From Data Integration to Big Data Integration

Sonia Bergamaschi[1], Domenico Beneventano[1], Federica Mandreoli[3], Riccardo Martoglia[3], Francesco Guerra[1], Mirko Orsini[2], Laura Po[1], Maurizio Vincini[1], Giovanni Simonini[1], Song Zhu[4], Luca Gagliardelli[4], Luca Magnotta[2,4],

[1] Dipartimento di Ingegneria "Enzo Ferrari"
Università di Modena e Reggio Emilia - name.surname@unimore.it
[2] Datariver s.r.l. - name.surname@datariver.it
[3] FIM - Università di Modena e Reggio - name.surname@unimore.it
[4] ICT School - Università di Modena e Reggio - name.surname@unimore.it

Abstract. The Database Group (DBGroup, www.dbgroup.unimore.it) and Information System Group (ISGroup, www.isgroup.unimore.it) research activities have been mainly devoted to the Data Integration Reserach Area. The DBGroup designed and developed the MOMIS data integration system, giving raise to a successful innovative enterprise DataRiver (www.datariver.it), distributing MOMIS as open source. MOMIS provides an integrated access to structured and semistructured data sources and allows a user to pose a single query and to receive a single unified answer. Description Logics, Automatic Annotation of schemata plus clustering techniques constitute the theoretical framework. In the context of data integration, the ISGroup addressed problems related to the management and querying of heterogeneous data sources in large-scale and dynamic scenarios. The reference architectures are the Peer Data Management Systems and its evolutions toward dataspaces. In these contexts, the ISGroup proposed and evaluated effective and efficient mechanisms for network creation with limited information loss and solutions for mapping management query reformulation and processing and query routing. The main issues of data integration have been faced: automatic annotation, mapping discovery, global query processing, provenance, multidimensional Information integration, keyword search, within European and national projects. With the incoming new requirements of integrating open linked data, textual and multimedia data in a big data scenario, the research has been devoted to the Big Data Integration Research Area. In particular, the most relevant achieved research results are: a scalable entity resolution method, a scalable join operator and a tool, LODEX, for automatically extracting metadata from Linked Open Data (LOD) resources and for visual querying formulation on LOD resources. Moreover, in collaboration with DATARIVER, Data Integration was successfully applied to smart e-health,

1 Data Integration at the DBGroup: State of the Art

Data Integration is the problem of combining data residing at different autonomous sources, and providing the user with a unified view of these data.

The DBGroup group has been investigating data integration for more than 20 years and almost all of the research activities have been centered around the MOMIS (Mediator EnvirOnment for Multiple Information Sources) Data Integration System; [5,20], a most recent description is in [18]. An open source version of the MOMIS system is delivered and maintained by Datariver (see section 4).

The MOMIS system is characterized by a classical wrapper/mediator architecture [60], where the local data sources contain the real data, while a *Global Virtual Schema* (*GVS*) provides a reconciled, integrated, and virtual view of the underlying data sources. MOMIS follows a *Global-As-View* approach: each class of the *GVS* is characterized in terms of a view over the sources; then a query over the *GVS* is rewritten on the data sources by *query unfolding* [36]. In the following we describe the main features and applications of the MOMIS system. For a complete version of this section see `http://dbgroup.ing.unimore.it/Momis`.

1.1 Automatic Global Schema Generation and Annotation

One of the main features of the MOMIS System is the rapid deployment of data integration projects, by means of a process that detects semantic similarities among the involved local source schemata, automatically generates a *GVS* and the mappings among the *GVS* and the local schemata. The theoretical framework of this data integration process is constituted by Description Logics, Automatic Annotation of schemata plus clustering techniques. In particular, with the Annotation step, schema labels (i.e. class/attribute names) are mapped to concepts of a lexical ontology, such as Wordnet [51], in order to perform *schema matching*, i.e., to discover relationships among the elements of different schemata [28]. However, performance of automatic annotation methods on real world schemata suffers from the abundance of non-dictionary words such as compound nouns, abbreviations and acronyms. In [58] we addressed this problem by introducing a *schema label normalization* method which expands abbreviations/acronyms and annotates compound nouns; the techniques was implemented into the NORMS tool, described in the next section.

1.2 The NORMS (NORMalizer of Schemata) tool

NORMS is a tool that provides automatic normalization and annotation for schema labels [57]; it was developed in collaboration with Datariver[5] and it can be integrated in the MOMIS framework to improve schema matching. To handle automatically a large number of non-dictionary words, NORMS uses, in addition to lexical ontologies, other resources: (1) complementary schemata (other schemata that have to be integrated with the current schema); (2) online abbreviation dictionary; (3) user-defined dictionaries. The main innovative aspect of NORMS is the *Compound Nouns (CN) Interpretation*: i.e., the task of determining the semantic relationships holding among its constituents. NORMS

[5] NORMS will be included in the next release of the MOMIS Open Source version, available at `http://www.datariver.it/data-integration/momis/`.

annotates each CN constituent and then performs automatic CN interpretation by using the set of nine semantic relationships (CAUSE, HAVE, MAKE etc.) defined by Levi in [38]. The performance of NORMS was tested in several data integration projects of different semantic domains; moreover, NORMS was also tested w.r.t. schema and data integration test suites, such as Almalgam [52] and TPC-H[6]. The experimental results have shown the effectiveness of NORMS, which significantly improves annotation, and, consequently, schema matching.

1.3 Data Fusion

Data fusion is the process of fusing multiple records representing the same real-world object into a single, consistent, and clean representation; merging different records into a single representation and at the same time resolving existing data conflicts was a problem addressed by several researchers in the past decades, but with few implemented solutions in integration systems [29]. To perform *Data Fusion*, we assumed that *Entity Resolution*, i.e., the identification of the same object in different sources, has been already performed and thus a shared object identifier (ID) exist among different sources. Multiple records with the same ID are fused by means of the Full Join Merge operator [18]. On 2016, we faced and solved the problem of Entity Resolution in a Big Data Integration scenario (see section 5.1). Moreover, as the increasing number of distributed data sources to be integrated introduces join scalability as a critical issue for the scalability of a Data Integration system, we devised a new scalable join operator (see section 5.2).

1.4 Data Provenance

Data Integration Systems deal with information coming from different sources, potentially uncertain or even inconsistent with each other, then the problem of "How to incorporate the notion of data quality (source reliability, accuracy, etc.)" is crucial. A common approach for evaluating the *quality* and *trustworthiness* of the data in systems involving a large number of sources is the analysis of the *provenance of information*. *Provenance* describes where data come from, how it is derived and how it was modified over time.

The notion of *PI-CS* provenance defined in [31] encodes all the *possible different derivations* of a output tuple in the query result, by storing a set of input tuples *for each derivation*. In [12] we extended the definition of *PI-CS* provenance to include *resolution functions* and, then, to distinguish between derivations where the contributing tuples have conflicts or not. We proved the usefulness of data provenance in a Data Integration context in several studies, from the use of data provenance in conflict resolution strategies [4] to the development of provenance-aware semantic search engines based on data integration[6].

1.5 Integration of Data and Multimedia Sources

The proliferation of multimedia data, and the consequent need of their integration with traditional information, represents nowadays a critical issue. In [13] we

[6] `http://www.tpc.org/tpch`

extended the MOMIS System to integrate “traditional” and “multimedia” data sources: multimedia queries can be expressed on the GVV without requiring multimedia processing capabilities at the *mediator* level since they are managed at the *local* level by the multimedia system. However, this solution requires a completely new query processing method. with respect to the one used for traditional data (based on a full join operation). Then, in [1] we discussed the question: “How can a multimedia local source supporting ranking queries be integrated into a mediator system without such capabilities?”. We first described a naïve approach to support ranking (*Top-K*) queries which keep substantially unchanged the query processing method; this approach does not guarantee the completeness of results for *Top-K* queries, i.e., less than K results might be returned. Then, we discussed two alternative solutions for extending a mediator system so as to support multimedia queries.

1.6 The MOMIS System: Applications and Extensions

The MOMIS system was used to integrate genotypic and phenotypic data, for the development of the CEREALAB database [50]. A specific user’s requirement coming from this application domain was data provenance capabilities, as discussed in [7]; moreover, in [11] we shown the publication of CEREALAB into the Linked Open Data cloud; finally, in [10] we applied to CEREALAB a fully automatic and semantic method for searching bibliographic data. We applied the MOMIS system either for building a tourism Information provider [19] and to define semantic mappings amongst product classification schemas [14,16].

In [9] we extended the MOMIS system with a multi-agent architecture based on the P2P model; while a single peer carries out data integration activities, it exchanges knowledge with other peers by means of specialized agents. In [59] we shown as the tests of the THALIA benchmark [35] are fulfilled by the MOMIS system; In [15] we presented a framework able to query an integrated view of data and to search for eServices related to retrieved data.

1.7 Integrating Multidimensional Information

Collaborative business making is emerging as a possible solution for the difficulties that Small and Medium Enterprises (SMEs) are having in the current difficult economic scenarios. Collaboration, as opposed to competition, provides a competitive advantage to companies and organizations that operate in a joint business structure. When dealing with multiple organizations, managers must access unified strategic information obtained from the knowledge repositories of each individual organization; unfortunately, traditional Business Intelligence (BI) tools are not designed with the aim of collaboration so the task becomes difficult from a managerial, organizational and technological point of view. To deal with this shortcoming, we provided an integration, mapping-based, methodology for heterogeneous Data Warehouses that aims at facilitating business stakeholders’ access to unified strategic information obtained from a network of heterogeneous collaborating SMEs [17]. In particular, we proposed a mapping-based integration methodology that is able to generate semantic mappings between dimensions of different DWs, either between dimension categories and between members of such categories. Our work is motivated by data-quality which is a crucial aspect when building an inter-organization DW; in fact, given the use of the DW, incorrect or low-quality data may not only make the DW useless, but its use may also lead to wrong decisions with potential negative impact on the organization. For these reasons, we analyzed data-quality requirements for the generated mappings, such as coherency and consistency, to ensure that the integrated information can be correctly aggregated (or disaggregated). This observation is relevant as the multidimensional data is usually explored along aggregation patterns, drilling-down or rolling-up from a starting analysis point.

1.8 MOMIS Dashboard

The MOMIS Dashboard[7] is an interactive visualization tool (available also as a mobile application) either for integrated data obtained with MOMIS or for data of several commercial DBMSs. It makes easier to compare data and capture useful information. It allows to filter the data and visualize the results through different charts. The available charts are : line, bar, pie, bubble on a Google Maps, tabular views; filters applicable on the charts are: date interval, numeric interval, set of different values, free text. The MOMIS Dashboard allows to define advanced access rules on the data, i.e different views. In this way, different users can access to the same charts but filtering only their proprietary data set. Another interesting feature is the multilingual support, i.e., it can manage different languages for the same Dashboard. The MOMIS Dashboard is developed in collaboration with Datariver that has already used it in several projects (see (http://www.datariver.it/data-integration/)), such as "Open Data Lavoro", a web application for analysis and monitoring of public open data related to job market, and the "Italian FSHD Registry", which will be described in section 4.1.

2 Data Integration at the ISGroup: State of the Art

Peer data management systems (PDMSs) represent a natural step beyond data integration systems, replacing their single logical schema with an interlinked collection of semantic mappings between peers' individual schemas [34]. Moreover, the synergy between PDMSs and Semantic Web technologies has paved the way for large-scale sharing of semantically rich data, like RDF or ontologies. Because of the lack of common understanding of the vocabulary used by peers, the resulting heterogeneity of data representations opens new challenges as to the efficient and effective retrieval of relevant information.

As opposed to viewing semantic misalignment as a limit for interoperability, the research of ISGROUP leveraged on the presence of semantic approximations between the peers' schemas as a means for giving effective hints along the following directions: 1) effective query processing [47,42]; 2) network creation [49].

[7] For a complete description see http://dbgroup.ing.unimore.it/MomisDashboard.

A PDMS underlies a potentially very large network able to handle huge amounts of data. For this reason, query routing, i.e. the process of selecting the most promising peers, is a fundamental issue for providing relevant answers to queries over distributed resources. In this context, we proposed a distributed index mechanism where each peer is provided with a novel kind of index, the Semantic Routing Index (SRI), for routing queries effectively. SRIs are also employed in the query answering phase for reducing the space of reformulations and for ranking answers. The proposed approach for query processing founds on a fuzzy settlement which provides a formal semantics of the approximations originated by the heterogeneity of the schemas in a PDMS.

When processing queries in a PDMS, the semantic path needs to relate the terms used in the query with the terms specified by the node providing the data. Hence it is likely that there will be information loss along long paths in the PDMS because of missing (or incomplete) mappings, leading to the well-known problem of how to boost a network of mappings in a PDMS. Our research investigated this issue and proposed the first efficient approach for the flexible creation and maintenance of the network structure, clustering together semantically related peers. The approach we propose is scalable, incremental, and self-adaptive in the creation and maintenance of clusters.

When it comes to network construction, every time a new peer joins the network, there is the need to make explicit the semantics of its schema by associating each schema's term with the right concept. In this parallel research branch, ISGROUP also proposed effective knowledge-based disambiguation techniques working on a variety of structures both at data and metadata level [39,46,45].

All the above techniques have been incorporated in a complete PDMS infrastructure, the SUNRISE (System for Unified Network Routing, Indexing and Semantic Exploration) system, that offers network construction and exploration facilities for XML and RDF data sharing [41,43]. Then, in the context of the FIRB NeP4B project and in collaboration with ISTI-CNR, SUNRISE was extended to multimedia data that needs the support of content-based, aka similarity, predicates in the query language [37].

Another research area where the main concepts of PDMSs where exploited profitably is inter-business collaborative where companies coordinate themselves to develop common and shared opportunities. In collaboration with the Business Intelligence Group of the University of Bologna we introduced a peer-to-peer data warehousing architecture, Business Intelligence Network (BIN) [32,48]. The original contributions we gave in this area are a language for the definition of semantic mappings between the schemata of peers, using predicates that are specifically tailored for the multidimensional model, a framework for OLAP query reformulation that relies on the translation of mappings and queries towards the underlying relational schemata, a query reformulation algorithm.

In the context of data integration systems, the dataspace model represent a further step towards the integration of loosely-connected information that support data co-existence and pay-as-you-go principles [33]. Such challenge arises in application scenarios that have emerged recently such as exploratory data analysis and personal information management. A very recent line of research is conducted in collaboration with George Fletcher from Tu/E and aims at investigating theoretical and engineering solutions for query-driven mapping discovery. [30]. The project proposes a shift of perspective in mapping management from the state-of-the-art data-centric approach to a user-centric approach where mappings contribute to users satisfaction in their information seeking activities.

Graph-based data models have recently gained much popularity as powerful means for data representation, especially in the dataspace scenario. The largeness and the heterogeneity of most graph-modeled datasets in several database application areas make flexible query answering capabilities an essential need. ISGROUP proposed an approach [40,44] for the approximate matching of complex queries on such kind of data, whose query answering model gracefully blends label approximation with structural relaxation, under the primary objective of delivering meaningfully approximated results only.

3 Keyword search over relational databases

Despite the effort that the research community has put in the field in the last fifteen years and the significant number of scientific publications and prototypes developed, keyword search over relational databases is still a hot and open challenge. No research prototypes have transitioned from proof-of-concept implementations into deployed systems, and the current research proposals fail in making users able to effectively and efficiently query relational databases. The problem is intrinsically complex and multifaceted, since the information in a database is spread over a number of tables connected with multiple paths. The simple exploitation of full-text search functionalities natively implemented in the DBMS (e.g., the`match-against` function in MySQL) enables users only to discover the attributes of the database containing the query keywords at run-time, thus providing just partial, and incomplete results to be joined to form complete answers. Nevertheless, the existence of a path joining two generic tuples is not assured even in a full connected database, and it can involve tuples in other tables. Moreover, in case of multiple connections, each path carries out different semantics which cannot match with the intended meaning of the user's query. The computation of all paths largely affect the time required for solving the query and the quality of the answer.

The DBGroup is an active member in this area since 2009 and developed techniques and tools in conjunction with other research groups (in particular with Prof. Velegrakis, University of Trento – IT, Prof. Ferro, University of Padua – IT, Prof. Trillo, University of Zaragoza – ES), and under the big umbrella of the KEYSTONE IC1302 COST Action (http://www.keystone-cost.eu). The main contributions are related to: (i) the introduction of a principled model for the representing the keyword search problem over structured databases; (ii) the development of techniques and prototypes based on heuristic rules and machine learning techniques; (iii) an analysis of the methodology adopted for evaluating keyword based systems and the definition of a roadmap and a reference architecture for an evaluation framework.

Representing keyword search over structured databases. In our research, we have developed a three-step principle model to represent the process of solving keyword queries. The fundamental steps we consider are first to match the keywords into the database structures, then to discover ways these matched structures can be combined, and finally to select the best matches and combinations such that the identified database structures represent what the user had in mind to discover when formulating the keyword query. The first step is focused on trying to capture the meaning of the keywords in the query as they are understood by the user. In some sense, it provides the *user perspective* of the keyword query and it does so by providing a mapping of the keywords into database terms. We call the mappings as *configurations* and the step is referred to as the *forward analysis step* since it starts from the keywords and moves towards the database. The second step tries to capture the meaning of the keywords as they can be understood from the point of view of the data engineers who designed that database structure. In this step we compute the paths connecting the keywords, thus forming the possible solutions (we refer them as the *interpretations*) to the user query. In some sense, this phase provides the *database perspective* of the keyword query and it does so by providing the relationships among the images of the keywords. This task is referred to as the *backward analysis step* since it starts from the database structures and moves towards the query keywords through their images. The third step provides a ranking of the interpretations produced by the second step. These are the *explanations*, i.e., "fully justified" answers to the keyword queries.

Techniques and prototypes to solve keyword queries. We have developed 3 prototypes implementing the heuristics and machine learning based techniques developed. The **KEYMANTIC** [22,21] system was focused on finding configurations. It provided a solution based on a bipartite graph matching model where user keywords were matched to database schema elements by using an extension of the Hungarian algorithm. In particular, weights qualifying the matches run-time change on the basis of the partial assignments computed by the algorithm. **KEYRY** [27,53] extended KEYMANTIC by providing a probabilistic framework, based on a Hidden Markov Model (HMM), to compute the configurations. In our approach, the HMM states are database elements, and the observations are the keywords that can be associated to the states. The application of the Viterbi algorithm allows the computation of sequences of states to be associated to the keyword queries. **QUEST** [26,25] provides a complete keyword search system for relational database, where the configurations, computed by means of a HMM approach, and the interpretations, computed by means of Steiner Trees, are joined by means of the probabilistic framework provided by the Dempster Shafer theory.

Evaluating keyword search systems. We noticed two main issues that are hampering the design and development of next generation systems for keyword search over structured data: (i) the lack of systemic approaches considering all of the issues of keyword search from the interpretation of the user needs, to the computation, retrieval, ranking and presentation of the results; and (ii) the absence of a shared and complete evaluation methodology measuring user satisfaction, achieved utility and required effort for carrying out informative tasks. In light of this, we proposed in [24] a reference architecture including all the components needed for a keyword search engine for relational databases and we have discussed a roadmap for the definition of a fair and solid evaluation framework.

4 Smart Health Care at DataRiver

DataRiver (`www.datariver.it`) is an Innovative SME, accredited as a Research Innovation Institution of the Emilia Romagna Region, founded on June 2009 as a Spin-Off of the University of Modena and Reggio Emilia from the initiative of professors and researchers of the DBGroup. The company develops innovative software solutions in the healthcare industry and offers specialized consulting services for Clinical Data Management, Big Data Integration (*BDI*) and Analytics concerned to Clinical Trials, Cancer and Rare Diseases Registries, Mobile Data Capture Apps for the collection and management of patient clinical data from mobile and wearable devices. The healthcare industry is naturally rich with data – clinical, patient, claim, hospital system, financial, pharmacy and, most recently, data from wearable technology. In the healthcare context, data flows through heterogeneous systems and are stored in disparate data sources. BDI techniques provide a complete and synthetic view of all the health information about a patient or a set of selected patients. The Big Data analytics applied to integrated patient's health data allow a widespread monitoring to prevent clinical events and plan personalized care treatments. The Internet of Medical Things (IoMT) is the collection of medical devices and applications that connect to healthcare IT systems through online computer networks, including comprehensive solutions that follow patients on their life places, enabling home care and telemedicine. DBGroup and DataRiver are involved in several Smart Health Care projects for the research and development of innovative solutions, addressing the issues for BDI in the healthcare context and providing state-of-the-art applications for the IoMT.

4.1 The Italian FSHD Registry: an enhanced data integration and analytics framework for Smart Health Care

Facioscapulohumeral muscular dystrophy (FSHD) is a common myopathy, that has been associated to the reduction of a string of DNA elements, named D4Z4. BDI and analytics techniques have been applied by DataRiver and DBGroup in the "Italian FSHD Registry“ project, to discover new research results by the unified analysis of patient's clinical and genomics data. The Italian FSHD Registry study (`www.fshd.it`) involves 13 medical research centers of the FSHD network to collect clinical and genomics data of more than 6.000 patients and 2.400 families; the objective is to develop an enhanced data integration and analytic framework to predict the risk of developing a severe form of the disease, and to identify factors that can help to prevent the disease worsening. In the "Italian FSHD Registry“ project we successfully used the MOMIS system and the MOMIS Dashboard (see section 1) jointly with OpenClinica (`www.openclinica.org`), an electronic data capture software for clinical research used to capture cleaner data, ensure compliance and promote patient engagement. The OpenClinica platform have been extended by DataRiver to collect family data and to be fully integrated with the MOMIS system. The MOMIS system provides a unified view of patient's data and biological data coming from different Biobank databases, to run queries both on clinical and genomics parameters and discover new information. The MOMIS Dashboard was used to easily monitor, query, visualize and extract statistical analysis of clinical and laboratory data.

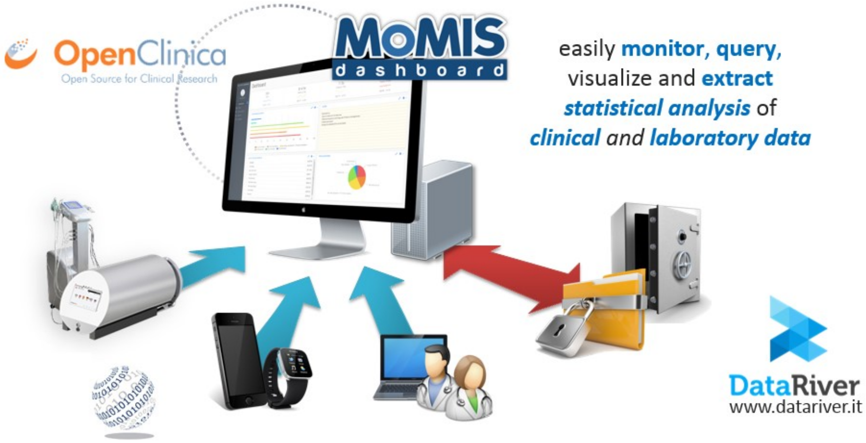

Fig. 1. Enhanced data integration and analytics for Smart Health Care

4.2 My Smart Age with HIV: a Mobile and IoMT platform for remote monitoring and empowerment of HIV patients

In the My Smart Age with HIV (MySAwH) clinical trial (`www.mysmartage.org`), an innovative IoMT mobile App has been developed to empower elderly HIV patients via health promotion, assessing reduction in health deficit and improvement in quality of life. DataRiver designed and developed the IoMT framework architecture and MySAwH App to collect patient's data from smartphone and wearable devices, elaborating and analyzing health indicators for the project; the IoMT framework has been designed to expand the traditional healthcare infrastructure and to provide patient monitoring and support outside the hospitals. The exploitation of IoMT, mobile App and wearable technologies, the integration and analysis of all collected patient's data in real time provide the physician with a continuous patient monitoring to measure the response to illness and the life quality improvement. The patient obtains an up to date insight of health condition and a constant support via the direct communication with caregiver.

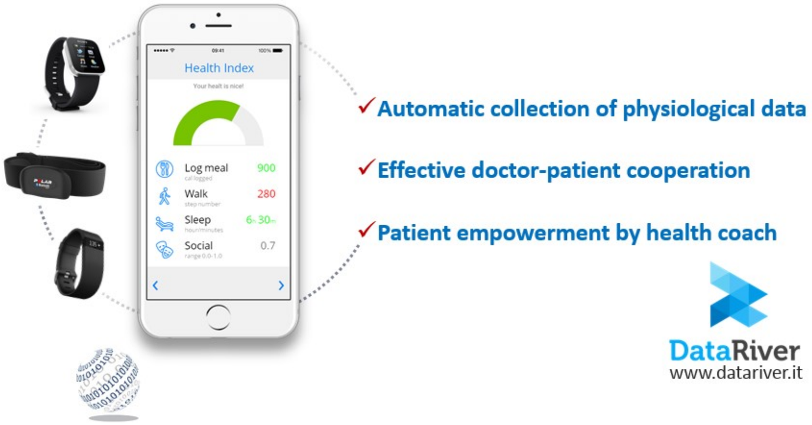

Fig. 2. Mobile, IoMT platform for remote monitoring and patient empowerment

5 Big Data Integration at the DBGroup

A huge amount of (semi-)structured data is available on the Web in the form of web tables, marked-up contents, and Linked Open Data. For enterprises, government agencies, and researcher of large scientific projects, this data can be even more valuable if integrated with their proprietary data. With the incoming of these new requirements, the research at the DBGroup has been devoted to the Big Data Integration Research Area. The main topics we are investigating are: scalable entity resolution methods, distributed scalable join methods and summarization and visual querying methods of LOD resources.

5.1 Entity Resolution

As introduced in section 1.3, *Entity Resolution* (ER) is the well-known problem to identify records that refer to the same entity. Generally, to perform ER and vocabulary-based topic detection, traditional techniques are based on schema-alignment among data sources (i.e., deriving a unique homogenous common schema from several heterogeneous ones). Unfortunately, the (semi-) structured data of the Web is usually characterized by high heterogeneity, volume and noise (missing/inconsistent data), making schema-alignment techniques no longer applicable. Therefore, Data Integration techniques dealing with this type of data typically renounce to exploit data source schemata.

At the DBgroup we developed a set of novel techniques [55,54,23] to induce loose schema information directly from the data, without exploiting the semantic of the schemas, able to scale to the huge data of the Web. This lose schema information can be employed as a surrogate of the schema-alignment and employed to enhance ER and vocabulary-based topic detection.

For ER, we proposed `Blast` [55] (Blocking with Loosely-Aware Schema Techniques), an approach to reduce the ER complexity with indexing techniques aiming to group similar records in blocks, and limit the comparison to only those records appearing in the same block. For the topic detection, we proposed Whatsit [23] a novel approach that generates signatures of sources that are matched against the signatures of a reference vocabulary. Thus, a description of the topics of the source in terms of this reference vocabulary is generated.

Finally, we developed an open source software[8] for both the approaches and we experimentally evaluated them on real world datasets. The results demonstrate that `Blast` outperforms the state-of-the-art blocking approaches for the big data scenario, and that Whatsit can actually be employed to detect topics of a given data source. Currently, we are developing ER methods for massively parallel and distributed systems (Apache Spark and Apache Flink), and we are performing on the same systems extensive benchmarks to assess their scalability for different scenarios related to big data integration problems.

5.2 Distributed Scalable Join

As we discussed in section 1.3, join scalability is a critical issue for the scalability of a Data Integration system and, then, a key requirement in a Big Data Integration scenario. To fulfill this goal we designed a scalable parallel join engine. In the distributed systems context, the main join paradigms/algorithms are [29]: Map Reduce Joins; Online Joins; Stream Joins.

Map Reduce is one of most popular paradigms for parallel computation. However, the Map Reduce paradigm uses a barrier between the Map and Reduce stages, this hurts performance. It means that the Map step has to be completed before the next step (Reduce) starts. Consequently, data coming by a big data source must be loaded in the system before the join operation can start. In addition, in case of multiple joins, the previous join step have to be completed before the next join can start. Otherwise, the Stream Join merges data in the stream paradigm, and the result of the Join can be used before it is completed. This join paradigm is useful to merge sensor data, where the input data is continuous. However, stream Joins are usually windows based to support infinite streams of data; this requires that input data sources be sorted and the sort is an expensive operation in a Big Data Integration scenario.

For these reasons, we are studying an Online Join algorithm based on a non-blocking join method. Our join engine receives stream inputs from data sources and merges tuples on the fly and, when the result of a tuple is complete, it can be used immediately in the next operation. Our join algorithm is a multi-way equi-join and it works on a cluster of computation nodes. The basis of the algorithm is composed by two steps: the first step is a the distribution step,

[8] http://stravanni.github.io/blast/

which distributes tuples on the nodes. We apply a hash function on the join attributes, and the result of the function determines to which computation node a tuple has to be sent. The second step is the local join step, where tuples are merged on the basis of the join condition. Each computation node performs the join locally: if the output tuple is complete, i.e. it have at least one tuple from all data sources, the node emits the result.

5.3 Summarization and visual querying methods for LOD resources

With more than one thousand of LOD sources available on the Web, we are assisting to an emerging trend in publication and consumption of LOD datasets. However, the pervasive use of external resources together with a deficiency in the definition of the internal structure of a dataset causes many LOD sources are extremely complex to understand. LOD tools lack in producing an high level representation of datasets and in supporting users in the exploration and querying of a source. To overcome the above problems we defined a method to unveil the implicit structure of a LOD dataset by building a *Schema Summary* which contains the main classes and properties used within the datasets, whether they are taken from external vocabularies or not, and is conceivable as an RDFS ontology. This method was implemented in the LODeX tool [2,3], which extracts statistical indexes for building the Schema Summary, by querying the SPARQL endpoint of a LOD source; LODeX allows users to compose visual queries by selecting objects from the Schema Summary and thus supporting users in exploring and understanding the contents of a LOD source. For a complete description see `http://dbgroup.ing.unimore.it/lod`.

The great majority of open data is normally published in an unstructured format and is typically accessed only by closed communities. In [56] we proposed a semi-automatic methodology for facilitating resource providers in publishing public data into the LOD cloud, and for helping consumers in efficiently accessing and querying them. The methodology was applied on the research project on Youth Policies of the Emilia-Romagna Region (”Open linked data Osservatorio Giovani della Regione Emilia-Romagna”) [8]. The project goals were to identify interesting data sources both from the open data community and from the private repositories of local governments of Emilia Romagna region related to the Youth Policies, to integrate them and, to show up the result of the integration by means of a useful navigator tool; in the end, to publish new information as LOD. We firstly have analyzed the useful open data sources, then using the MOMIS system, we have integrated them with the proprietary data provided by the project partners; the MOMIS Dashboard was then used to provide an easy access to the data. Finally we published the integrated data as LOD. This project has exemplified how a Public Administrations can benefit from the use of Open Data and can effectively extract new and important information by integrating its own datasets with open data sources.

References

1. I. Bartolini, D. Beneventano, S. Bergamaschi, P. Ciaccia, A. Corni, M. Orsini, M. Patella, and M. M. Santese. MOMIS goes multimedia: WINDSURF and the case of top-k queries. In *SEBD'15, Gaeta, June 14-17, 2015.*, pages 200–207, 2015.
2. F. Benedetti, S. Bergamaschi, and L. Po. Lodex: A tool for visual querying linked open data. In *ISWC'15 Posters & Demonstrations Track*, 2015.
3. F. Benedetti, S. Bergamaschi, and L. Po. Visual querying LOD sources with lodex. In *K-CAP'15, Palisades, NY, USA, October 7-10, 2015*, pages 12:1–12:8, 2015.
4. D. Beneventano. Provenance based conflict handling strategies. In *DASFAA'12, Busan, South Korea, 15-18 April 2012*, pages 286–297, 2012.
5. D. Beneventano and S. Bergamaschi. The momis methodology for integrating heterogeneous data sources. In *IFIP 18th World Computer Congress 22–27 August 2004 Toulouse, France*, pages 19–24. Springer US, 2004.
6. D. Beneventano and S. Bergamaschi. Provenance-aware semantic search engines based on data integration systems. *IJOCI*, 4(2):1–30, 2014.
7. D. Beneventano, S. Bergamaschi, and A. R. Dannaoui. Integration and provenance of cereals genotypic and phenotypic data. In *SEBD'12*, pages 91–98, 2012.
8. D. Beneventano, S. Bergamaschi, L. Gagliardelli, and L. Po. Driving innovation in youth policies with open data. In *IC3K'15, Revised Selected Papers*, Communications in Computer and Information Science. Springer, 2016.
9. D. Beneventano, S. Bergamaschi, F. Guerra, and M. Vincini. The SEWASIE network of mediator agents for semantic search. *J. UCS*, 13(12):1936–1969, 2007.
10. D. Beneventano, S. Bergamaschi, and R. Martoglia. Exploiting semantics for searching agricultural bibliographic data. *Journal of Information Science*, 42(6):748–762, 2016.
11. D. Beneventano, S. Bergamaschi, S. Sorrentino, M. Vincini, and F. Benedetti. Semantic annotation of the CEREALAB database by the AGROVOC linked dataset. *Ecological Informatics*, 26(2):119–126, 2015.
12. D. Beneventano, A. R. Dannaoui, and A. Sala. On provenance of data fusion queries. In *SEBD'11, June 26-29, 2011*, pages 84–94, 2011.
13. D. Beneventano, C. Gennaro, S. Bergamaschi, and F. Rabitti. A mediator-based approach for integrating heterogeneous multimedia sources. *Multimedia Tools Appl.*, 62(2):427–450, 2013.
14. D. Beneventano, F. Guerra, S. Magnani, and M. Vincini. A web service based framework for the semantic mapping amongst product classification schemas. *Journal of Electronic Commerce Research*, 5(2):114–127, 2004.
15. D. Beneventano, F. Guerra, A. Maurino, M. Palmonari, G. Pasi, and A. Sala. Unified semantic search of data and services. In *MTSR'09*, pages 95–107, 2009.
16. D. Beneventano, S. E. Haoum, and D. Montanari. Mapping of heterogeneous schemata, business structures, and terminologies. In *Workshop at DEXA'07*, pages 412–418, 2007.
17. D. Beneventano, M. Olaru, and M. Vincini. Analyzing dimension mappings and properties in data warehouse integration. In *OTM'13*, pages 616–623, 2013.
18. S. Bergamaschi, D. Beneventano, F. Guerra, and M. Orsini. Data integration. In D. W. Embley and B. Thalheim, editors, *Handbook of Conceptual Modeling: Theory, Practice and Research Challenges*. Springer Verlag, 2011.
19. S. Bergamaschi, D. Beneventano, F. Guerra, and M. Vincini. Building a tourism information provider with the MOMIS system. *Journal of Information Technology & Tourism*, 7(3-4):221–238, 2004.

20. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Records*, 28(1), March 1999.
21. S. Bergamaschi, E. Domnori, F. Guerra, M. Orsini, R. Trillo-Lado, and Y. Velegrakis. Keymantic: Semantic Keyword-based Searching in Data Integration Systems. *PVLDB*, 3(2), 2010.
22. S. Bergamaschi, E. Domnori, F. Guerra, R. Trillo-Lado, and Y. Velegrakis. Keyword search over relational databases: a metadata approach. In *SIGMOD*, pages 565–576. ACM, 2011.
23. S. Bergamaschi, D. Ferrari, F. Guerra, G. Simonini, and Y. Velegrakis. Providing insight into data source topics. *J. Data Semantics*, 5(4):211–228, 2016.
24. S. Bergamaschi, N. Ferro, F. Guerra, and G. Silvello. Keyword-based search over databases: A roadmap for a reference architecture paired with an evaluation framework. *Trans. Computational Collective Intelligence*, 21:1–20, 2016.
25. S. Bergamaschi, F. Guerra, M. Interlandi, R. T. Lado, and Y. Velegrakis. QUEST: A keyword search system for relational data based on semantic and machine learning techniques. *PVLDB*, 6(12):1222–1225, 2013.
26. S. Bergamaschi, F. Guerra, M. Interlandi, R. T. Lado, and Y. Velegrakis. Combining user and database perspective for solving keyword queries over relational databases. *Inf. Syst.*, 55:1–19, 2016.
27. S. Bergamaschi, F. Guerra, S. Rota, and Y. Velegrakis. A hidden markov model approach to keyword-based search over relational databases. In *ER*, LNCS 6998, pages 411–420. Springer, 2011.
28. S. Bergamaschi, L. Po, and S. Sorrentino. Automatic annotation for mapping discovery in integration systems. In *SEBD'08*, pages 334–341, 2008.
29. J. Bleiholder and F. Naumann. Data fusion. *ACM Comp. Surv.*, 41:1–41, 2008.
30. G. H. L. Fletcher and F. Mandreoli. No users no dataspaces! query-driven dataspace orchestration? In *Proc. of SEBD*, pages 150–157, 2016.
31. B. Glavic, G. Alonso, R. J. Miller, and L. M. Haas. Tramp: Understanding the behavior of schema mappings through provenance. *PVLDB*, 3(1):1314–1325, 2010.
32. M. Golfarelli, F. Mandreoli, W. Penzo, S. Rizzi, and E. Turricchia. Towards OLAP query reformulation in peer-to-peer data warehousing. In *Proc. of ACM (DOLAP)*, pages 37–44, 2010.
33. A. Y. Halevy, M. J. Franklin, and D. Maier. Principles of dataspace systems. In *ACM PODS*, pages 1–9, 2006.
34. A. Y. Halevy, Z. G. Ives, D. Suciu, and I. Tatarinov. Schema mediation for large-scale semantic data sharing. *VLDB J.*, 14(1):68–83, 2005.
35. J. Hammer, M. Stonebraker, and O. Topsakal. Thalia: Test harness for the assessment of legacy information integration. In *ICDE*, pages 485–486, 2005.
36. M. Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246, 2002.
37. R. Lenzi, C. Gennaro, F. Mandreoli, R. Martoglia, M. Mordacchini, W. Penzo, and S. Sassatelli. A unified multimedia and semantic perspective for data retrieval in the semantic web. *Inf. Syst.*, 36(2):174–191, 2011.
38. J. N. Levi. *The syntax and semantics of complex nominals.* Academic Press, 1978.
39. F. Mandreoli and R. Martoglia. Knowledge-based sense disambiguation (almost) for all structures. *Inf. Syst.*, 36(2):406–430, 2011.
40. F. Mandreoli, R. Martoglia, and W. Penzo. Approximating expressive queries on graph-modeled data: The gex approach. *J. of Systems and Software*, 2015(109):106–123, 2015.
41. F. Mandreoli, R. Martoglia, W. Penzo, and S. Sassatelli. Data-sharing p2p networks with semantic approximation capabilities. *IEEE IC*, 13(5):60–70, 2009.

42. F. Mandreoli, R. Martoglia, W. Penzo, S. Sassatelli, and G. Villani. Sri@work: Efficient and effective routing strategies in a pdms. In *WISE*, pages 285–297, 2007.
43. F. Mandreoli, R. Martoglia, W. Penzo, S. Sassatelli, and G. Villani. Building a pdms infrastructure for xml data sharing with sunrise. In *EDBT-DATAX*, 2008.
44. F. Mandreoli, R. Martoglia, W. Penzo, and G. Villani. Flexible query answering on graph-modeled data. In *Proc. of EDBT 2009*, pages 216–227, 2009.
45. F. Mandreoli, R. Martoglia, and E. Ronchetti. Versatile structural disambiguation for semantic-aware applications. In *Proc. of ACM CIKM*, pages 209–216, 2005.
46. F. Mandreoli, R. Martoglia, and E. Ronchetti. Strider: a versatile system for structural disambiguation. In *Proc. of EDBT 2006*, pages 1194–1197, 2006.
47. F. Mandreoli, R. Martoglia, S. Sassatelli, and W. Penzo. Sri: Exploiting semantic information for effective query routing in a pdms. In *Proc. of the ACM CIKM Workshop WIDM*, pages 19–26, 2006.
48. F. Mandreoli, W. Penzo, S. Rizzi, M. Golfarelli, and E. Turricchia. Olap query reformulation in peer-to-peer data warehousing. *Inf. Syst.*, 37(5):393–411, 2012.
49. F. Mandreoli, W. Penzo, S. Sassatelli, S. Lodi, and R. Martoglia. Semantic peer, here are the neighbors you want! In *Proc. of EDBT 2008*, pages 26–37, 2008.
50. J. Milc, A. Sala, S. Bergamaschi, and N. Pecchioni. A genotypic and phenotypic information source: the cerealab database. *Database*, 2011.
51. G. A. Miller. Wordnet: a lexical database for english. *C. ACM*, 38(11):39–41, 1995.
52. R. J. Miller, D. Fisla, M. Huang, F. Kymlicka, and V. Lee. The amalgam schema and data integration test suite. *www.cs.toronto.edu/ miller/amalgam*, 2001.
53. S. Rota, S. Bergamaschi, and F. Guerra. The list viterbi training algorithm and its application to keyword search over databases. In *CIKM*, pages 1601–1606, 2011.
54. G. Simonini and S. Bergamaschi. Enhancing entity resolution efficiency with loosely schema-aware techniques. pages 270–277, 2016.
55. G. Simonini, S. Bergamaschi, and H. V. Jagadish. BLAST: a loosely schema-aware meta-blocking approach for entity resolution. *PVLDB*, 9(12):1173–1184, 2016.
56. S. Sorrentino, S. Bergamaschi, E. Fusari, and D. Beneventano. Semantic annotation and publication of linked open data. In *Computational Science and Its Applications - ICCSA 2013*, pages 462–474, 2013.
57. S. Sorrentino, S. Bergamaschi, and M. Gawinecki. NORMS: an automatic tool to perform schema label normalization. In *ICDE'11*, pages 1344–1347, 2011.
58. S. Sorrentino, S. Bergamaschi, M. Gawinecki, and L. Po. Schema label normalization for improving schema matching. *DKE*, 69(12):1254–1273, 2010.
59. M. Vincini, D. Beneventano, and S. Bergamaschi. Semantic integration of heterogeneous data sources in the momis data transformation system. *J. UCS - Journal of Universal Computer Science*, 19(13):1986–2012, 2013.
60. G. Wiederhold. Intelligent integration of information. In *SIGMOD'93, Washington, D.C., May 26-28, 1993*, pages 434–437. ACM Press, 1993.