Data Exchange, Integration, and Streams

A Volume from DEIS'10 – GI-Dagstuhl Seminar 10452

Edited by
Phokion G. Kolaitis
Maurizio Lenzerini
Nicole Schweikardt

Dagstuhl Follow-Ups – Vol. 5 www.dagstuhl.de/dfu

Editors

Phokion G. Kolaitis
UC Santa Cruz & IBM Research – Almaden

Maurizio Lenzerini
Universitá di Roma La Sapienza

Nicole Schweikardt
Goethe-Universität Frankfurt am Main

ACM Classification 1998
H.2.5 Heterogeneous Databases, H.2.4 Systems, H.2.8 Database Applications, F.2.2 Nonnumerical Algorithms and Problems

ISBN 978-3-939897-61-3

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at http://www.dagstuhl.de/dagpub/978-3-939897-61-3.

Publication date
October, 2013

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Digital Object Identifier: 10.4230/DFU.Vol5.10452.i

DFU – Dagstuhl Follow-Ups

The series *Dagstuhl Follow-Ups* is a publication format which offers a frame for the publication of peer-reviewed papers based on Dagstuhl Seminars. DFU volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

ISSN 1868-8977

www.dagstuhl.de/dfu

Contents

Data Exchange, Integration, and Streams. *Dagstuhl Follow-Ups*, Volume 5, ISBN 978-3-939897-61-3.
Editors: Phokion G. Kolaitis, Maurizio Lenzerini, and Nicole Schweikardt
Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

Preface

This volume is based on GI-Dagstuhl Seminar 10452 on "Data Exchange, Integration, and Streams" (DEIS'10) held in November 2010. Before discussing the volume itself, we present some background and an overview of the DEIS'10 event, which we co-organized.

Background

The Schloss Dagstuhl – Leibniz Center for Informatics or, simply, Dagstuhl is known as the place "where computer scientists meet". Many computer scientists are familiar with the Dagstuhl seminars in which participants spend a week interacting with colleagues in an informal setting by sharing new results and work in progress, exchanging ideas, or embarking on new collaborations. Alongside these year-round seminars, however, Dagstuhl also hosts a different and less frequent type of event that is expressly geared towards students and postdoctoral scholars. Specifically, Dagstuhl is also the home of the GI-Dagstuhl Seminars[1], which are sponsored jointly by the German Informatics Society (GI) and the Schloss Dagstuhl – Leibniz Center for Informatics. The designated purpose of GI-Dagstuhl Seminars is to enable young researchers to learn about new developments in a particular area of research through active engagement in the seminar, which is typically organized by an international team of senior researchers. GI-Dagstuhl Seminars are typically limited to at most 20–25 participants, including the organizers.

In November of 2009, we submitted a proposal for a GI-Dagstuhl Seminar in the form of an advanced school on data exchange, data integration, and data streams. These are three different, yet inter-related, facets of information integration that have been investigated in depth by the research community in recent years.

Data exchange and data integration deal with the execution of information integration, but they adopt distinctly different approaches. Data exchange is the problem of transforming data residing in different sources into data structured under a target schema; in particular, data exchange entails the materialization of data, after the data have been extracted from the sources and re-structured into the unified format. In contrast, data integration can be described as symbolic or virtual integration: users are provided with the capability to pose queries and obtain answers via the unified format interface, while the data remain in the sources and no materialization of the restructured data is required.

In the basic data stream model, the input data consists of one or several streams of data items that can be read only sequentially, one after the other. This scenario is relevant for a large number of applications where massive amounts of data need to be processed. Typically, algorithms have to work with one or few passes over the data and a memory buffer of size significantly smaller than the input size.

Overview of the DEIS'10 Event

After our proposal was accepted, we disseminated the plan for the Advanced School on Data Exchange, Integration, and Streams (DEIS'10) via postings to a number of forums, including DBWorld, and through a dedicated web page at
`http://www.tks.cs.uni-frankfurt.de/events/deis10`

[1] `http://www.dagstuhl.de/en/program/gi-dagstuhl-seminars/`

Data Exchange, Integration, and Streams. *Dagstuhl Follow-Ups*, Volume 5, ISBN 978-3-939897-61-3.
Editors: Phokion G. Kolaitis, Maurizio Lenzerini, and Nicole Schweikardt
Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

Potential applicants were asked to submit by July 15, 2010 an application consisting of a letter of interest, a curriculum vitae, up to three representative papers or theses authored by the applicant, and a letter of recommendation from an academic supervisor or other senior colleague. We received 31 applications, out of which 22 applicants were selected to participate in DEIS'10; together with the organizers, this brought the total number of DEIS'10 participants to 25, which is the maximum that can be accommodated in a GI-Dagstuhl Seminar. The great majority of the applications received were of very high quality. In fact, we would have gladly accepted more applicants had there been more room. Of the 22 successful applicants, 18 were graduate students and 4 were postdoctoral scholars. In terms of geography, 18 were located in Europe, 3 in North America, and 1 in South America.

The participants were notified of their selection in early September 2010. Each participant was asked to study the relevant literature in a specialized topic that was assigned to him or her by the organizers of DEIS'10, based on the interests and expertise of the participants. Moreover, each participant was assigned one of the three organizers as mentor. Mentors and mentees interacted via email during September and October 2010. In particular, participants were asked to send their mentors a progress report with an outline of their presentation by the beginning of October 2010, which was followed by a semi-final draft of the slides of their presentation a week before DEIS'10 took place.

During the first day of DEIS'10, each of the three organizers gave a 90-minute tutorial on one of the three main themes of the school. Specifically, there was a tutorial on "Schema Mappings and Data Exchange" by Phokion Kolaitis, a tutorial on "Data Integration" by Maurizio Lenzerini, and a tutorial on "Data Streams" by Nicole Schweikardt. The rest of the program consisted of the presentations by the participants. Each participant was given 45 minutes to present an overview of the specialized topic assigned to her or him; the presentations were followed by or were interspersed with questions by the audience, so that a total of one hour was allotted to each specialized topic. The specialized topics covered during DEIS'10 were as follows.

Data Exchange: "The chase procedure and its applications to data exchange" by Andrian Onet; "Algorithms for computing the core of universal solutions" by Vadim Savenkov; "The inverse operator on schema mappings and its uses in data exchange" by Jorge Pérez; "Integrity constraints in data exchange" by Víctor Guttiérez-Basulto; "Semantics of query answering in data exchange and closed world reasoning" by André Hernich; "Analyzing, comparing and debugging schema mappings" by Emanuel Salinger; and "XML data exchange" by Amélie Gheerbrant.

Data Integration: "Query answering in data integration" by Piotr Wieczorek; "Data integration: consistent query answering" by Sławomir Staworko; "Data cleaning for data integration" by Ekaterini Ioannou; "Description logics for data integration" by Y. Angélica Ibáñez-Garcia; "View-based query processing" by Paolo Guagliardo; "Probabilistic data integration and probabilistic data exchange" by Livia Predoiu; "Learning and discovering queries and mappings" by Marie Jacob; "Theory of peer data management" by Sebastian Skritek; "Peer data management systems" by Armin Roth; and "XML data integration" by Lucja Kot.

Data Streams: "Basic algorithmic techniques for processing data streams" by Mariano Zelke; "Data stream management systems and query languages" by Sandra Geisler; "Querying and mining data streams" by Elena Ikonomovska; "Distributed processing of data streams and large data sets" by Marwan Hassani; and "Stream-based processing of XML documents" by Cristian Riveros.

While a small number of participants presented some of their own research work, most of the presentations were a synthesis of papers studied by the participants in the months before DEIS'10 took place. In total, well over 100 published papers were distilled and synthesized by the participants in their presentations. The slides of these presentations and the relevant bibliographical references can be found at the web page of DEIS'10.

In addition to the tutorials and the presentations of specialized topics, an after-dinner problem session was held in the second day of DEIS'10. In this session, both the organizers and the participants presented selected open problems in each of the three main themes of DEIS'10. The last time slot of DEIS'10 was a wrap-up session during which feedback about the event was solicited and tentative plans for a follow-up event were discussed.

Follow-Up

For some of the topics presented at DEIS'10, excellent survey articles already exist. Some other topics are still too nascent to justify survey articles at this point of time. For several more mature topics for which no survey articles presently exist, we felt that the time is ripe to produce such survey articles as a follow-up to DEIS'10. To this effect, we invited a number of DEIS'10 participants to contribute chapters to this volume. In several cases, we paired authors and asked them to co-author chapters that constitute a synthesis of their individual presentations at DEIS'10. Each draft chapter was peer-reviewed and subsequently revised to take into account the suggestions of the reviewers.

Overview of the Volume

The first four chapters in this volume examine several different, yet inter-related, aspects of data exchange. The underlying thread in these chapters is the systematic use of schema mappings, which are are high-level syntactic specifications that describe the relationship between two database schemas. Schema mappings have turned out to be the essential building blocks in formalizing and analyzing data inter-operability tasks, such as data exchange and data integration.

The first chapter of this volume, which is authored by Adrian Onet, gives a comprehensive overview of the properties of the chase procedure, an important algorithm that has been widely used to construct "good" solutions in data exchange and also to reason about schema mappings. The study of "good" solutions in data exchange is pursued in more depth in the second chapter, which is authored by Vadim Savenkov. Here, the focus is on the algorithmic properties of core universal solutions, which, intuitively, are the "best" solutions to materialize in data exchange. The third chapter, which is authored by Jorge Pérez, examines the various approaches that have been taken towards giving precise semantics and studying the properties of the inverse operator on schema mappings, an operator that, as the name suggests, is intended to "reverse" the action of the given schema mapping. The fourth chapter, authored by Emanuel Sallinger, presents an overview of the concepts introduced and the methods developed to reason about schema mappings with emphasis on optimality and equivalence between schema mappings.

The next four chapters explore different aspects of data integration. All chapters are centered around what is considered the main problem in this form of information integration, namely processing queries posed to the data integration system.

The first one, which is the fifth chapter of this volume, authored by Paolo Guagliardo and Piotr Wieczorek, provides an overview of the techniques for computing the answers to queries posed to the global schema of a virtual data integration system, both in the case where the

global schema is expressed in the relational model, and in the case where a semi-structured data model is used instead. While most papers on query processing in data integration concentrate on positive queries, the sixth chapter of this volume, authored by André Hernich, addresses the issue of selecting the right semantics and the right algorithms for answering non-monotone queries, both in data integration and in data exchange. The seventh paper, which is authored by Armin Roth and Sebastian Skritek, deals with a sophisticated form of information integration that is receiving great attention in the last years, namely peer data integration. Unlike traditional data integration systems, peer data integration systems do not rely on a unique global schema. Instead, they allow for full autonomy of a set of data sources, with mappings between them, and no need of central coordinator. The eighth chapter, authored by Ekaterini Ioannou and Sławek Staworko, provides a discussion on techniques introduced for handling inconsistencies. Query processing in data integration is often studied under the assumption that the data integration system is logically consistent. However, this is an unrealistic assumption in many real world contexts. The chapter illustrates two main approaches to deal with inconsistencies, based on "on-line" consistent query answering, and methods for resolving the inconsistencies "off-line", respectively.

The last two chapters deal with data streams. Both chapters are centered around the question of how to efficiently process massive amounts of data in a real-time manner, using memory buffers that are significantly smaller than the input data.

The first one, which is the ninth chapter of this volume, authored by Elena Ikonomovska and Mariano Zelke, gives an overview of algorithmic techniques for data stream processing. It presents abstract models for data stream processing and contains a tutorial on fundamental techniques for sampling and sketching data, as well as a survey of algorithmic approaches for similarity mining, group testing, clustering, and summarizing data streams. The tenth chapter, authored by Sandra Geisler, gives an overview of data stream management systems (DSMS), i.e., database management systems specifically designed for processing data streams. It gives details on the architecture of DSMS, surveys existing systems and query languages, and discusses methods for monitoring the data quality of DSMS.

Acknowledgments

We are grateful to the Schloss Dagstuhl – Leibniz Center for Informatics and to the German Informatics Society (GI) for giving us the opportunity to organize DEIS'10. We also wish to acknowledge the German Research Foundation (DFG) and the ACM Special Interest Group on Management Of Data (SIGMOD) for their generous financial support, which made it possible to provide travel support to participants of DEIS'10. We are particularly grateful to Dr. Marc Herbstritt, Member of the Scientific Staff of the Schloss Dagstuhl – Leibniz Center for Informatics, for his help and encouragement throughout the process of organizing DEIS'10 and preparing this volume. Finally, we wish to thank the participants of DEIS'10, the authors of the chapters in this volume, and the anonymous reviewers of the chapters. This volume would not have been possible without the contributions of all these colleagues.

June 2013 *Phokion G. Kolaitis, Maurizio Lenzerini, and Nicole Schweikardt*