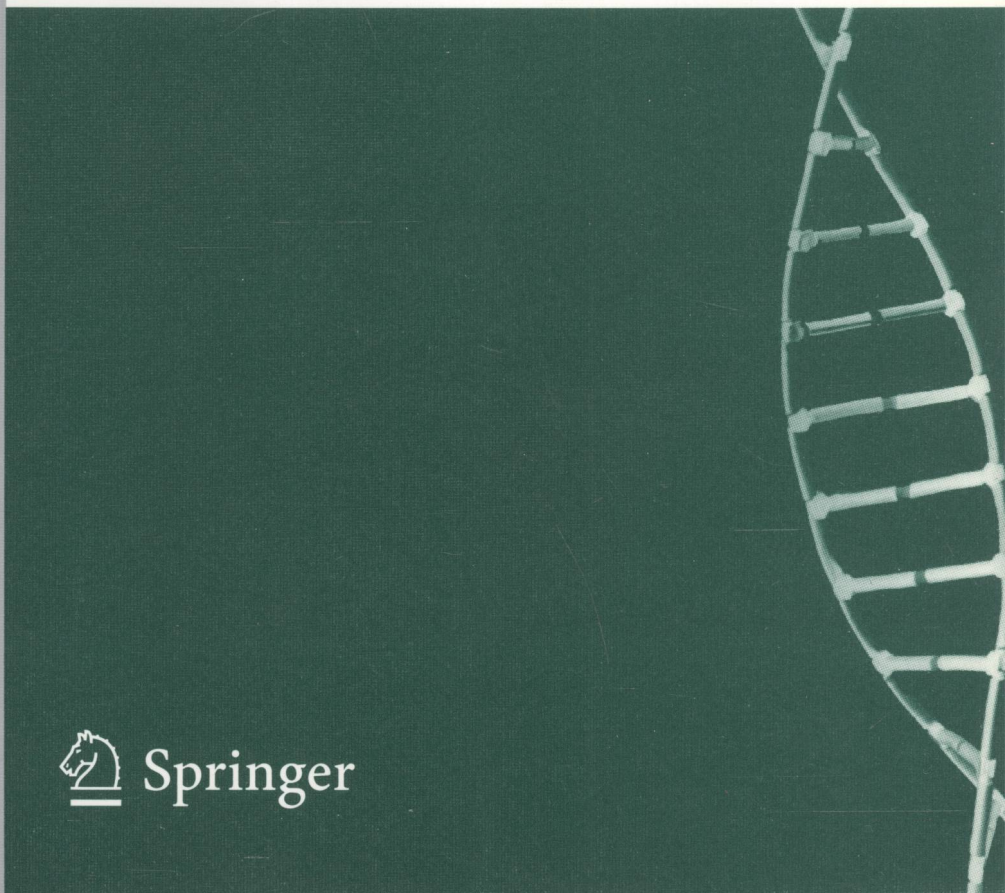
Ulf Leser
Felix Naumann
Barbara Eckman (Eds.)

LNBI 4075

Data Integration in the Life Sciences

**Third International Workshop, DILS 2006
Hinxton, UK, July 2006
Proceedings**



Springer

Ulf Leser Felix Naumann
Barbara Eckman (Eds.)

Data Integration in the Life Sciences

Third International Workshop, DILS 2006
Hinxton, UK, July 20-22, 2006
Proceedings

Springer

Series Editors

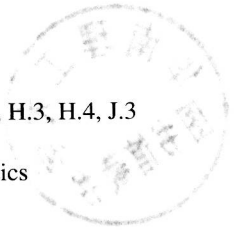
Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Ulf Leser
Felix Naumann
Humboldt-Universität zu Berlin, Institut für Informatik
Unter den Linden 6, 10099 Berlin, Germany
E-mail: {leser, naumann}@informatik.hu-berlin.de

Barbara Eckman
IBM Application and Integration Middleware
1475 Phoenixville Pike, West Chester, PA 19380, USA
E-mail: baeckman@us.ibm.com

Library of Congress Control Number: 2006928955



CR Subject Classification (1998): H.2, H.3, H.4, J.3

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-36593-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-36593-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11799511 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4075

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Preface

Data management and data integration are fundamental problems in the life sciences. Advances in molecular biology and molecular medicine are almost universally underpinned by enormous efforts in data management, data integration, automatic data quality assurance, and computational data analysis. Many hot topics in the life sciences, such as systems biology, personalized medicine, and pharmacogenomics, critically depend on integrating data sets and applications produced by different experimental methods, in different research groups, and at different levels of granularity. Despite more than a decade of intensive research in these areas, there remain many unsolved problems. In some respects, these problems are becoming more severe, both due to continuous increases in data volumes and the growing diversity in types of data that need to be managed. And the next big challenge is already upon us: the need to integrate the different "omics" data sets with the vast amounts of clinical data, collected daily in thousands of hospitals and physicians' offices all over the world.

DILS 2006 is the third in an annual workshop series that aims at fostering discussion, exchange, and innovation in research and development in the areas of data integration and data management for the life science. DILS 2004 in Leipzig and DILS 2005 in San Diego each attracted around 100 researchers from all over the world. This year the number of submitted papers again increased. The Program Committee selected 23 papers out of 50 strong full submissions. In an effort to include contributions that do not present a new method but that describe innovative and up-and-running practical systems, we distinguished "research papers" and "systems papers." The seven systems papers can be found in the sections Systems I and Systems II. Among the research papers there are four short papers and 12 full papers.

In addition to the presented papers, DILS 2006 featured two invited talks by Victor M. Markowitz and James H. Kaufmann and a session with updates on projects of world-wide importance: the Taverna eScience project, the BioMoby integration framework, and the BioMart integrated genomics data warehouse. Finally, there was a lively poster session.

The workshop was held at the Wellcome Trust Conference Center on the campus of the European Bioinformatics Institute (EBI) in Hinxton, UK. It was kindly sponsored by Microsoft Research, who also made available their conference management system, IBM Research, metanomics, metanomicshealth, the EBI industry programme, and by Schering AG. We are grateful for the help of Springer in putting together and publishing these proceedings. As Program Co-chairs we thank all authors who submitted their work, and the Program Committee members for their careful (and timely) reviews.

We particularly thank Paul Kersey of the EBI, who served as Local Chair of the workshop, and thus did all the hard work.

June 2006

Ulf Leser
Felix Naumann
Barbara Eckman

Organization

DILS 2006 Co-chairs

Ulf Leser	Humboldt-Universität zu Berlin, Germany
Felix Naumann	Humboldt-Universität zu Berlin, Germany
Barbara Eckman	IBM Healthcare and Life Sciences, USA

Local Chair

Paul Kersey, European Bioinformatics Institute, Hinxton, UK

Program Committee

Emmanuel Barillot	Institut Curie	France
David Benton	GlaxoSmithKline	USA
Laure Berti-Equille	Universitaire de Beaulieu	France
Peter Bunemann	University of Edinburgh	UK
Terence Critchlow	Lawrence Livermore National Laboratory	USA
Jürgen Eils	Deutsches Krebsforschungszentrum DKFZ	Germany
Floris Geerts	University of Edinburgh and Limburgs Universitair Centrum	UK
Amarnath Gupta	San Diego Supercomputer Center	USA
Joachim Hammer	University of Florida	USA
Henning Hermjakob	European Bioinformatics Institute	UK
Mike Hogarth	UC Davis	USA
Stefan Jablonski	Univ. Erlangen-Nuernberg	Germany
H V Jagadish	University of Michigan	USA
Hasan Jamil	Wayne State University	USA
Jacob Köhler	Rothamsted Research	UK
Peter Karp	SRI International	USA
Vipul Kashyap	Partners HealthCare System	USA
Arek Kasprzyk	European Bioinformatics Institute	UK
Anthony Kosky	Axiope Inc	USA
Bertram Ludäscher	UC Davis	USA
Paula Matuszek	GlaxoSmithKline Beecham	USA
Peter Mork	The MITRE Corporation	USA
Jignesh Patel	University of Michigan	USA
Norman Paton	University of Manchester	UK
Christian Piepenbrock	Epigenomics AG	Germany
Erhard Rahm	Universität Leipzig	Germany

Louiqa Raschid	University of Maryland	USA
Otto Ritter	AstraZeneca	USA
Monica Scannapieco	University of Rome "La Sapienza"	Italy
Dennis Paul Wall	Harvard Medical School	USA
Sharon Wang	IBM Healthcare and Life Sciences	USA
Bertram Weiss	Schering AG	Germany
Limsoon Wong	Institute for Infocomm Research	Singapore

Additional Reviewers

Shawn Bowers	Andrew Jones	Heiko Müller
Adriane Chapman	Toralf Kirsten	Eugene Novikov
Hon Nian Chua	Judice Koh	Loic Royer
Heiko Dietze	Jörg Lange	Donny Soh
Andreas Doms	Mario Latendresse	Silke Trissl
Nan Guo	Christian Lawerenz	Thomas Wächter
Michael Hartung	Timothy M. McPhillips	

Sponsoring Institutions

Microsoft Research	http://research.microsoft.com/
metanomics	http://www.metanomics.de/
metanomicshealth	http://www.metanomics-health.de/
IBM Research	http://www.research.ibm.com/
EBI Industry Program	http://industry.ebi.ac.uk/
Schering	http://www.schering.de/

Website

For more information please visit the DILS 2006 website at `http://www.informatik.hu-berlin.de/dils2006/`.

Lecture Notes in Bioinformatics

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R.C. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D. M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Table of Contents

Keynotes

Data Integration

Text Mining

Systems I

Potpourri

Systems II

Short Papers

Workflow

An Application Driven Perspective on Biological Data Integration
(Keynote Presentation)

Victor M. Markowitz

Lawrence Berkeley National Lab
Biological Data Management and Technology
1 Cyclotron Road, Berkeley, CA 94720
VMMarkowitz@lbl.gov

Data integration is an important part of biological applications that acquire data generated using evolving technologies and methods or involve data analysis across diverse specialized databases that reflect the expertise of different groups in a specific domain. The increasing number of such databases, the emergence of new types of data that need to be captured, as well as the evolving biological knowledge add to the complexity of already challenging integration problems. Furthermore, devising solutions to these problems requires technical expertise in several areas, such as database management systems, database administration and software engineering, as well as data modeling and analysis.

In practice, biological data integration is less daunting when considered in the context of scientific applications that address specific research questions. Established technologies and methods, such as database management systems, data warehousing tools, and statistical methods, have been employed successfully in developing systems that address such questions. The key challenge is marshaling the scientific and technical expertise required for formulating research questions, determining the integrated data framework for answering them, and addressing the underlying data semantics problems.

Evidence suggests that an iterative strategy based on gradually accumulating domain specific knowledge throughout the integration process is effective in devising solutions for application specific biological data integration problems. This strategy will be discussed in the context of two recently developed integrated genome systems, IMG (http://img.jgi.doe.gov) and IMG/M (http://img.jgi.doe.gov/m).

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, p. 1, 2006.

Towards a National Healthcare Information Infrastructure
(Keynote Presentation)

Sarah Knoop

IBM Healthcare Information Management,
IBM Almaden Research Center,
650 Harry Rd, San Jose, CA 95120
seknoop@us.ibm.com

Many countries around the world have placed an increased focus on the need to modernize their healthcare information infrastructure. This is particularly challenging in the United States. The U.S. healthcare industry is by far the largest in the world in both absolute dollars and in percentage of GDP (>$1.7T - 15% of GDP). It is also quite fragmented and complex. This complexity, coupled with an antiquated infrastructure for the collection of and access to medical data, leads to enormous inefficiencies and sources of error. Driven by consumer, regulatory, and governmental pressure, there is a growing consensus that the time has come to modernize the US Healthcare Information Infrastructure (HII). A modern HII will provide care givers with better and timelier access to data. The launch of a National Health Infrastructure Initiative (NHII) in the US in May 2004 - with the goal of providing an electronic health record for every American within the next decade- will eventually transform the healthcare industry in general...just as I/T has transformed other industries in the past. While such transformation may be disruptive in the short term, it will in the future significantly improve the quality, efficiency, and successful delivery of healthcare while decreasing costs to patients and payers and improving the overall experiences of consumers and providers. The key to this successful outcome will be based on the way we apply I/T to healthcare data and to the services delivered through that I/T. This must be accomplished in a way that protects individuals, allows competition, but gives caregivers reliable and efficient access to the data required to treat patients and to improve the practice of medical science.

In this talk we will describe the IBM Research HII project and our implementation of the standards for interoperability. We will also discuss how the same infrastructure required for interoperable electronic patient records must support the needs of medical science and public health. This can be accomplished by building higher level services upon a National Health Information Network, including discovery services for medical research and data mining and modeling services to protect populations against emerging infectious disease.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, p. 2, 2006.

Data Access and Integration in the ISPIDER Proteomics Grid

Lucas Zamboulis[1,2], Hao Fan[1,2,*], Khalid Belhajjame[3], Jennifer Siepen[3], Andrew Jones[3], Nigel Martin[1], Alexandra Poulovassilis[1], Simon Hubbard[3], Suzanne M. Embury[4], and Norman W. Paton[4]

[1] School of Computer Science and Information Systems, Birkbeck, Univ. of London
[2] Department of Biochemistry and Molecular Biology, University College London
[3] Faculty of Life Sciences, University of Manchester
[4] School of Computer Science, University of Manchester

Abstract. Grid computing has great potential for supporting the integration of complex, fast changing biological data repositories to enable distributed data analysis. One scenario where Grid computing has such potential is provided by proteomics resources which are rapidly being developed with the emergence of affordable, reliable methods to study the proteome. The protein identifications arising from these methods derive from multiple repositories which need to be integrated to enable uniform access to them. A number of technologies exist which enable these resources to be accessed in a Grid environment, but the independent development of these resources means that significant data integration challenges, such as heterogeneity and schema evolution, have to be met. This paper presents an architecture which supports the combined use of Grid data access (OGSA-DAI), Grid distributed querying (OGSA-DQP) and data integration (AutoMed) software tools to support distributed data analysis. We discuss the application of this architecture for the integration of several autonomous proteomics data resources.

1 Introduction

Grid computing technologies are becoming established which enable distributed computational and data resources to be accessed in a service-based environment. In the life sciences, these technologies offer the possibility of analysis of complex distributed post-genomic resources. To support transparent access, however, such heterogeneous resources need to be integrated rather than simply accessed in a distributed fashion. This paper presents an architecture for such integration and discusses the application of this architecture for the integration of several autonomous proteomics resources.

Proteomics is the study of the protein complement of the genome. It is a rapidly expanding group of technologies adopted by laboratories around the world as it is an essential component of any comprehensive functional genomics

* Currently at International School of Software, Wuhan University, China.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 3–18, 2006.

study targeted at the elucidation of biological function. This popularity stems from the increased availability and affordability of reliable methods to study the proteome, as well as the ever growing numbers of tertiary structures and genome sequences emerging from structural genomics and sequencing projects.

The *In Silico Proteome Integrated Data Environment Resource* (ISPIDER) project[1] aims to develop an integrated platform of proteome-related resources, using existing standards from proteomics, bioinformatics and e-Science. The integration of such resources would be extremely beneficial for a number of reasons. First, having access to more data leads to more reliable analyses; for example, performing protein identifications over an integrated resource would reduce the chances of false negatives. Second, bringing together resources containing different but closely related data increases the breadth of information the biologist has access to. Furthermore, the integration of these resources, as opposed to merely providing a common interface for accessing them, enables data from a range of experiments, tissues, or different cell states to be brought together in a form which may be analysed by a biologist in spite of the widely varying coverage and underlying technology of each resource.

In this paper we present an architecture which supports the combined use of Grid data access (OGSA-DAI), Grid distributed querying (OGSA-DQP) and data integration (AutoMed) software tools, together with initial results from the integration of three distributed, autonomous proteomics resources, namely gpmDB[2], Pedro[3] and PepSeeker[4]. The emergence of databases on experimental proteomics, capturing data from experiments on protein separation and identification, is very recent and we know of no previous work that combines data access, distributed querying and data integration of multiple proteomics databases as described here.

Paper outline: Section 2 gives an overview of the OGSA-DAI, OGSA-DQP and AutoMed technologies and introduces the three proteomics resources we have integrated. Section 3 discusses the development of the global schema integrating the proteomics resources within the ISPIDER project, Section 4 presents our new architecture, Section 5 discusses related work and Section 6 gives our conclusions and directions of further work.

2 Background

2.1 OGSA-DAI and OGSA-DQP

OGSA-DAI (Open Grid Services Architecture - Data Access and Integration) is an open-source, extendable middleware product exposing data resources on Grids via web services [2]. OGSA-DAI[5] supports both relational (MySQL, DB2,

[1] See http://www.ispider.man.ac.uk
[2] See http://gpmdb.thegpm.org
[3] See http://pedrodb.man.ac.uk:8080/pedrodb
[4] See http://nwsr.smith.man.ac.uk/pepseeker
[5] See http://www.ogsadai.org.uk/

SQL Server, Oracle, PostgreSQL), XML (Xindice, plans for eXist) and text data sources. It provides a uniform request format for a number of operations on data sources, including querying/updating, data transformation (XSLT), compression (ZIP/GZIP), and data delivery (FTP/SOAP).

OGSA-DQP (Open Grid Services Architecture - Distributed Query Processor) is a service-based distributed query processor [1], offering parallelism to support efficient querying of OGSA-DAI resources available in a grid environment. OGSA-DQP[6] offers two services, the Grid Distributed Query Service (GDQS) or Coordinator, and the Query Evaluation Service (QES) or Evaluator. The Coordinator uses resource metadata and computational resource information to compile, optimise, partition and schedule distributed query execution plans over multiple execution nodes in the Grid. The distributed evaluator services execute query plans generated by the Coordinator. Each Evaluator evaluates a partition of the query execution plan assigned to it by a Coordinator. A set of Evaluators participating in a query form a tree through which data flows from leaf Evaluators which interact with Grid data services, up the tree to reach its destination.

The following steps are needed for a client to set up a connection with OGSA-DQP and execute queries over OGSA-DAI resources. First, the client configures an appropriate GDQS data service resource. As a result of this process, the schemas of the resources are imported and the client is able to access one or more of the databases whose schemas have been referenced within a single query. The client then submits a Perform Document to OGSA-DQP containing an OQL [5] query. The Polar* [21] compiler parses, optimises and schedules the query. The query is partitioned, and each partition is sent to a different Evaluator. The Evaluators then interact with the OGSA-DAI resources and with each other, and send their results back to the GDQS, and, finally, the client.

2.2 AutoMed

AutoMed[7] is a heterogeneous data transformation and integration system which offers the capability to handle virtual, materialised and indeed hybrid data integration across multiple data models. It supports a low-level *hypergraph-based data model (HDM)* and provides facilities for specifying higher-level modelling languages in terms of this HDM. An HDM schema consists of a set of nodes, edges and constraints, and each modelling construct of a higher-level modelling language is specified as some combination of HDM nodes, edges and constraints. For any modelling language $\mathcal{M}$ specified in this way (via the API of AutoMed's Model Definitions Repository), AutoMed provides a set of primitive schema transformations that can be applied to schema constructs expressed in $\mathcal{M}$. In particular, for every construct of $\mathcal{M}$ there is an `add` and a `delete` primitive transformation which add to/delete from a schema an instance of that construct. For those constructs of $\mathcal{M}$ which have textual names, there is also a `rename` primitive transformation.

[6] See http://www.ogsadai.org.uk/about/ogsa-dqp/

[7] See http://www.doc.ic.ac.uk/automed

AutoMed schemas can be incrementally transformed by applying to them a sequence of primitive transformations, each adding, deleting or renaming just one schema construct (thus, in general, AutoMed schemas may contain constructs of more than one modelling language). A sequence of primitive transformations from one schema S_1 to another schema S_2 is termed a *pathway* from S_1 to S_2. All source, intermediate, and integrated schemas, and the pathways between them, are stored in AutoMed's Schemas & Transformations Repository.

Each `add` and `delete` transformation is accompanied by a query specifying the extent of the added or deleted construct in terms of the rest of the constructs in the schema. This query is expressed in a functional query language, IQL, and we will see some examples of IQL queries in Section 4.2. Also available are `extend` and `contract` primitive transformations which behave in the same way as `add` and `delete` except that they state that the extent of the new/removed construct cannot be precisely derived from the other constructs present in the schema. More specifically, each `extend` and `contract` transformation takes a pair of queries that specify a lower and an upper bound on the extent of the construct. The lower bound may be `Void` and the upper bound may be `Any`, which respectively indicate no known information about the lower or upper bound of the extent of the new construct.

The queries supplied with primitive transformations can be used to translate queries or data along a transformation pathway — we refer the reader to [15,14] for details. The queries supplied with primitive transformations also provide the necessary information for these transformations to be automatically *reversible*, in that each `add`/`extend` transformation is reversed by a `delete`/`contract` transformation with the same arguments, while each `rename` is reversed by a `rename` with the two arguments swapped.

As discussed in [15], this means that AutoMed is a *both-as-view (BAV)* data integration system: the `add`/`extend` steps in a transformation pathway correspond to Global-As-View (GAV) rules as they incrementally define target schema constructs in terms of source schema constructs; while the `delete` and `contract` steps correspond to Local-As-View (LAV) rules since they define source schema constructs in terms of target schema constructs. An in-depth comparison of BAV with other data integration approaches can be found in [15,14].

2.3 The Proteomics Resources

Thus far we have integrated three autonomous proteomics resources, all of which contain information on protein/peptide identification:

The Proteome Experimental Data RepOsitory (PEDRo [9]) provides access to a collection of descriptions of experimental data sets in proteomics. PEDRo was one of the first databases used for storing proteomics experimental data. It has also been used as a format for exchanging proteomics data, and in this respect has influenced the standardisation activities of the Proteomics Standards Initiative (PSI[8]).

The Global Proteome Machine Database (gpmDB [6]) is a publicly available database with over 2,200,000 proteins and almost 470,000 unique peptide

[8] See http://psidev.sourceforge.net