Data Integration

Datenintegration

Melanie Herschel, Université Paris Sud 11, France

Data integration aims at combining data that resides in distributed, autonomous, and heterogeneous databases into a single consistent view of the data. Its applications are abundant, ranging from data integration in scientific data (helping scientists share, understand, reuse and complement past results) to data integration in enterprises (for instance to set up data warehouses, perform business intelligence or implement master data management) and data integration on the Web (comparison online shopping or linking open data). In order to achieve data integration, three major problems have been of particular interest to both the database research community and IT industry. First, the heterogeneity between data models and schemas of data sources has to be overcome. Second, data sources may overlap in the sets of real-world entities such as persons or products they represent and their multiple and usually different representations of a same entity, so called duplicates, need to be identified. Finally, in the integrated result, every entity should be represented exactly once, so duplicates need to be merged to a single representation, a problem also referred to as data fusion. This special issue covers some solutions and new challenges related to data integration, both from a research and an industrial perspective.

The article by Mecca and Papotti describes the state-of-the art of schema mapping and data exchange solutions, employed to address the first of the above steps by bridging the heterogeneity between data models and schemas of data sources to be integrated. Schema mapping techniques have acquired great popularity due to their declarative nature, clean semantics, easy to use design tools and their efficiency and modularity in the deployment step. The article divides the surveyed approaches into three ages: the heroic age that produced the theoretical foundations and early tools, the silver age when schema mapping tools have grown their way into complex systems and have been translated into both commercial and open-source tools, and a forthcoming golden age with novel research opportunities and a new generation of systems capable of dealing with a significantly larger class of real-life applications.

Maier, Oberhofer and Schwarz describe a commercial approach to data integration that addresses the three main problems above (among others) and that has been used in large data integration projects worldwide. This approach is based on the observation that the largest part of a typical data integration effort is dedicated to the implementation of transformation, cleansing, and data validation logic in robust and highly performing commercial systems. This effort is simple and does not demand skills beyond commercial product knowledge, but it is very labour-intensive and error prone. Their approach helps to industrialize data integration projects and significantly lowers the amount of simple, but labour-intensive work. The key idea is that the target landscape for a data integration project has pre-defined data models and associated meta data which can be leveraged for building and automating the data integration process.

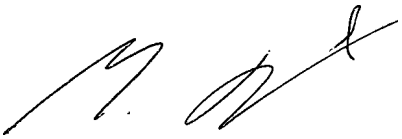
In her article on multi-scale data integration, Berti-Equille presents some challenging research directions for integrating massive multi-scale scientific data from the observational science domain. This data is intensively collected in order to measure various properties of the Earth. For instance, scientists observe environmental conditions, ecosystems, or biological species. The ability to understand complex phenomena such as global warming and to predict trends from spatio-temporal data have become a major issue in observational science, for which theoretical and technical advances in multi-scale data integration are essential. The paper describes several use cases of data integration in observational sciences and outlines challenge due to temporal, spatial, structural, semantic or analytic dependencies, different levels of data granularity or data abstraction from raw measurement data to processed data and derived statistics, various data interpretations or usages depending on the disciplines, quality heterogeneity of spatio-temporal data, and scaling issues.

The next article by Braunschweig, Ebarius, Thiele and Lehner describes an integration system targeted towards open web data. This data is nowadays only insufficiently examined or not considered at all in decision making processes. This is because of the lack of end-user friendly tools that help to reuse this public data and to derive knowledge from it. To address this shortcoming, they propose a data repository that does not necessitate a schema, unlike traditional structured data sources commonly used by decision making processes. This provides the flexibility necessary to store and gradually integrate heterogeneous web data. Based on this repository, semi-automatic schema enrichment efficiently augments the data in a "pay-as-you-go" fashion. To resolve ambiguities in the integrated data, the authors propose a crowd-based verification component that is able to resolve such conflicts in a scalable manner.

The final article by van Keulen describes how managing uncertainty may lead to better data interoperability. Data interoperability encompasses the many data management activities needed for effective information management in anyone's or any organization's everyday work such as data cleaning, coupling, fusion, mapping, and information extraction. Semantic uncertainty appears in all these tasks, because sometimes data is subjective, incomplete, not current, incorrect or it can be interpreted in different ways. The paper proposes to treat data quality problems as a fact of life, not as something to be repaired afterwards. Van Keulen first describes for several data interoperability use cases, including duplicate detection and data fusion, how to formally model the associated data quality problems as semantic uncertainty. Furthermore, he provides an argument why the proposed approach leads to better data interoperability in terms of natural problem exposure and risk assessment, more robustness and automation, reduced development costs, and potential for natural and effective feedback loops leveraging human attention.

Despite covering diverse aspects of data integration within the five articles, this special issue can only provide a glimpse of all topics relevant to that area. The goal of this special issue was to provide the reader with some insight into both research problems and real-life applications of data integration, ranging from state-of-the-art approaches to new emerging trends, thus conveying a perspective on the past, the present and the future of data integration.

This special issue being the result of many people's work, I especially thank the editors of "it", the authors and the reviewers for their excellent work and fruitful collaboration.



Melanie Herschel

Melanie Herschel graduated in Information Technology from the University of Cooperative Education Stuttgart in 2003. She performed research on duplicate detection and data integration at the Humboldt University of Berlin and at the Hasso Plattner Institute in Potsdam. She completed her PhD in 2007 and from 2008 to 2009 worked at the IBM Almaden Research Center on topics related to data provenance. In 2009, she joined the University of Tübingen to pursue her research on data quality, data integration, and data provenance. Since 2011, she is an associate professor at the University of Paris 11.

Address: Laboratoire de Recherche en Informatique (LRI), Université Paris Sud 11, F-91405 Orsay Cedex, e-mail: melanie.herschel@lri.fr