compare them with strategies implemented in the alignment system SAMBO. We also evaluate the combination of the proposed strategies and the SAMBO strategies.

Paper 5: A method for similarity-based grouping of biological data

In paper 5 a method for similarity-based grouping is proposed. As the main steps the method contains specification of grouping rules, pairwise grouping between entries, actual grouping of similar entries, and evaluation and analysis of the results. Often, different strategies can be used in the different steps. The method enables exploration of the influence of the choices and supports evaluation of the results with respect to given classifications. The grouping method is illustrated by test cases based on different strategies and classifications. The results show the complexity of the similarity-based grouping tasks and give deeper insights in the selected grouping tasks, the analyzed data source, and the influence of different strategies on the results.

Paper 6: A Tool for evaluating strategies for grouping of biological data

In paper 6 we present KitEGA, an environment supporting the evaluation of grouping strategies. Based on the method presented in paper 5, we propose a framework for comparative evaluation of strategies for grouping data based on the method, and present its current implementation. Further, we illustrate the use of KitEGA by comparing grouping strategies for classifying proteins regarding biological function and isozymes.

5 Related Work

5.1 Integration of biological data sources

Requirements for general purpose information integration systems for biological data sources on the Web were discussed in [DOB95], [Kar96], [Won02] and [HK04]. The first two papers were written a decade ago. Since then, the area of life sciences has evolved fast: many more data sources and tools are publicly available and new tasks have to be solved. While some of the earlier defined requirements for information integration systems are still valid in the changed environment, other requirements need to be reconsidered and new requirements need to be specified. The more recent paper [Won02] argues for a general purpose information integration system that supports core functionality needed for information integration in life sciences. Therefore, the defined requirements do not cover some of the issues specific to the area. The authors of [HK04] point out a few high level requirements for information systems emphasizing the need to automate a maximum number of tasks while minimizing the amount of time and interactions for the user. The requirements provided in [HK04] are in line with the requirements specified in paper 2. In paper 2 the requirements are specified at a more detailed level by looking at different information integration aspects and focusing on systems providing virtual integration of data sources.

Within the area of life sciences several integration approaches have been proposed and systems have been implemented. This includes systems based on database technology, i.e. virtual and materialized (data warehouses) integration approaches. Also, systems based on the Semantic Web, web services, grid and agents technologies are developed. In this thesis we focused on issues related to virtual integration. For an overview of such systems see paper 2. For solving specialized tasks, the use of warehouses is a widely adopted integration solution (e.g. [TRM05]). During the recent years Semantic Web technologies are being used for resolving scalability, heterogeneity and reusability problems in the life sciences. In these approaches biological data and knowledge is represented using Semantic Web languages, e.g. XML, RDF and OWL [Muk05]. A number of studies are conducted to explore integrated use of data represented in these formats, e.g. [CYS05] and [SLD06]. Also, the use of ontologies is proposed to resolve semantic heterogeneity problems and to support knowledge discovery based on biological data [Gar05]. Further, work is ongoing in applying web services and grid technologies in the area [VS05]. BioMOBY [WL02] and myGrid [SRG03]

are example projects based on these technologies. Also, agent technology is shown to be useful for meeting integration challenges in the life sciences. The authors in [KBB04] argue that advanced communication supported by agent technology can complement the Semantic Web and grid technologies.

Some of the available information integration systems use ontology-based technologies to support querying (e.g. BACIIS [MWL03], KIND [LGM03], SEMEDA [KPL03] and TAMBIS [GSN01]). A common feature is that the integrated schemas used in these systems are seen as ontologies. In contrast, in the approach described in paper 3, we expect ontologies to be agreed upon and shared by many users [Lam04]. As in our approach, the integrated schemas include domain knowledge and information on data structures at the data sources. All the systems use the maintained ontology to describe the content of data sources. Though it is not explicitly stated, cross-references between data sources are probably used to join the retrieved data items. KIND uses two ontologies describing static and process knowledge, respectively. The ontologies combine domain knowledge from neuroanatomy and neurophysiology. In SEMEDA controlled vocabularies can be used to specify semantics of data type values. Also, data source content descriptions can be refined with integrated schema terms. Ontological annotations and mappings between ontology terms are not taken into account in any of the systems.

5.2 Integration of ontologies

Different strategies can be used to perform alignment of ontologies. [LT06b] describes a general strategy for aligning two ontologies. One of the main component types is a matcher responsible for computing similarities between the terms from the different source ontologies. The matchers can implement strategies based on linguistic matching, structure-based strategies, constraint-based approaches, instance-based strategies, strategies that use auxiliary information or a combination of these. By using different matchers and combining and filtering the results in different ways we obtain different alignment strategies. Tools for ontology alignment are discussed in [LT06a].

Some ontology alignment and merging systems provide alignment strategies using literature, such as ArtGen [MW02], FCA-Merge [SM01] and OntoMapper [PPF02]. Also, there are systems that implement alignment algorithms based on the structure of the ontologies. Most systems rely on the

existence of previously aligned concepts. For instance, Anchor-PROMPT [NM01] determines the similarity of concepts by the frequency of their appearance along the paths between previously aligned concepts. The paths may be composed of any kind of relations. Also SAMBO as described in [LT05] provides such a component where the similarity between concepts is augmented based on their location in the is-a hierarchy relative to already aligned concepts. In contrast, the methods proposed in this thesis do not require previously aligned concepts.

OntoMapper implements the most similar approach to the strategies described in paper 4. OntoMapper provides an ontology alignment algorithm using Bayesian learning. A set of documents (abstracts of technical papers taken from ACM's digital library and Citeseer) is assigned to each concept in the ontologies. Two raw similarity scores matrices for the ontologies are computed by the Rainbow text classifier. The similarity between the concepts is calculated based on these two matrices using the Bayesian method. When analyzing structure of the ontologies, OntoMapper does not require previously aligned concepts and takes the documents from the sub-concepts into account when computing the similarity between two concepts. However, as this is hard-coded in the method, it is not clear how the structure of the ontologies influences the result of the computation.

In contrast to most other approaches, [CTL06] uses the structural information not to compute similarity between ontological terms, but as a method for filtering wrong results generated by matchers. The approach gives good results when many initial suggestions are available and the time for filtering is often only a small fraction of the time for the similarity computation.

5.3 Grouping of biological data entries

There are two kinds of related work: evaluations of grouping algorithms and tools for supporting evaluation of grouping algorithms.

A number of evaluations of different kinds of grouping algorithms have been performed. For instance, regarding clustering of gene expression data [YHR01] proposes a measure to estimate the predictive power of a clustering algorithm and compares two partitional and three hierarchical clustering algorithms based on this measure. [DD03] proposes three validation strategies and compares six algorithms. Also [GSS03] proposes a new validation measure and compares four clustering methods. Five biclustering methods for gene expression data are evaluated in [PBZ06]. Common to all these evalu-

ations is the fact that they focus on cluster validation for the evaluation and comparison of algorithms. They use synthetic and real data sources. Some of the papers also aim to propose new validation measures. Further, in all these evaluations, most of the evaluated algorithms needed to be re-implemented for the purpose of the evaluations.

[CRF03] presents the SecondString Toolkit for name-matching methods which could be used, for instance, in duplicate detection. Several distance functions for strings are implemented. The algorithms are compared on a data set regarding non-interpolated average precision.

A system that goes some way into providing an environment for clustering and validation is the Machaon Cluster Validation Environment [BAC05]. This system is intended for clustering of microarray data and evaluating the quality of the obtained clusters. The system focuses on cluster validation for new data sets and therefore uses internal measures based on compactness and isolation. The system implements several clustering algorithms, metrics (distance), and internal measures [BA03]. The user can choose among these to run a cluster task on a data set. The results are shown as a tree. The highest level nodes represent the chosen cluster algorithms with particular parameter selection. The next level represents the results of applying different validity measures to the clusters generated by the algorithm.

The framework and system (KitEGA) that we propose in papers 5 and 6 aims to go one step further. KitEGA is a platform for evaluating and comparing similarity-based grouping strategies. Evaluators can plug in their own algorithms related to the grouping strategies and the evaluation measures, as well as their own data sets. KitEGA provides then the support for running the algorithms, and summarizing and analyzing the results.

6 Future Work

6.1 Integration of biological data sources

As we observed in section 5.1 the focus of the research on integrating data in the life sciences is reorienting from the use of classical database approaches to the use of web and Semantic Web technologies. [Muk05] mentions challenges to make the best use of the new technologies. First, most of the biological data and knowledge should be available in the Semantic Web. To achieve this, tools supporting automatic extraction of biological data from the literature is needed. Also, most of the currently available tools for the

Semantic Web are research prototypes. Further studies are needed on how to extend these prototypes into systems supporting real-world applications for effective retrieval of information and discovery of hidden knowledge on the Semantic Web. For instance, to guarantee scalability, inference engines available for querying the Semantic Web and graph theory based algorithms used to explore associations between objects on the Semantic Web may have to be reconsidered.

Paper 2 enumerates other challenges for information integration systems for the life sciences. To allow users to view and specify different types of information, more powerful modules for supporting interaction between the users and information integration systems are needed. Also, the need for further research on how to resolve semantic heterogeneity is emphasized. For instance, the available approaches, like the ontology-based data integration approach proposed in paper 4, could be tested in the context of the real Semantic Web. Also, paper 2 states the need for tools supporting the development and maintenance of information integration systems. Such tools are essential to cope with the scale and dynamics of the life sciences.

6.2 Integration of ontologies

Alignment and merging of ontologies is an important research topic and new systems and strategies for ontology alignment should be developed. More studies are needed that explore which strategies work well for which types of ontologies and a system as KitAMO [LT06c] can provide a good environment to perform these studies. In the future we will see an increase of available alignments between ontologies. This will provide a type of ontological information that can be used in, for instance, data integration as discussed in paper 3. Further, there are efforts to promote interoperability of ontologies, such as the OBO Foundry where it is required that the ontologies use relations which are unambiguously defined following the pattern of definitions defined in the OBO Relation Ontology [SCK05]. The results of such efforts will provide information that should be taken into account during the alignment process.

There are a number of issues related to the algorithms in paper 4 that would be interesting to further investigate. A limitation of our algorithms is that abstracts of research articles are only classified to one concept. We want to extend our strategies by allowing abstracts to be classified to 0, 1 or more concepts. We are also interested in looking at other classification algorithms.

Regarding the structure the ontologies in the current experiments are reasonably simple taxonomies. We want to investigate whether the structure-based strategies lead to similar results for other types of ontologies. Further, our matchers could be enhanced to use synonyms and domain knowledge.

6.3 Grouping of biological data entries

Similarity-based grouping of data entries is not a trivial task. In order to find the most suitable grouping strategies for given tasks, tools are needed to support the evaluation and comparison of different grouping procedures. An example of such system is KitEGA (paper 6). We intend to extend the current KitEGA implementation in several ways. We will extend the system to fully comply with our framework. Further, we will provide a number of libraries for components that are common. This could include, for instance, different evaluation measures or grouping methods. We will also use KitEGA for studies in data integration.

References

[BA03] Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. *Signal Processing*, 83:825-833, 2003.

[BAC05] Bolshakova N, Azuaje F, Cunningham P. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21(4):451-455, 2005.

[CGG03] Collins F, Green E, Guttmacher A, Guyer M. A Vision for the Future of Genomics Research. *Nature*, 422:835-847, 2003.

[CRF03] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. *Proceedings of the KDD Workshop on Data Cleaning and Object Consolidation*, 2003.

[CTL06] Chen B, Tan H, Lambrix P. Structure-based filtering for ontology alignment. *Proceedings of the IEEE WETICE Workshop on Semantic Technologies in Collaborative Applications*, 2006.

[CYS05] Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* 21(1):i85-96, 2005.

[DD03] Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459-466, 2003.

[DOB95] Davidson S, Overton C, Buneman P. Challenges in Integrating Biological Data Sources. *Journal of Computational Biology*, 2(4):557-572, 1995.

[Gar05] Gardner SP. Ontologies and semantic data integration. *Drug Discovery Today*, 10(14):1001-1007, 2005.

[GSN01] Goble CA, Stevens R, Ng G, Bechhofer S, Paton N, Baker P, Peim M, Brass A. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40(2), 2001.

[GSS03] Gat-Viks I, Sharan R, Shamir R. Scoring clustering solutions by their biological relevance. *Bioinformatics*, 19(18):2381-2389, 2003.

[HK04] Hernandez T, Kambhampati S. Integration of biological sources: Current systems and challenges. *ACM SIGMOD Record*, 33(3):51-60, 2004.

[Jak05] Jakonienė V. *A Study in Integrating Multiple Biological Data Sources*. Licentiate thesis No 1149, Linköpings universitet, Sweden, 2005.

[Kar96] Karp P. A strategy for database interoperation. *Journal of Computational Biology*, 2(4):573-586, 1996.

[KBB04] Karasavvas KA, Baldock R, Burger A. Bioinformatics integration and agent technology. *Journal of Biomedical Informatics*, 37(3):205-219, 2004.

[KPL03] Köhler J, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics*, 19(18):2420-2427, 2003.

[Lam04] Lambrix P. Ontologies in Bioinformatics and Systems Biology. Chapter 8 in Dubitzky W, Azuaje F (eds) *Artificial Intelligence Methods and Tools for Systems Biology*, Springer, pp 129-146, 2004.

[LGM03] Ludäscher B, Gupta A, Martone ME. A Model-Based Mediator System for Scientific Data Management. Chapter 12 in Lacroix Z,

Critchlow T (eds) *Bioinformatics: Managing Scientific Data*, Morgan Kaufmann Publishers, pp 335-370, 2003.

[LMN04] Lacroix Z, Murthy H, Naumann F, Raschid L. Links and Paths through Life Science Data Sources. *Proceedings of the International Workshop on Data Integration in the Life Sciences*, pp 203-211, 2004. LNCS 2994.

[LT05] Lambrix P, Tan H. A Framework for Aligning Ontologies. *Proceedings of the Workshop on Principles and Practice of Semantic Web Reasoning*, pp 17-31, 2005. LNCS 3703.

[LT06a] Lambrix P, Tan H. SAMBO - A System for Aligning and Merging Biomedical Ontologies. *Journal of Web Semantics, Special issue on Semantic Web for the Life Sciences*, 2006.

[LT06b] Lambrix P, Tan H. Ontology alignment and merging. Chapter in Burger A, Davidson D, Baldock R (eds) *Anatomy Ontologies for Bioinformatics: Principles and Practice*, Springer, 2006. To appear.

[LT06c] Lambrix P, Tan H. A Tool for Evaluating Ontology Alignment Strategies. *Journal on Data Semantics*, VIII, 2006. To appear.

[LTJ06] Lambrix P, Tan H, Jakonienė V, Strömbäck L. Biological Ontologies. Chapter in Baker CJO, Cheung KH (eds) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, Springer, 2006. To appear.

[MIN04] Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvarova J, Hermosilla I, Vicente F. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *Journal of Biomedical Informatics*, 37:30-42, 2004.

[Muk05] Mukherjea S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Briefings in Bioinformatics*, 6(3):252-62, 2005.

[MW02] Mitra P, Wiederhold G. Resolving terminological heterogeneity in ontologies. *Proceedings of the ECAI Workshop on Ontologies and Semantic Interoperability*, 2002.

[MWL03] Miled ZB, Webster YW, Liu Y, Li N. An Ontology for Semantic Integration of Life Science Web Databases. *International Journal of Cooperative Information Systems*, 12(2):275-294, 2003.

[NAR] NAR. Nucleic Acids Research. http://nar.oupjournals.org

[NM01] Noy N, Musen M. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. *Proceedings of the IJCAI Workshop on Ontologies and Information Sharing*, pp 63-70, 2001.

[OBO] OBO. Open Biomedical Ontologies. http://obo.sourceforge.net/

[PBZ06] Prelić A, Bleuler S, Zimmermann Ph, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression. *Bioinformatics*, 22(9):1122-1129, 2006.

[PPF02] Prasad S, Peng Y, Finin T. Using Explicit Information To Map Between Two Ontologies. *Proceedings of the AAMAS Workshop on Ontologies in Agent Systems*, 2002.

[SCK05] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies. *Genome Biology*, 6(5):R46, 2005.

[SLD06] Stephens S, LaVigna D, DiLascio M, Luciano J. Aggregations of Bioinformatics Data Using Semantic Web Technology. *Journal Web Semantics*, 4(3), 2006.

[SM01] Stumme G, Mädche A. FCA-Merge: Bottom-up merging of ontologies. *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp 225-230, 2001.

[SRG03] Stevens RD, Robinson AJ, Goble CA. MyGrid: personalised bioinformatics on the information. *Bioinformatics*, 19(1):i302-i304, 2003.

[TRM05] Trißl S, Rother K, Müller H, Steinke T, Koch I, Preissner R, Frömmel C, Leser U. Columba: An Integrated Database of Proteins, Structures, and Annotations. *BMC Bioinformatics*, 6:81, 2005.

[VS05] Vyas H, Summers R. Interoperability of bioinformatics resources. *VINE: The journal of information and knowledge management systems*, 35(3):132-139, 2005.

[WL02] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics*, 3(4):331-41, 2002.

[Won02] Wong L. Technologies for integrating Biological Data. *Briefings in Bioinformatics*, 3(4):389-404, 2002.

[YHR01] Yeung K, Haynor D, Ruzzo W. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309-318, 2001.

Department of Computer and Information Science
Linköpings universitet

Dissertations

Linköping Studies in Science and Technology

No 14 **Anders Haraldsson:** A Program Manipulation System Based on Partial Evaluation, 1977, ISBN 91-7372-144-1.

No 17 **Bengt Magnhagen:** Probability Based Verification of Time Margins in Digital Designs, 1977, ISBN 91-7372-157-3.

No 18 **Mats Cedwall:** Semantisk analys av process-beskrivningar i naturligt språk, 1977, ISBN 91-7372-168-9.

No 22 **Jaak Urmi:** A Machine Independent LISP Compiler and its Implications for Ideal Hardware, 1978, ISBN 91-7372-188-3.

No 33 **Tore Risch:** Compilation of Multiple File Queries in a Meta-Database System 1978, ISBN 91-7372-232-4.

No 51 **Erland Jungert:** Synthesizing Database Structures from a User Oriented Data Model, 1980, ISBN 91-7372-387-8.

No 54 **Sture Hägglund:** Contributions to the Development of Methods and Tools for Interactive Design of Applications Software, 1980, ISBN 91-7372-404-1.

No 55 **Pär Emanuelson:** Performance Enhancement in a Well-Structured Pattern Matcher through Partial Evaluation, 1980, ISBN 91-7372-403-3.

No 58 **Bengt Johnsson, Bertil Andersson:** The Human-Computer Interface in Commercial Systems, 1981, ISBN 91-7372-414-9.

No 69 **H. Jan Komorowski:** A Specification of an Abstract Prolog Machine and its Application to Partial Evaluation, 1981, ISBN 91-7372-479-3.

No 71 **René Reboh:** Knowledge Engineering Techniques and Tools for Expert Systems, 1981, ISBN 91-7372-489-0.

No 77 **Östen Oskarsson:** Mechanisms of Modifiability in large Software Systems, 1982, ISBN 91-7372-527-7.

No 94 **Hans Lunell:** Code Generator Writing Systems, 1983, ISBN 91-7372-652-4.

No 97 **Andrzej Lingas:** Advances in Minimum Weight Triangulation, 1983, ISBN 91-7372-660-5.

No 109 **Peter Fritzson:** Towards a Distributed Programming Environment based on Incremental Compilation,1984, ISBN 91-7372-801-2.

No 111 **Erik Tengvald:** The Design of Expert Planning Systems. An Experimental Operations Planning System for Turning, 1984, ISBN 91-7372-805-5.

No 155 **Christos Levcopoulos:** Heuristics for Minimum Decompositions of Polygons, 1987, ISBN 91-7870-133-3.

No 165 **James W. Goodwin:** A Theory and System for Non-Monotonic Reasoning, 1987, ISBN 91-7870-183-X.

No 170 **Zebo Peng:** A Formal Methodology for Automated Synthesis of VLSI Systems, 1987, ISBN 91-7870-225-9.

No 174 **Johan Fagerström:** A Paradigm and System for Design of Distributed Systems, 1988, ISBN 91-7870-301-8.

No 192 **Dimiter Driankov***:* Towards a Many Valued Logic of Quantified Belief, 1988, ISBN 91-7870-374-3.

No 213 **Lin Padgham:** Non-Monotonic Inheritance for an Object Oriented Knowledge Base, 1989, ISBN 91-7870-485-5.

No 214 **Tony Larsson:** A Formal Hardware Description and Verification Method, 1989, ISBN 91-7870-517-7.

No 221 **Michael Reinfrank:** Fundamentals and Logical Foundations of Truth Maintenance, 1989, ISBN 91-7870-546-0.

No 239 **Jonas Löwgren:** Knowledge-Based Design Support and Discourse Management in User Interface Management Systems, 1991, ISBN 91-7870-720-X.

No 244 **Henrik Eriksson:** Meta-Tool Support for Knowledge Acquisition, 1991, ISBN 91-7870-746-3.

No 252 **Peter Eklund:** An Epistemic Approach to Interactive Design in Multiple Inheritance Hierarchies,1991, ISBN 91-7870-784-6.

No 258 **Patrick Doherty:** NML3 - A Non-Monotonic Formalism with Explicit Defaults, 1991, ISBN 91-7870-816-8.

No 260 **Nahid Shahmehri:** Generalized Algorithmic Debugging, 1991, ISBN 91-7870-828-1.

No 264 **Nils Dahlbäck:** Representation of Discourse-Cognitive and Computational Aspects, 1992, ISBN 91-7870-850-8.

No 265 **Ulf Nilsson:** Abstract Interpretations and Abstract Machines: Contributions to a Methodology for the Implementation of Logic Programs, 1992, ISBN 91-7870-858-3.

No 270 **Ralph Rönnquist:** Theory and Practice of Tense-bound Object References, 1992, ISBN 91-7870-873-7.

No 273 **Björn Fjellborg:** Pipeline Extraction for VLSI Data Path Synthesis, 1992, ISBN 91-7870-880-X.

No 276 **Staffan Bonnier:** A Formal Basis for Horn Clause Logic with External Polymorphic Functions, 1992, ISBN 91-7870-896-6.

No 277 **Kristian Sandahl:** Developing Knowledge Management Systems with an Active Expert Methodology, 1992, ISBN 91-7870-897-4.

No 281 **Christer Bäckström:** Computational Complexity

of Reasoning about Plans, 1992, ISBN 91-7870-979-2.

No 292 **Mats Wirén:** Studies in Incremental Natural Language Analysis, 1992, ISBN 91-7871-027-8.

No 297 **Mariam Kamkar:** Interprocedural Dynamic Slicing with Applications to Debugging and Testing, 1993, ISBN 91-7871-065-0.

No 302 **Tingting Zhang:** A Study in Diagnosis Using Classification and Defaults, 1993, ISBN 91-7871-078-2.

No 312 **Arne Jönsson:** Dialogue Management for Natural Language Interfaces - An Empirical Approach, 1993, ISBN 91-7871-110-X.

No 338 **Simin Nadjm-Tehrani**: Reactive Systems in Physical Environments: Compositional Modelling and Framework for Verification, 1994, ISBN 91-7871-237-8.

No 371 **Bengt Savén:** Business Models for Decision Support and Learning. A Study of Discrete-Event Manufacturing Simulation at Asea/ABB 1968-1993, 1995, ISBN 91-7871-494-X.

No 375 **Ulf Söderman:** Conceptual Modelling of Mode Switching Physical Systems, 1995, ISBN 91-7871-516-4.

No 383 **Andreas Kågedal:** Exploiting Groundness in Logic Programs, 1995, ISBN 91-7871-538-5.

No 396 **George Fodor:** Ontological Control, Description, Identification and Recovery from Problematic Control Situations, 1995, ISBN 91-7871-603-9.

No 413 **Mikael Pettersson:** Compiling Natural Semantics, 1995, ISBN 91-7871-641-1.

No 414 **Xinli Gu:** RT Level Testability Improvement by Testability Analysis and Transformations, 1996, ISBN 91-7871-654-3.

No 416 **Hua Shu:** Distributed Default Reasoning, 1996, ISBN 91-7871-665-9.

No 429 **Jaime Villegas:** Simulation Supported Industrial Training from an Organisational Learning Perspective - Development and Evaluation of the SSIT Method, 1996, ISBN 91-7871-700-0.

No 431 **Peter Jonsson:** Studies in Action Planning: Algorithms and Complexity, 1996, ISBN 91-7871-704-3.

No 437 **Johan Boye:** Directional Types in Logic Programming, 1996, ISBN 91-7871-725-6.

No 439 **Cecilia Sjöberg:** Activities, Voices and Arenas: Participatory Design in Practice, 1996, ISBN 91-7871-728-0.

No 448 **Patrick Lambrix:** Part-Whole Reasoning in Description Logics, 1996, ISBN 91-7871-820-1.

No 452 **Kjell Orsborn:** On Extensible and Object-Relational Database Technology for Finite Element Analysis Applications, 1996, ISBN 91-7871-827-9.

No 459 **Olof Johansson:** Development Environments for Complex Product Models, 1996, ISBN 91-7871-855-4.

No 461 **Lena Strömbäck:** User-Defined Constructions in Unification-Based Formalisms,1997, ISBN 91-7871-857-0.

No 462 **Lars Degerstedt:** Tabulation-based Logic Programming: A Multi-Level View of Query Answering, 1996, ISBN 91-7871-858-9.

No 475 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av hur ekonomiska styrsystem utformas och används efter företagsförvärv, 1997, ISBN 91-7871-914-3.

No 480 **Mikael Lindvall:** An Empirical Study of Requirements-Driven Impact Analysis in Object-Oriented Software Evolution, 1997, ISBN 91-7871-927-5.

No 485 **Göran Forslund**: Opinion-Based Systems: The Cooperative Perspective on Knowledge-Based Decision Support, 1997, ISBN 91-7871-938-0.

No 494 **Martin Sköld**: Active Database Management Systems for Monitoring and Control, 1997, ISBN 91-7219-002-7.

No 495 **Hans Olsén**: Automatic Verification of Petri Nets in a CLP framework, 1997, ISBN 91-7219-011-6.

No 498 **Thomas Drakengren:** Algorithms and Complexity for Temporal and Spatial Formalisms, 1997, ISBN 91-7219-019-1.

No 502 **Jakob Axelsson:** Analysis and Synthesis of Heterogeneous Real-Time Systems, 1997, ISBN 91-7219-035-3.

No 503 **Johan Ringström:** Compiler Generation for Data-Parallel Programming Langugaes from Two-Level Semantics Specifications, 1997, ISBN 91-7219-045-0.

No 512 **Anna Moberg:** Närhet och distans - Studier av kommunikationsmmönster i satellitkontor och flexibla kontor, 1997, ISBN 91-7219-119-8.

No 520 **Mikael Ronström:** Design and Modelling of a Parallel Data Server for Telecom Applications, 1998, ISBN 91-7219-169-4.

No 522 **Niclas Ohlsson:** Towards Effective Fault Prevention - An Empirical Study in Software Engineering, 1998, ISBN 91-7219-176-7.

No 526 **Joachim Karlsson:** A Systematic Approach for Prioritizing Software Requirements, 1998, ISBN 91-7219-184-8.

No 530 **Henrik Nilsson:** Declarative Debugging for Lazy Functional Languages, 1998, ISBN 91-7219-197-x.

No 555 **Jonas Hallberg:** Timing Issues in High-Level Synthesis,1998, ISBN 91-7219-369-7.

No 561 **Ling Lin:** Management of 1-D Sequence Data - From Discrete to Continuous, 1999, ISBN 91-7219-402-2.

No 563 **Eva L Ragnemalm:** Student Modelling based on Collaborative Dialogue with a Learning Companion, 1999, ISBN 91-7219-412-X.

No 567 **Jörgen Lindström:** Does Distance matter? On geographical dispersion in organisations, 1999, ISBN 91-7219-439-1.

No 582 **Vanja Josifovski:** Design, Implementation and

Evaluation of a Distributed Mediator System for Data Integration, 1999, ISBN 91-7219-482-0.

No 589 **Rita Kovordányi**: Modeling and Simulating Inhibitory Mechanisms in Mental Image Reinterpretation - Towards Cooperative Human-Computer Creativity, 1999, ISBN 91-7219-506-1.

No 592 **Mikael Ericsson:** Supporting the Use of Design Knowledge - An Assessment of Commenting Agents, 1999, ISBN 91-7219-532-0.

No 593 **Lars Karlsson:** Actions, Interactions and Narratives, 1999, ISBN 91-7219-534-7.

No 594 **C. G. Mikael Johansson:** Social and Organizational Aspects of Requirements Engineering Methods - A practice-oriented approach, 1999, ISBN 91-7219-541-X.

No 595 **Jörgen Hansson:** Value-Driven Multi-Class Overload Management in Real-Time Database Systems, 1999, ISBN 91-7219-542-8.

No 596 **Niklas Hallberg:** Incorporating User Values in the Design of Information Systems and Services in the Public Sector: A Methods Approach, 1999, ISBN 91-7219-543-6.

No 597 **Vivian Vimarlund:** An Economic Perspective on the Analysis of Impacts of Information Technology: From Case Studies in Health-Care towards General Models and Theories, 1999, ISBN 91-7219-544-4.

No 598 **Johan Jenvald:** Methods and Tools in Computer-Supported Taskforce Training, 1999, ISBN 91-7219-547-9.

No 607 **Magnus Merkel:** Understanding and enhancing translation by parallel text processing, 1999, ISBN 91-7219-614-9.

No 611 **Silvia Coradeschi:** Anchoring symbols to sensory data, 1999, ISBN 91-7219-623-8.

No 613 **Man Lin:** Analysis and Synthesis of Reactive Systems: A Generic Layered Architecture Perspective, 1999, ISBN 91-7219-630-0.

No 618 **Jimmy Tjäder:** Systemimplementering i praktiken - En studie av logiker i fyra projekt, 1999, ISBN 91-7219-657-2.

No 627 **Vadim Engelson:** Tools for Design, Interactive Simulation, and Visualization of Object-Oriented Models in Scientific Computing, 2000, ISBN 91-7219-709-9.

No 637 **Esa Falkenroth:** Database Technology for Control and Simulation, 2000, ISBN 91-7219-766-8.

No 639 **Per-Arne Persson:** Bringing Power and Knowledge Together: Information Systems Design for Autonomy and Control in Command Work, 2000, ISBN 91-7219-796-X.

No 660 **Erik Larsson:** An Integrated System-Level Design for Testability Methodology, 2000, ISBN 91-7219-890-7.

No 688 **Marcus Bjäreland:** Model-based Execution Monitoring, 2001, ISBN 91-7373-016-5.

No 689 **Joakim Gustafsson:** Extending Temporal Action Logic, 2001, ISBN 91-7373-017-3.

No 720 **Carl-Johan Petri:** Organizational Information Provision - Managing Mandatory and Discretionary Use of Information Technology, 2001, ISBN-91-7373-126-9.

No 724 **Paul Scerri:** Designing Agents for Systems with Adjustable Autonomy, 2001, ISBN 91 7373 207 9.

No 725 **Tim Heyer**: Semantic Inspection of Software Artifacts: From Theory to Practice, 2001, ISBN 91 7373 208 7.

No 726 **Pär Carlshamre:** A Usability Perspective on Requirements Engineering - From Methodology to Product Development, 2001, ISBN 91 7373 212 5.

No 732 **Juha Takkinen:** From Information Management to Task Management in Electronic Mail, 2002, ISBN 91 7373 258 3.

No 745 **Johan Åberg:** Live Help Systems: An Approach to Intelligent Help for Web Information Systems, 2002, ISBN 91-7373-311-3.

No 746 **Rego Granlund:** Monitoring Distributed Teamwork Training, 2002, ISBN 91-7373-312-1.

No 757 **Henrik André-Jönsson:** Indexing Strategies for Time Series Data, 2002, ISBN 917373-346-6.

No 747 **Anneli Hagdahl:** Development of IT-suppor-ted Inter-organisational Collaboration - A Case Study in the Swedish Public Sector, 2002, ISBN 91-7373-314-8.

No 749 **Sofie Pilemalm:** Information Technology for Non-Profit Organisations - Extended Participatory Design of an Information System for Trade Union Shop Stewards, 2002, ISBN 91-7373-318-0.

No 765 **Stefan Holmlid:** Adapting users: Towards a theory of use quality, 2002, ISBN 91-7373-397-0.

No 771 **Magnus Morin:** Multimedia Representations of Distributed Tactical Operations, 2002, ISBN 91-7373-421-7.

No 772 **Pawel Pietrzak:** A Type-Based Framework for Locating Errors in Constraint Logic Programs, 2002, ISBN 91-7373-422-5.

No 758 **Erik Berglund:** Library Communication Among Programmers Worldwide, 2002, ISBN 91-7373-349-0.

No 774 **Choong-ho Yi:** Modelling Object-Oriented Dynamic Systems Using a Logic-Based Framework, 2002, ISBN 91-7373-424-1.

No 779 **Mathias Broxvall:** A Study in the Computational Complexity of Temporal Reasoning, 2002, ISBN 91-7373-440-3.

No 793 **Asmus Pandikow:** A Generic Principle for Enabling Interoperability of Structured and Object-Oriented Analysis and Design Tools, 2002, ISBN 91-7373-479-9.

No 785 **Lars Hult:** Publika Informationstjänster. En studie av den Internetbaserade encyklopedins bruksegenskaper, 2003, ISBN 91-7373-461-6.

No 800 **Lars Taxén:** A Framework for the Coordination of Complex Systems´ Development, 2003, ISBN 91-7373-604-X

No 808 **Klas Gäre:** Tre perspektiv på förväntningar och förändringar i samband med införande av informa-

tionsystem, 2003, ISBN 91-7373-618-X.

No 821 **Mikael Kindborg:** Concurrent Comics - programming of social agents by children, 2003, ISBN 91-7373-651-1.

No 823 **Christina Ölvingson:** On Development of Information Systems with GIS Functionality in Public Health Informatics: A Requirements Engineering Approach, 2003, ISBN 91-7373-656-2.

No 828 **Tobias Ritzau:** Memory Efficient Hard Real-Time Garbage Collection, 2003, ISBN 91-7373-666-X.

No 833 **Paul Pop:** Analysis and Synthesis of Communication-Intensive Heterogeneous Real-Time Systems, 2003, ISBN 91-7373-683-X.

No 852 **Johan Moe:** Observing the Dynamic Behaviour of Large Distributed Systems to Improve Development and Testing - An Emperical Study in Software Engineering, 2003, ISBN 91-7373-779-8.

No 867 **Erik Herzog:** An Approach to Systems Engineering Tool Data Representation and Exchange, 2004, ISBN 91-7373-929-4.

No 872 **Aseel Berglund:** Augmenting the Remote Control: Studies in Complex Information Navigation for Digital TV, 2004, ISBN 91-7373-940-5.

No 869 **Jo Skåmedal:** Telecommuting's Implications on Travel and Travel Patterns, 2004, ISBN 91-7373-935-9.

No 870 **Linda Askenäs:** The Roles of IT - Studies of Organising when Implementing and Using Enterprise Systems, 2004, ISBN 91-7373-936-7.

No 874 **Annika Flycht-Eriksson:** Design and Use of Ontologies in Information-Providing Dialogue Systems, 2004, ISBN 91-7373-947-2.

No 873 **Peter Bunus:** Debugging Techniques for Equation-Based Languages, 2004, ISBN 91-7373-941-3.

No 876 **Jonas Mellin:** Resource-Predictable and Efficient Monitoring of Events, 2004, ISBN 91-7373-956-1.

No 883 **Magnus Bång:** Computing at the Speed of Paper: Ubiquitous Computing Environments for Healthcare Professionals, 2004, ISBN 91-7373-971-5

No 882 **Robert Eklund:** Disfluency in Swedish human-human and human-machine travel booking dialogues, 2004. ISBN 91-7373-966-9.

No 887 **Anders Lindström:** English and other Foreign Linquistic Elements in Spoken Swedish. Studies of Productive Processes and their Modelling using Finite-State Tools, 2004, ISBN 91-7373-981-2.

No 889 **Zhiping Wang:** Capacity-Constrained Production-inventory systems - Modellling and Analysis in both a traditional and an e-business context, 2004, ISBN 91-85295-08-6.

No 893 **Pernilla Qvarfordt:** Eyes on Multimodal Interaction, 2004, ISBN 91-85295-30-2.

No 910 **Magnus Kald:** In the Borderland between Strategy and Management Control - Theoretical Framework and Empirical Evidence, 2004, ISBN 91-85295-82-5.

No 918 **Jonas Lundberg:** Shaping Electronic News: Genre Perspectives on Interaction Design, 2004, ISBN 91-85297-14-3.

No 900 **Mattias Arvola:** Shades of use: The dynamics of interaction design for sociable use, 2004, ISBN 91-85295-42-6.

No 920 **Luis Alejandro Cortés:** Verification and Scheduling Techniques for Real-Time Embedded Systems, 2004, ISBN 91-85297-21-6.

No 929 **Diana Szentivanyi:** Performance Studies of Fault-Tolerant Middleware, 2005, ISBN 91-85297-58-5.

No 933 **Mikael Cäker:** Management Accounting as Constructing and Opposing Customer Focus: Three Case Studies on Management Accounting and Customer Relations, 2005, ISBN 91-85297-64-X.

No 937 **Jonas Kvarnström:** TALplanner and Other Extensions to Temporal Action Logic, 2005, ISBN 91-85297-75-5.

No 938 **Bourhane Kadmiry:** Fuzzy Gain-Scheduled Visual Servoing for Unmanned Helicopter, 2005, ISBN 91-85297-76-3.

No 945 **Gert Jervan:** Hybrid Built-In Self-Test and Test Generation Techniques for Digital Systems, 2005, ISBN: 91-85297-97-6.

No 946 **Anders Arpteg:** Intelligent Semi-Structured Information Extraction, 2005, ISBN 91-85297-98-4.

No 947 **Ola Angelsmark:** Constructing Algorithms for Constraint Satisfaction and Related Problems - Methods and Applications, 2005, ISBN 91-85297-99-2.

No 963 **Calin Curescu:** Utility-based Optimisation of Resource Allocation for Wireless Networks, 2005. ISBN 91-85457-07-8.

No 972 **Björn Johansson:** Joint Control in Dynamic Situations, 2005, ISBN 91-85457-31-0.

No 974 **Dan Lawesson:** An Approach to Diagnosability Analysis for Interacting Finite State Systems, 2005, ISBN 91-85457-39-6.

No 979 **Claudiu Duma:** Security and Trust Mechanisms for Groups in Distributed Services, 2005, ISBN 91-85457-54-X.

No 983 **Sorin Manolache:** Analysis and Optimisation of Real-Time Systems with Stochastic Behaviour, 2005, ISBN 91-85457-60-4.

No 986 **Yuxiao Zhao:** Standards-Based Application Integration for Business-to-Business Communications, 2005, ISBN 91-85457-66-3.

No 1004 **Patrik Haslum:** Admissible Heuristics for Automated Planning, 2006, ISBN 91-85497-28-2.

No 1005 **Aleksandra Tesanovic:** Developing Reusable and Reconfigurable Real-Time Software using Aspects and Components, 2006, ISBN 91-85497-29-0.

No 1008 **David Dinka:** Role, Identity and Work: Extending the design and development agenda, 2006, ISBN 91-85497-42-8.

No 1009 **Iakov Nakhimovski:** Contributions to the Modeling and Simulation of Mechanical Systems with Detailed Contact Analysis, 2006, ISBN 91-85497-43-X.

No 1013 **Wilhelm Dahllöf:** Exact Algorithms for Exact Satisfiability Problems, 2006, ISBN 91-85523-97-6.

No 1016 **Levon Saldamli:** PDEModelica - A High-Level Language for Modeling with Partial Differential Equations, 2006, ISBN 91-85523-84-4.

No 1017 **Daniel Karlsson:** Verification of Component-based Embedded System Designs, 2006, ISBN 91-85523-79-8.

No 1018 **Ioan Chisalita:** Communication and Networking Techniques for Traffic Safety Systems, 2006, ISBN 91-85523-77-1.

No 1019 **Tarja Susi:** The Puzzle of Social Activity - The Significance of Tools in Cognition and Cooperation, 2006, ISBN 91-85523-71-2.

No 1021 **Andrzej Bednarski:** Integrated Optimal Code Generation for Digital Signal Processors, 2006, ISBN 91-85523-69-0.

No 1022 **Peter Aronsson:** Automatic Parallelization of Equation-Based Simulation Programs, 2006, ISBN 91-85523-68-2.

No 1023 **Sonia Sangari:** Some Visual Correlates to Focal Accent in Swedish, 2006, ISBN 91-85523-67-4.

No 1035 **Vaida Jakoniene:** Integration of Biological Data, 2006, ISBN 91-85523-28-3.

Linköping Studies in Information Science

No 1 **Karin Axelsson:** Metodisk systemstrukturering- att skapa samstämmighet mellan informa-tionssystemarkitektur och verksamhet, 1998. ISBN-9172-19-296-8.

No 2 **Stefan Cronholm:** Metodverktyg och användbarhet - en studie av datorstödd metodbaserad systemutveckling, 1998. ISBN-9172-19-299-2.

No 3 **Anders Avdic:** Användare och utvecklare - om anveckling med kalkylprogram, 1999. ISBN-91-7219-606-8.

No 4 **Owen Eriksson:** Kommunikationskvalitet hos informationssystem och affärsprocesser, 2000. ISBN 91-7219-811-7.

No 5 **Mikael Lind:** Från system till process - kriterier för processbestämning vid verksamhetsanalys, 2001, ISBN 91-7373-067-X

No 6 **Ulf Melin:** Koordination och informationssystem i företag och nätverk, 2002, ISBN 91-7373-278-8.

No 7 **Pär J. Ågerfalk:** Information Systems Actability - Understanding Information Technology as a Tool for Business Action and Communication, 2003, ISBN 91-7373-628-7.

No 8 **Ulf Seigerroth:** Att förstå och förändra systemutvecklingsverksamheter - en taxonomi för metautveckling, 2003, ISBN91-7373-736-4.

No 9 **Karin Hedström:** Spår av datoriseringens värden - Effekter av IT i äldreomsorg, 2004, ISBN 91-7373-963-4.

No 10 **Ewa Braf:** Knowledge Demanded for Action - Studies on Knowledge Mediation in Organisations, 2004, ISBN 91-85295-47-7.

No 11 **Fredrik Karlsson:** Method Configuration - method and computerized tool support, 2005, ISBN 91-85297-48-8.

No 12 **Malin Nordström:** Styrbar systemförvaltning - Att organisera systemförvaltningsverksamhet med hjälp av effektiva förvaltningsobjekt, 2005, ISBN 91-85297-60-7.

No 13 **Stefan Holgersson:** Yrke: POLIS - Yrkeskunskap, motivation, IT-system och andra förutsättningar för polisarbete, 2005, ISBN 91-85299-43-X.