1

Data Integration — Problems, Approaches, and Perspectives

Patrick Ziegler and Klaus R. Dittrich
{pziegler, dittrich}@ifi.unizh.ch

Database Technology Research Group
Department of Informatics, University of Zurich

Summary. Data integration is one of the older research fields in the database area and has emerged shortly after database systems were first introduced into the business world. In this paper, we briefly introduce the problem of integration and, based on an architectural perspective, give an overview of approaches to address the integration issue. We discuss the evolution from structural to semantic integration and shortly present our own research in the SIRUP (Semantic Integration Reflecting User-specific semantic Perspectives) approach. Finally, an outlook to challenging areas of future research in the realm of data integration is given.

1.1 Introduction

In today's business world, it is typical that enterprises run different but coexisting information systems. Employing these systems, enterprises struggle to realize business opportunities in highly competitive markets. In this setting, the integration of existing information systems is becoming more and more indispensable in order to dynamically meet business and customer needs while leveraging long-term investments in existing IT infrastructure.

In general, integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. The reason for integration is twofold: First, given a set of existing information systems, an integrated view can be created to facilitate information access and reuse through a single information access point. Second, given a certain information need, data from different complementing information systems is combined to gain a more comprehensive basis to satisfy the need.

There is a manifold of applications that benefit from integrated information. For instance, in the area of business intelligence (BI), integrated information can be used for querying and reporting on business activities, for statistical analysis, online analytical processing (OLAP), and data mining in order to enable forecasting, decision making, enterprise-wide planning, and, in the

end, to gain sustainable competitive advantages. For customer relationship management (CRM), integrated information on individual customers, business environment trends, and current sales can be used to improve customer services. Enterprise information portals (EIP) present integrated company information as personalized web sites and represent single information access points primarily for employees, but also for customers, business partners, and the public. Last, but not least, in the area of e-commerce and e-business, integrated information enables and facilitates business transactions and services over computer networks.

Similar to information, IT services and applications can be integrated, either to provide a single service access point or to provide more comprehensive services to meet business requirements. For instance, integrated workflow and document management systems can be used within enterprises to leverage intraorganizational collaboration. Based on the ideas of business process reengineering (BPR), integrated IT services and applications that support business processes can help to reduce time-to-market and to provide added-value products and services. Thereby, interconnecting building blocks from selected IT services and applications enables supply chain management within individual enterprises as well as cooperation beyond the boundaries of traditional enterprises, as in interorganizational cooperation, business process networks (BPN), and virtual organizations. For instance, in e-procurement, supply and demand for producer goods are provided with integrated information and services to streamline the purchasing process for institutional buyers. Thus, it is possible to bypass intermediaries and to enable direct interaction between supply and demand, as in business-to-business (B2B), business-to-consumer (B2C), and business-to-employee (B2E) transactions. These trends are fueled by XML that is becoming *the* industry standard for data exchange as well as by web services that provide interoperability between various software applications running on different platforms.

In the enterprise context, the integration problem is commonly referred to as enterprise integration (EI). Enterprise integration denotes the capability to integrate information and functionalities from a variety of information systems in an enterprise. This encompasses enterprise information integration (EII) that concerns integration on the data and information level and enterprise application integration (EAI) that considers integration on the level of application logic. In this paper, we focus on the integration of information and, in particular, highlight integration solutions that are provided by the database community. Our goal is to give, based on an architectural perspective, a database-centric overview of principal approaches to the integration problem and to illustrate some frequently used approaches. Additionally, we introduce semantic integration that is needed in all integration examples given above and that forms a key factor for current and future integration solutions. An outlook to our own approach to personal semantic data integration and future research challenges round off this paper which is an extension of [37].

The structure of this paper is as follows: In the following Section, we sketch the problem of integration. Section 1.3 presents principal approaches to address the integration issue and in Section 1.4, the evolution from structural to current semantic integration approaches is discussed. Our work in the SIRUP project is outlined in Section 1.5 and then, an outlook to challenging areas of future data integration research is given. Finally, Section 1.7 concludes the paper.

1.2 The Problem of Integration

Integration of multiple information systems generally aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. Users are provided with a homogeneous logical view of data that is physically distributed over heterogeneous data sources. For this, all data has to be represented using the same abstraction principles (unified global data model and unified semantics). This task includes detection and resolution of schema and data conflicts regarding structure and semantics.

In general, information systems are not designed for integration. Thus, whenever integrated access to different source systems is desired, the sources and their data that do not fit together have to be coalesced by additional adaptation and reconciliation functionality. Note that there is not *the* one single integration problem. While the goal is always to provide a homogeneous, unified view on data from different sources, the particular integration task may depend on:

- the architectural view of an information system (see Figure 1.1),
- the content and functionality of the component systems,
- the kind of information that is managed by component systems (alphanumeric data, multimedia data; structured, semi-structured, unstructured data),
- requirements concerning autonomy of component systems,
- intended use of the integrated information system (read-only or write access),
- performance requirements, and
- the available resources (time, money, human resources, know-how, etc.) [12].

Additionally, several kinds of heterogeneity typically have to be considered. These include differences in:

- hardware and operating systems,
- data management software,
- data models, schemas, and data semantics,
- middleware,

- user interfaces, and
- business rules and integrity constraints.

1.3 Approaches to Integration

In this section, we apply an architectural perspective to give an overview of the different ways to address the integration problem. The presented classification is based on [12] and distinguishes integration approaches according to the level of abstraction where integration is performed.

Information systems can be described using a layered architecture, as shown in Figure 1.1: On the topmost layer, users access data and services through various interfaces that run on top of different applications. Applications may use middleware — transaction processing (TP) monitors, message-oriented middleware (MOM), SQL-middleware, etc. — to access data via a data access layer. The data itself is managed by a data storage system. Usually, database management systems (DBMS) are used to combine the data access and storage layer.

In general, the integration problem can be addressed on each of the presented system layers. For this, the following principal approaches — as illustrated in Figure 1.1 — are available:

Manual Integration

Here, users directly interact with all relevant information systems and manually integrate selected data. That is, users have to deal with different user interfaces and query languages. Additionally, users need to have detailed knowledge on location, logical data representation, and data semantics.

Common User Interface

In this case, the user is supplied with a common user interface (e.g., a web browser) that provides a uniform look and feel. Data from relevant information systems is still separately presented so that homogenization and integration of data yet has to be done by the users (for instance, as in search engines).

Integration by Applications

This approach uses integration applications that access various data sources and return integrated results to the user. This solution is practical for a small number of component systems. However, applications become increasingly fat as the number of system interfaces and data formats to homogenize and integrate grows.

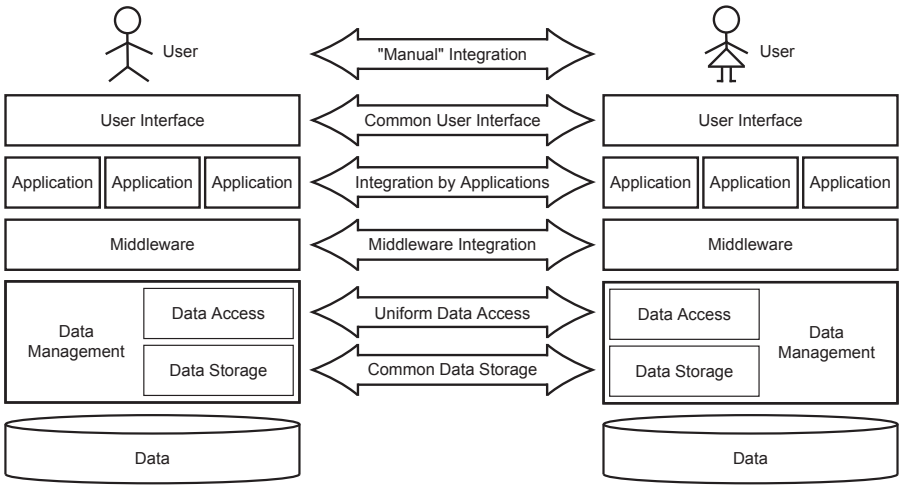

Fig. 1.1. General Integration Approaches on Different Architectural Levels

Integration by Middleware

Middleware provides reusable functionality that is generally used to solve dedicated aspects of the integration problem, e.g., as done by SQL-middleware. While applications are relieved from implementing common integration functionality, integration efforts are still needed in applications.[1] Additionally, different middleware tools usually have to be combined to build integrated systems.

Uniform Data Access

In this case, a logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available on this level. Local information systems keep their autonomy and can support additional data access layers for other applications. However, global provision of physically integrated data can be time-consuming since data access, homogenization, and integration have to be done at runtime.

Common Data Storage

Here, physical data integration is performed by transferring data to a new data storage; local sources can either be retired or remain operational. In

[1] For instance, SQL-middleware provides a single access point to send SQL queries to all connected component systems. However, query results are not integrated into one single, homogeneous result set.

general, physical data integration provides fast data access. However, if local data sources are retired, applications that access them have to be migrated to the new data storage as well. In case local data sources remain operational, periodical refreshing of the common data storage needs to be considered.

In practice, concrete integration solutions are realized based on the presented six general integration approaches. Important examples include:

- *Mediated query systems* represent a uniform data access solution by providing a single point for read-only querying access to various data sources, e.g., as in TSIMMIS [9]. A mediator [34] that contains a global query processor is employed to send subqueries to local data sources; returned local query results are then combined.
- *Portals* as another form of uniform data access are personalized doorways to the internet or intranet where each user is provided with information according to his detected information needs. Usually, web mining is applied to determine user-profiles by click-stream analysis; thereby, information the user might be interested in can be retrieved and presented.
- *Data warehouses* realize a common data storage approach to integration. Data from several operational sources (on-line transaction processing systems, OLTP) are extracted, transformed, and loaded (ETL) into a data warehouse. Then, analysis, such as online analytical processing (OLAP), can be performed on cubes of integrated and aggregated data.
- *Operational data stores* are a second example of a common data storage. Here, a "warehouse with fresh data" is built by immediately[2] propagating updates in local data sources to the data store. Thus, up-to-date integrated data is available for decision support. Unlike in data warehouses, data is neither cleansed nor aggregated nor are data histories supported.
- *Federated database systems (FDBMS)* achieve a uniform data access solution by logically integrating data from underlying local DBMS. Federated database systems are fully-fledged DBMS; that is, they implement their own data model, support global queries, global transactions, and global access control. Usually, the five-level reference architecture by [30] is employed for building FDBMS.
- *Workflow management systems (WFMS)* allow to implement business processes where each single step is executed by a different application or user. Generally, WFMS support modeling, execution, and maintenance of processes that are comprised of interactions between applications and human users. WFMS represent an integration-by-application approach.
- *Integration by web services* performs integration through software components (i.e., web services) that support machine-to-machine interaction over a network by XML-based messages that are conveyed by internet protocols. Depending on their offered integration functionality, web services

[2] That is, not within the same transaction but within a period of time that is reasonable according to the particular application requirements.

either represent a uniform data access approach or a common data access interface for later manual or application-based integration.

- *Model management* introduces high-level operations between models (such as database schemas, UML models, and software configurations) and model mappings; such operations include matching, merging, selection, and composition [6]. Using a schema algebra that encompasses all these operations, it is intended to reduce the amount of hand-crafted code required for transformations of models and mappings as needed for schema integration. Model management falls into the category of manual integration.
- *Peer-to-peer (P2P) integration* is a decentralized approach to integration between distributed, autonomous peers where data can be mutually shared and integrated through mappings between local schemas of peers. P2P integration constitutes, depending on the provided integration functionality, either a uniform data access approach or a data access interface for subsequent manual or application-based integration.
- *Grid data integration* provides the basis for hypotheses testing and pattern detection in large amounts of data in grid environments, i.e., interconnected computing resources being used for high-throughput computing. Here, often unpredictable and highly dynamic amounts of data have to be dealt with to provide an integrated view over large (scientific) data sets. Grid data integration represents an integration by middleware approach.
- *Personal data integration systems* (e.g., [38]) are a special form of manual integration. Here, tailored integrated views are defined (e.g., by a declarative integration language), either by users themselves or by dedicated integration engineers. Each integrated view precisely matches the information needs of a user by encompassing all relevant entities with real-world semantics as intended by the particular user; thereby, the integrated view reflects the user's personal way to perceive his application domain of interest.
- *Collaborative integration* (e.g., [25]), another special form of manual integration, is based on the idea to have users to contribute to a data integration system for using it. Here, initial partial schema mappings are presented to users who answer questions concerning the mappings; these answers are then taken to refine the mappings and to expand the system capabilities. Similar to folksonomies, where data is collaboratively labeled for later retrieval, the task of schema mapping is distributed over participating users.
- In *Dataspace systems* [13], co-existence of all data (i.e., both structured and unstructured) is propagated rather than full integration. A dataspace system is used to provide the same basic functionality, e.g., search facilities, over all data sources independently of their degree of integration. Only when more sophisticated services are needed, such as relational-style queries, additional efforts are made to integrate the required data sources more closely. In general, dataspace systems may simultaneously use every one of the presented six general integration approaches.

1.4 From Structural to Semantic Integration

Database technology was introduced in enterprises since the late 1960s to support (initially rather simple) business applications. As the number of applications and data repositories rapidly grew, the need for integrated data became apparent. As a consequence, first integration approaches in the form of multidatabase systems [21] were developed around 1980 — e.g., MULTIBASE [24]. This was a first cornerstone in a remarkable history of research in the area of data integration. The evolution continued over mediators (e.g., Garlic [8]) and agent systems (e.g., InfoSleuth [4]) to ontology-based (e.g., OBSERVER [26]), peer-to-peer (P2P) (e.g., Piazza [19]), and web service-based integration approaches (e.g., Active XML [1]). Recently, tailored personal data integration (e.g., SIRUP [38]), collaborative integration (e.g., MOBS [25]), and dataspace systems [13] are being addressed by the research community (see Figure 1.2).

In general, early integration approaches were based on a relational or functional data model and realized rather tightly-coupled solutions by providing one single global schema. To overcome their limitations concerning the aspects of abstraction, classification, and taxonomies, object-oriented integration approaches [7] were adopted to perform structural homogenization and integration of data. With the advent of the internet and web technologies, the focus shifted from integrating purely well-structured data to also incorporating semi- and unstructured data while architecturally, loosely-coupled mediator and agent systems became popular.

However, integration is more than just a structural or technical problem. Technically, it is rather easy to connect different relational DBMS (e.g., via ODBC or JDBC). More demanding is to integrate data described by different data models; even worse are the problems caused by data with heterogeneous semantics. For instance, having only the name "loss" to denote a relation in an enterprise information system does not provide sufficient information to doubtlessly decide whether the represented loss is a book loss, a realized loss, or a future expected loss and whether the values of the tuples reflect only a roughly estimated loss or a precisely quantified loss. Integrating two "loss" relations with (implicit) heterogeneous semantics leads to erroneous results and completely senseless conclusions. Therefore, explicit and precise semantics of integratable data are essential for semantically correct and meaningful integration results. Note that none of the principal integration approaches in Section 1.3 helps to resolve semantic heterogeneity; neither is XML that only provides structural information a solution.

In the database area, semantics can be regarded as people's interpretation of data and schema items according to their understanding of the world in a certain context. In data integration, the type of semantics considered is generally real-world semantics that are concerned with the "mapping of objects in the model or computational world onto the real world [...] [and] the issues that involve human interpretation, or meaning and use of data and information" [27]. In this setting, semantic integration is the task of grouping,

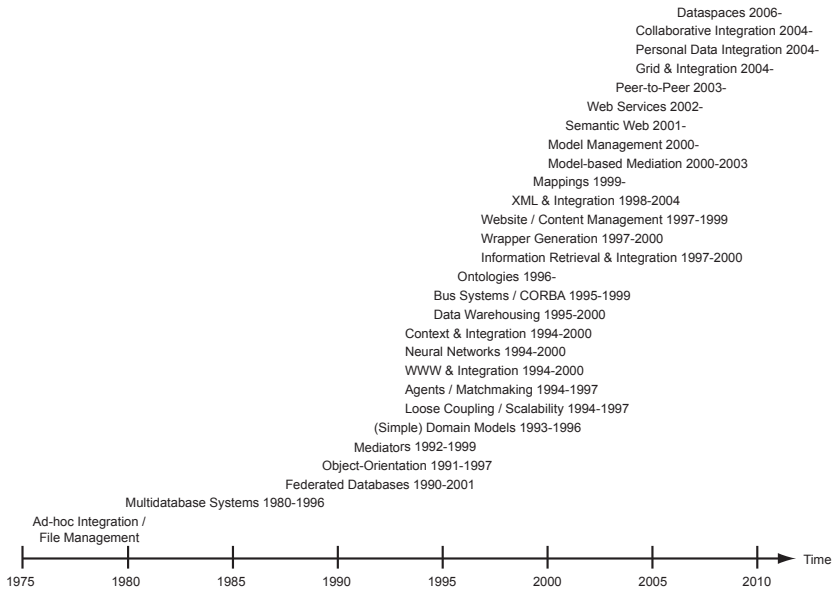

Fig. 1.2. Data Integration Research Trends over Time

combining or completing data from different sources by taking into account explicit and precise data semantics in order to avoid that semantically incompatible data is structurally merged. That is, semantic integration has to ensure that only data related to the same or sufficiently[3] similar real-world entity or concept is merged. A prerequisite for this is to resolve semantic ambiguity concerning integratable data by explicit metadata to elicit all relevant implicit assumptions and underlying context information.

One idea to overcome semantic heterogeneity in the database area is to exhaustively specify the intended real-world semantics of all data and schema elements. Unfortunately, it is impossible to completely define what a data or schema element denotes or means in the database world [29]. Therefore, database schemas do typically not provide enough explicit semantics to interpret data always consistently and unambiguously [30]. These problems are further worsened by the fact that semantics may be embodied in data models, conceptual schemas, application programs, the data itself, and the minds of users. Moreover, there are no absolute semantics that are valid for all potential users; semantics are relative [15]. These difficulties concerning semantics are the reason for many still open research challenges in the area of data integration.

[3] How much similarity is considered as sufficient depends on the particular information need and application area.

Ontologies — which can be defined as explicit, formal descriptions of concepts and their relationships that exist in a certain universe of discourse, together with a shared vocabulary to refer to these concepts — can contribute to solve the problem of semantic heterogeneity. Compared with other classification schemes, such as taxonomies, thesauri, or keywords, ontologies allow more complete and more precise domain models [20]. With respect to an ontology a particular user group commits to, the semantics of data provided by data sources for integration can be made explicit. Based on this shared understanding, the danger of semantic heterogeneity can be reduced. For instance, ontologies can be applied in the area of the Semantic Web to explicitly connect information from web documents to its definition and context in machine-processable form; thereby, semantic services, such as semantic document retrieval, can be provided.

In database research, *single* domain models and ontologies were first applied to overcome semantic heterogeneity. As in SIMS [3], a domain model is used as a single ontology to which the contents of data sources are mapped. Thus, queries expressed in terms of the global ontology can be asked. In general, single-ontology approaches are useful for integration problems where all information sources to be integrated provide nearly the same view on a domain [33]. In case the domain views of the sources differ, finding a common view becomes difficult. To overcome this problem, multi-ontology approaches like OBSERVER [26] describe each data source with its own ontology; then, these local ontologies have to be mapped, either to a global ontology or between each other, to establish a common understanding. Thus, it is now state of the art that information systems "carry with them an explicit model of the world that they operate in, a model of what the data that they carry stand for." [32].

1.5 Personal Semantic Data Integration in the SIRUP Approach

Mapping all data to one single domain model or ontology forces users to adapt to one single conceptualization of the world. This contrasts to the fact that receivers of integrated data widely differ in their conceptual interpretation of and preference for data — they are generally situated in various real-world contexts and have different conceptual models of the world in mind [17]. These models do not only vary between different people in the same domain, but even for the same individual over time [14]. COIN [17] was one of the first research projects to consider the different contexts data providers and data receivers are situated in.

In our own research, we continue the trend of taking into account user-specific aspects in the process of semantic integration. We address the problem how individual mental domain models and personal semantics of concepts can be reflected in data integration to provide tailor-made integration

for personal information needs. In the SIRUP (Semantic Integration Reflecting User-specific semantic Perspectives) approach [38], we investigate how data — equipped with explicit, queryable semantics — can be effectively pre-integrated on a conceptual level. Thereby, we aim at enabling users to perform declarative data integration by conceptual modeling of their individual ways to perceive a domain of interest.

Origin of our research is the observation that different users often have diverse views of reality — i.e., they perceive and conceptualize the same real-world part differently, according to their relative points of view, their information needs, and expectations [23]. Additionally, none of these co-existing views of the real world can be regarded as being more correct than another because each view is intended for a worthy purpose [31]. In general, we refer to this phenomenon as data receiver heterogeneity. Imposing a single global schema for all users can have severe limitations that seriously interfere with the users' individual work because thereby, data receiver sovereignty is violated. Sovereignty of data receivers refers to the fact that using integrated data must be non-intrusive [28]; i.e., users should not be forced to adapt to any standard concerning structure and semantics of data they desire. Therefore, to take a "one integrated schema fits all" approach is definitely not a satisfactory solution. We generally subsume problems that cause a single global schema to be inappropriate for particular users as perspectual integration mistakes [38]. These include:

- *Data selection mistakes* are caused when data that is available through the global schema is, from the users' perspective, inappropriately collected and selected from a given data source — for example, by only including particular local relations in the global schema.
- *Source selection mistakes* occur when the decision of the global schema designer, which data sources to incorporate into the global schema, differs from individual users' preferences for data from various origins (e.g., due to quality or reliability).
- *Entity granularity mistakes* refer to the fact that the degree of granularity in which information is represented in the global schema can be too coarse-grained (general) or too fine-grained (specialized) according to the requirements of individual users — e.g., by integrating a "seminar" and a "colloquium" relation into a general global "course" relation.
- *Attribute granularity mistakes* are problems of inadequate granularity concerning attributes of entities in the global schema.
- *Data semantics mistakes* arise when the global schema provides an integrated view on data that is semantically not related according to the individual perception of specific users. For instance, data concerning lectures and seminars may be globally merged since both represent similar forms of teaching. However, this is not useful for people who are only interested in seminars because seminar information is blurred with lectures.

- Last, but not least, *data taxonomy mistakes* occur when generalization / specialization hierarchies given by the global schema do not fit the perspective of the particular domain according to individual users.

In general, all six integration mistakes presented can be independently combined to form combined perspectual integration mistakes.

To avoid perspectual integration mistakes, we advocate user-specific, personal semantic data integration. However, to be suitable for this, data integration approaches have to meet certain requirements. We summarize these requirements with the ASME criteria [36]:

- ***A**bstraction* refers to shielding users from low-level heterogeneities of underlying data sources;
- ***S**election* means the possibility of user-specific selection of data and data sources for individual integration;
- ***M**odeling* corresponds to the availability of means to incorporate user-specific perception of the domain for which integrated data is desired in the process of data integration;
- ***E**xplicit semantics* refers to means for explicitly representing the intended real-world semantics of data.

As shown in [36], current data integration approaches fail to completely meet these requirements. In response to this, we propose the SIRUP approach to personal semantic data integration to fulfill all the ASME criteria entirely.

In SIRUP, data providers declaratively link groups of attributes representing alphanumeric data for particular real-world concepts (e.g., "database lecture at University of Zurich") to so-called IConcepts (short for "Intermediate Concept"). Each IConcept represents a single, distinct concept of the real world, and for each real-world concept, there is only one single IConcept in a SIRUP integration system. To make its meaning explicit for both, humans and computers, every IConcept is connected to an ontological concept (through the SOQA ontology API [40]) that precisely represents its intended semantics. Thus, by connecting attribute data from diverse data sources to IConcepts, data from these sources is pre-integrated on a conceptual level and its intended semantics made explicit. In order to allow more than one data source to provide data concerning a particular concept of the real world and to distinguish the origin of data, all the attributes from each data source are organized as separate attribute groups in their respective IConcept. In addition, data providers annotate all attributes they provide for IConcepts so that metadata on attribute meaning, data types, key constraints, measurement units, etc. is explicitly available for users.

Based on these foundations, we provide a declarative integration and query language so that users, equipped with suitable IConcept search tools (see [39]), can derive user-specific concepts (UserConcepts) that are tailored to their information needs from the available set of IConcepts. These UserConcepts can be organized in hierarchies so that individually integrated, virtual

views (so-called Semantic Perspectives) representing user-specific conceptual domains models to precisely meet personal information needs can be built. In the whole process of UserConcept modeling and combination, all available metadata including ontology links is automatically maintained and propagated; thus, Semantic Perspectives are annotated individual schemas over diverse data sources with explicit semantics. Finally, queries against Semantic Perspectives can be formulated that are processed by the respective SIRUP integration system. If desired, resulting data can be exported in a variety of formats, such as XML documents, relational tuples (through JDBC), and Excel spreadsheets.

1.6 Outlook

Albeit there is a remarkable history of research in the field of data integration and in spite of significant progress that has been made since the mid-1990s, ranging from concepts and algorithms to systems and commercial aspects, significant challenges still remain [18]. In this section, we present some areas that exhibit such challenges for future data integration research from our own perspective.

First of all, dynamic markets and increased competition demand for higher degrees of flexibility concerning data access and interoperability in the business domain. Thus, enterprises are faced with the requirement to provide multiple co-existing integrated views on their distributed corporate data sources to flexibly support different information needs. For instance, to enable banks to precisely assess credit risks according to the Basel II standard for risk management[4], a comprehensive and sound basis of integrated customer data is necessary. While in most banks, the needed data is available, it is often scattered over distributed sources, can be inconsistent and partially available only in hard paper copies. This alone is a challenging integration task for many banks; however, it is aggravated by the fact that alternative ways to organize the integrated data can simultaneously be necessary to support distinct information needs (e.g., categorization of credit risks according to geographical criteria or based on customer types). Here, personal data integration approaches like SIRUP can contribute.

Fostering agile cross-enterprise cooperation is another area that imposes challenges for data integration. For example, for virtual organizations as sets of organizational units that work towards a common goal, on-the-fly data integration is extremely important due to their dynamic nature [35]. To effectively provide the needed information by all the cooperating partners in a timely manner, each of them being situated in a different real-world context having his own conceptual model of the world in mind, flexible and tailored data integration is a prerequisite. Based on adequately integrated data, required

[4] See `http://www.bis.org/publ/bcbsca.htm`

applications like supply chain management (SCM), enterprise resource planning (ERP), and customer relationship management (CRM) can be realized.

Another area of inter-organizational cooperation between organizational units is e-science. Here, virtual experiments based on intensive computations and huge amounts of data are performed in grid environments, as, for example, in earth sciences, particle physics, and bioinformatics. Not only is data integration in this field required to meet diverse scientific information needs, but also scalability and manageability issues rise due to the fact that masses of data need to be handled efficiently. A key factor for interdisciplinary multi-national e-science projects is the ability to precisely satisfy the data integration and sharing needs of the involved research groups from diverse disciplines. Similarly, successful work in life sciences and e-health relies on integrated access to disparate forms of data that are spread over many biological and medical institutions by taking into account local data semantics. For these areas, user- and group-specific integration approaches like SIRUP can be useful.

As one of the goals of data integration is the provision of unified access to multiple data sources, privacy and security are important issues. Thus, flexible yet effective means for access control in integrated systems are necessary [22]. Despite the fact that integration can provide many benefits, data integration and data sharing are often hampered by privacy concerns [10]. For instance, companies abstain from exchanging data because of fear to be exploited by competitors or regulatory institutions. Similarly, integrated access to patient data can advance medical research but may be impossible without proven measures for privacy protection and access control. Therefore, the development of techniques to guarantee data integration and data sharing without loss of privacy is essential.

Data quality, that can be characterized through accuracy, completeness, timeliness, and consistency of data, is of major interest for the usability of integrated data. In the realm of data integration, however, often complex data flows between data producers, data integrators, and consumers of integrated data have to be taken into account to provide appropriate data quality solutions. Fortunately, ontology-enhanced schemas, as used in semantic data integration, represent an important prerequisite for high quality integrated data and can thus ease quality related issues [16]. In particular, the possibility for users to verify where data originates from and how it was combined and converted into its current form are central in enabling users to distinguish between facts and beliefs and, in consequence, to establish trust in integrated data [16]. Therefore, data lineage and traceability issues are likely to play an important role in future integrations systems, especially when complex data transformations over widely distributed data sources are involved. In addition, globally enforcing integrity constraints can help users to trust integrated data from diverse sources [11].

In our own work in the SIRUP project, we focus on personal semantic integration of structured and annotated alphanumeric data. However, unstructured data, such as letters, reports, presentations, emails, and web pages

constitute about 80-90% of all the data in enterprises according to current estimates by analyst firms, such as Gartner. Thus, there is a big challenge to transform this into valueable integrated information that precisely serves the needs in a dynamic business world. One approach to manage this may be provided by the emerging concept of dataspaces that postulates co-existing structured and unstructured data sets without initially requiring to integrate all data. Similar loosely-coupled approaches to data integration are represented by social networks and data sharing communities who collaboratively and incrementally contribute to building an integrated set of data. Here, the vision is to provide ease of use in community data sharing so that also non-expert users can manage and share their diverse data with minimal effort [2]. However, the future needs to show to what extent these approaches can contribute to reach the grand challenge as formulated in the Asilomar report on database research [5], i.e., to make it easy for everyone to store, organize, access, and analyze the majority of human information online.

1.7 Conclusions

In this paper, we gave an overview of issues and principal approaches in the area of integration seen from a database perspective. Even though data integration is one of the older research topics in the database area, there is yet no silver bullet solution and there is none to be expected in the near future. The most difficult integration problems are caused by semantic heterogeneity; they are being addressed in current research focusing on applying explicit, formalized data semantics to provide semantics-aware integration solutions. Despite this, considerable work remains to be done for the vision of truly personal semantic integration in form of easy to use and scalable solutions to become true.

References

1. S. Abiteboul, O. Benjelloun, and T. Milo. Web Services and Data Integration. In *Third International Conference on Web Information Systems Engineering (WISE 2002)*, pages 3–7, Singapore, December 12-14, 2002. IEEE Computer Society.
2. S. Abiteboul and N. Polyzotis. The Data Ring: Community Content Sharing. In *Third Biennial Conference on Innovative Data Systems Research (CIDR 2007)*, Asilomar, CA, USA, January 7-10, 2007. Online Proceedings.
3. Y. Arens, C. Y. Chee, C.-N. Hsu, and C. A. Knoblock. Retrieving and Integrating Data from Multiple Information Sources. *International Journal of Cooperative Information Systems (IJCIS)*, 2(2):127–158, 1993.
4. R. J. Bayardo, B. Bohrer, R. S. Brice, A. Cichocki, J. Fowler, A. Helal, V. Kashyap, T. Ksiezyk, G. Martin, M. H. Nodine, M. Rashid, M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, and D. Woelk. InfoSleuth: Agent-Based

Semantic Integration of Information in Open and Dynamic Environments. In *1997 ACM SIGMOD International Conference on Management of Data (SIGMOD 1997)*, pages 195–206, Tucson, Arizona, USA, 1997. ACM.

5. P. Bernstein, M. Brodie, S. Ceri, D. DeWitt, M. Franklin, H. Garcia-Molina, J. Gray, J. Held, J. Hellerstein, H. V. Jagadish, M. Lesk, D. Maier, J. Naughton, H. Pirahesh, M. Stonebraker, and J. Ullman. The Asilomar Report on Database Research. *SIGMOD Record*, 27(4):74–80, 1998.
6. P. A. Bernstein, A. Y. Halevy, and R. A. Pottinger. A Vision for Management of Complex Models. *ACM SIGMOD Record*, 29(4):55–63, 2000.
7. O. A. Bukhres and A. K. Elmagarmid, editors. *Object-Oriented Multidatabase Systems: A Solution for Advanced Applications*. Prentice-Hall, 1996.
8. M. Carey, L. Haas, P. Schwarz, M. Arya, W. Cody, R. Fagin, M. Flickner, A. Luniewski, W. Niblack, D. Petkovic, J. Thomas, J. Williams, and E. Wimmers. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach. In *5th International Workshop on Research Issues in Data Engineering-Distributed Object Management (RIDE-DOM 1995)*, pages 124–131, Taipei, Taiwan, March 6-7, 1995.
9. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *16th Meeting of the Information Processing Society of Japan (IPSJ)*, pages 7–18, Tokyo, Japan, October, 1994.
10. C. Clifton, M. Kantarcioglu, A. Doan, G. Schadow, J. Vaidya, A. K. Elmagarmid, and D. Suciu. Privacy-Preserving Data Integration and Sharing. In *9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2004)*, pages 19–26, Paris, France, June 13, 2004. ACM.
11. S. Conrad, M. Höding, G. Saake, I. Schmitt, and C. Türker. Schema Integration with Integrity Constraints. In *15th British National Conference on Databases (BNCOD 1997)*, pages 200–214, London, UK, July 7-9, 1997. Springer.
12. K. R. Dittrich and D. Jonscher. All Together Now — Towards Integrating the World's Information Systems. In *Advances in Multimedia and Databases for the New Century*, pages 109–123, Kyoto, Japan, November 30 - December 2, 1999. World Scientific Press.
13. M. J. Franklin, A. Y. Halevy, and D. Maier. From Databases to Dataspaces: A New Abstraction for Information Management. *SIGMOD Record*, 34(4):27–33, 2005.
14. B. R. Gaines and M. L. G. Shaw. Comparing the Conceptual Systems of Experts. In *11th International Joint Conference on Artificial Intelligence (IJCAI 1989)*, pages 633–638, Detroit, Michigan, USA, August, 1989. Morgan Kaufmann.
15. M. García-Solaco, F. Saltor, and M. Castellanos. Semantic Heterogeneity in Multidatabase Systems. In O. A. Bukhres and A. K. Elmagarmid, editors, *Object-Oriented Multidatabase Systems. A Solution for Advanced Applications*, pages 129–202. Prentice-Hall, 1996.
16. M. Gertz, M. T. Özsu, G. Saake, and K.-U. Sattler. Report on the Dagstuhl Seminar "Data Quality on the Web". *SIGMOD Record*, 33(1):127–132, 2004.
17. C. H. Goh, S. E. Madnick, and M. Siegel. Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment. In *Third International Conference on Information and Knowledge Management (CIKM 1994)*, pages 337–346, Gaithersburg, USA, November 29 - December 2, 1994. ACM.

18. A. Y. Halevy. Data Integration: A Status Report. In *Datenbanksysteme in Business, Technologie und Web (BTW 2003)*, volume 26, pages 24–29, Leipzig, Germany, February 26-28, 2003. Gesellschaft für Informatik (GI).
19. A. Y. Halevy, Z. G. Ives, D. Suciu, and I. Tatarinov. Schema Mediation in Peer Data Management Systems. In *19th International Conference on Data Engineering (ICDE 2003)*, pages 505–518, Bangalore, India, March 5-8, 2003. IEEE Computer Society.
20. M. N. Huhns and M. P. Singh. Agents on the Web: Ontologies for Agents. *IEEE Internet Computing*, 1(6):81–83, 1997.
21. A. R. Hurson and M. W. Bright. Multidatabase Systems: An Advanced Concept in Handling Distributed Data. *Advances in Computers*, 32:149–200, 1991.
22. D. Jonscher and K. R. Dittrich. An Approach for Building Secure Database Federations. In *20th International Conference on Very Large Data Bases (VLDB 1994)*, pages 24–35, Santiago de Chile, Chile, September 12-15, 1994. Morgan Kaufmann.
23. W. Kent. *Data and Reality. Basic Assumptions in Data Processing Reconsidered.* North-Holland, Amsterdam, 1978.
24. T. Landers and R. L. Rosenberg. An Overview of MULTIBASE. In *Second International Symposium on Distributed Data Bases (DDB 1982)*, pages 153–184, Berlin, Germany, September 1-3, 1982. North-Holland.
25. R. McCann, A. Doan, V. Varadaran, A. Kramnik, and C. Zhai. Building Data Integration Systems: A Mass Collaboration Approach. In *Sixth International Workshop on Web and Databases (WebDB 2003)*, pages 25–30, San Diego, California, USA, June 12-13, 2003.
26. E. Mena, V. Kashyap, A. P. Sheth, and A. Illarramendi. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In *First IFCIS International Conference on Cooperative Information Systems (CoopIS 1996)*, pages 14–25, Brussels, Belgium, June 19-21, 1996. IEEE Computer Society.
27. A. M. Ouksel and A. P. Sheth. Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area and the Special Section. *SIGMOD Record*, 28(1):5–12, 1999.
28. P. Scheuermann, A. K. Elmagarmid, H. Garcia-Molina, F. Manola, D. McLeod, A. Rosenthal, and M. Templeton. Report on the Workshop on Heterogenous Database Systems held at Northwestern University, Evanston, Illinois, December 11-13, 1989. *SIGMOD Record*, 19(4):23–31, 1990.
29. A. P. Sheth, S. K. Gala, and S. B. Navathe. On Automatic Reasoning for Schema Integration. *International Journal of Intelligent and Cooperative Information Systems*, 2(1):23–50, 1993.
30. A. P. Sheth and J. A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22(3):183–236, 1990.
31. A. Sølvberg. Data and What They Refer to. In *Conceptual Modeling, Current Issues and Future Directions, Selected Papers from the Symposium on Conceptual Modeling, Los Angeles, California, USA, held before ER 1997*, pages 211–226. Springer, 1997.
32. A. Sølvberg. Conceptual Modeling in a World of Models. In R. Kaschek, editor, *Entwicklungsmethoden für Informationssysteme und deren Anwendung, EMISA 1999*, pages 63–77, Fischbachau, Germany, 1999. Teubner.

33. H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI-2001 Workshop on Ontologies and Information Sharing*, pages 108–117, Seattle, USA, April 4-5, 2001.
34. G. Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3):38–49, 1992.
35. M. Winslett. Databases in Virtual Organizations: A Collective Interview and Call for Researchers. *SIGMOD Record*, 34(1):86–89, 2005.
36. P. Ziegler. User-Specific Semantic Integration of Heterogeneous Data: What Remains to be Done? Technical Report ifi-2004.01, Department of Informatics, University of Zurich. http://www.ifi.unizh.ch/techreports/TR_2004.html, 2004.
37. P. Ziegler and K. R. Dittrich. Three Decades of Data Integration - All Problems Solved? In *18th IFIP World Computer Congress (WCC 2004), Volume 12, Building the Information Society*, pages 3–12, Toulouse, France, August 22-27, 2004. Kluwer.
38. P. Ziegler and K. R. Dittrich. User-Specific Semantic Integration of Heterogeneous Data: The SIRUP Approach. In *First International IFIP Conference on Semantics of a Networked World (ICSNW 2004)*, pages 44–64, Paris, France, June 17-19, 2004. Springer.
39. P. Ziegler, C. Kiefer, C. Sturm, K. R. Dittrich, and A. Bernstein. Detecting Similarities in Ontologies with the SOQA-SimPack Toolkit. In *10th International Conference on Extending Database Technology (EDBT 2006)*, pages 59–76, Munich, Germany, March 26-31, 2006. Springer.
40. P. Ziegler, C. Sturm, and K. R. Dittrich. Unified Querying of Ontology Languages with the SIRUP Ontology Query API. In *Datenbanksysteme in Business, Technologie und Web (BTW 2005)*, pages 325–344, Karlsruhe, Germany, March 2-4, 2005. Gesellschaft für Informatik (GI).