Univ. of Leipzig
Database group

UNIVERSITÄT LEIPZIG

Research Report 2016/2017

Web: http://dbs.uni-leipzig.de

Overview

Database group in May 2017. *f.l.t.r:* Prof. Dr. Erhard Rahm, Dr. Eric Peukert, Ying-Chi Lin, Martin Grimmer, Alieh Saeedi, Giacomo Bergami (guest), Victor Christen, Dr.- M. Ali Rostami, Markus Nentwig, Elias Saalmann (student), Dr. Anika Groß, Matthias Kricke, Johannes Geisler (student), Stephan Kemper (student), Martin Franke, Ziad Sehili, André Petermann

1. Staff

Professor	Prof. Dr. Rahm, Erhard
Research associate (BMWI)	Alkhouri, Georges (since Sep. 2017)
Research associate	Christen, Victor
Research associate	Franke, Martin (since May 2017)
Research associate (BMBF)	Gladbach, Marcel (since Oct. 2017)
Research associate (BMBF)	Grimmer, Martin (since Sep. 2016)
Research associate	Dr. Groß, Anika (until Sep. 2017)
Secretary	Hesse, Andrea
Research associate	Junghanns, Martin
Research associate (BMBF)	Kricke, Matthias (since Sep. 2016)
Research associate (DFG)	Lin, Ying-Chi (since Apr. 2016)
Research associate (DFG)	Nentwig, Markus
Research associate (BMBF)	Petermann, André
Postdoctoral Researcher (BMBF)	Dr. Peukert, Eric
Research associate (BMBF)	Dr. Rostami, Mohammad Ali (since Apr. 2017)
Ph. D. student	Saeedi, Alieh (since May 2016)
Research associate	Sehili, Ziad
Associated team member	Prof. Dr. Thor, Andreas (HfTL Leipzig)

2. Highlights

There have been several highlights in 2016 and 2017:

1. Microsoft Research, one of the world-leading research organizations in computer science, has recognized Prof. Rahm with an Outstanding Collaborator Award in 2016. This was in recognition of his research together with Phil Bernstein on schema matching and related topics.
2. In July 2016, the Big Data Center of Excellence ScaDS (Competence Center for Scalable Data Services and Solutions) Dresden/Leipzig has hosted its first summer school for Big Data in Leipzig. About 80 participants attended it.
3. The BMBF-funded Big Data center ScaDS Dresden/Leipzig was successfully evaluated after three years in Sep. 2017 so that its funding will likely be continued until Sep. 2021
4. Several ScaDS-related third-party research projects could be secured and started, in particular on security, graph analytics and data integration topics *(Exploids*, *BIGGR*, *TIQ* and *KORBA*).
5. In 2016 the collaborative BMBF project Leipzig Health Atlas – LHA has started to provide a platform for sharing highly annotated data, usable models and software tools for medical data analysis. The database group is contributing to this project for the topic of semantic annotations, especially for medical forms.

6. The project *SMITH* (Smart Medical Information Technology for Health Care) as part of the *Medical Informatics Initiative* has been granted and will start in 2018. The database group will work on privacy-preserving record linkage within this consortium.
7. In the beginning of 2017, Prof. Rahm, A. Groß, Z. Sehili and V. Christen stayed at the Australian National University (ANU) to continue the research collaboration with Prof. Peter Christen, especially on privacy-preserving data integration.
8. The following guests visited the database group: Peter Christen, Dinusha Vatsalan, Qing Wang, Dennis E. Shasha and Giacomo Bergami.
9. The demo on graph grouping in Gradoop has won the Best Demo Award at the BTW conference 2017 in Stuttgart.
10. In May 2017, the traditional Zingst research seminar of the database group took place for the 15th time at the Leipzig University branch in Zingst / Baltic Sea.

Participants of the international Big Data summer school in Leipzig (July 2016)

3. Research topics and projects

ScaDS – Competence Center for Scalable Data Services and Solutions Dresden/Leipzig

E. Rahm, E. Peukert, Z., Sehili, M. Junghanns, A. Petermann, M. Kricke

The „Competence Center for Scalable Data Services and Solutions Dresden/Leipzig (ScaDS)“ lead by Prof. Nagel from the TU Dresden and Prof. Rahm from the University of Leipzig is a nationwide competence center for Big Data in Germany. The center is funded by the BMBF (German Federal Ministry of Education and Research). In addition to the Univ. of Leipzig, there are three further funded partners: TU Dresden, Leibniz Institute of Ecological Urban and Regional Development, and the Max Planck Institute of Molecular Cell Biology and Genetics.

ScaDS Dresden/Leipzig bundles the diverse research competences of the participating institutes and addresses Big Data challenges in a holistic and application-oriented manner. The initial research areas include “Efficient Big Data Architectures”, “Data Quality and Integration”, “Knowledge Extraction”, “Visual Analysis” as well as “Data Life Cycle Management and Workflows”. Furthermore, the competence center tackles a broad range of application from the Life Sciences, Material Sciences, Environmental and Transport Sciences, Digital Humanities and from the business world.

ScaDS has become a success story at the University of Leipzig and its partner institutions. About 120 scientific publications were published and ScaDS members presented more than 200 keynotes and talks in the first three years (2014-2017). New scientific Big Data methods and solutions have been developed such as for large-scale graph analytics, data integration and visual analysis. Together with application partner many challenges were solved which partly resulted in a number of Big Data Service that are offered by ScaDS. Through ScaDS a number of industry contacts were established and many collaborations were started which partly already lead to further research projects that run in close collaboration with ScaDS.

An exemplary application project has been conducted with the Department of Analytical Chemistry at Helmholtz Centre for Environmental Research (UFZ) to manage and analyze mass spectrometry data. To substantially improve the flexibility and scalability of analysis workflows for such data, a new end-to-end analytics platform for such data has been developed that utilizes the KNIME workflow system, an Oracle database and the statistical programming language R.

The database group of Prof. Rahm is involved in many ScaDS projects, in particular related to graph analytics (Gradoop) and data integration including privacy-preserving record linkage.

Distributed Large-Scale Graph Analytics

M. Junghanns, A. Petermann, E. Peukert, E. Rahm

Processing highly connected data as graphs becomes more and more important in many different domains. Prominent examples are social networks, e.g., Facebook and Twitter, as well as information networks like the World Wide Web or biological networks. One important similarity of these domain specific data is their inherent graph structure, which makes them eligible for analytics using graph algorithms. Besides that, the datasets share two more similarities: they are huge in size, making it hard or even impossible to process them on a single machine and they are heterogeneous in terms of the objects they represent and their attached data. With the objective of analyzing these large-scale, heterogeneous graphs, we developed a framework called “Gradoop” (Graph Analytics on Hadoop). Gradoop is built around the so-called Extended Property Graph Model (EPGM) which supports not only single graphs but also collections of heterogeneous graphs and includes a wide range of composable operators. These operators allow the definition of complex analytical programs as they take single graphs or graph collections as input and result in single graphs or graph collections. Gradoop is built on top of Apache Flink and Apache HBase, and makes use of the provided APIs to implement the EPGM and its operators.

The prototype is publicly available (www.gradoop.com), a first use case is the BIIIG project for graph analytics in business information networks. Furthermore, a benchmark on artificial social network data with up to 11 billion edges has been conducted and published. Additional operators, e.g., for graph grouping and pattern matching, have been implemented and the corresponding articles published.

In our ongoing work, we focus on generalization for industry use cases, performance optimization and the addition of graph data integration operators and algorithms. Two separate projects could be started in cooperation with industry partners to facilitate the use of Gradoop for interactive graph analysis. In particular, the BIGGR project studies how to make Gradoop widely usable within workflows defined with the well-known Knime workflow approach. The project "Development of an interactive tool for the visual analysis of very large graph data", funded by the Sächsische Aufbau Bank (SAB), will focus on interactive analysis of networked data for business intelligence. This project is a cooperation with TIQ Solutions GmbH, Leipzig.

BIGGR – Big Graph Data Analysis Workflows

M. Ali Rostami, S. Dienst, E. Peukert, E. Rahm

The analysis of big and network-structured data is a current trend in different fields like biological or social networks. This data, which can be interpreted as graphs, is central to many fields for extracting different information. Typical processes are data import, integration, transformation, analysis of corresponding graphs, and finally the visualization with the goal of identifying the relations and influences of data. However, the classical databases are not flexible enough and do not have a suitable support of analysis workflows and algorithms. Also, the modeled dependencies and parameters in the analysis cannot be supported. On the other hand, the existing graph databases are very technical for the analysis big graphs by end users.

The goal of BIGGR is to develop a new software system for user-friendly and efficient analysis and visualization of big graphs. This system should be usable without deep knowledge. More clearly, the graph analysis workflows can be defined and executed graphically from simple basic operators. Then, the user can see a graphical view of the results at the end. For this purpose, the KNIME Analytics platform and the Gradoop framework are available as open source systems from both partners of this project, which are specifically adapted, expanded and combined. In addition, the system should be easy to extend with new operators, execution target systems and visualization techniques. Practical suitability is evaluated for different application cases. Being open-source makes the results widely usable for data analysts in Germany and worldwide.

Distributed Large-Scale Graph Data Integration

E. Peukert, A. Saeedi, M. Kricke, E. Rahm

Graph analytics and business intelligence processes as introduced in Gradoop and BIIIG are highly data driven processes. It is essential for the process to use different data sources containing heterogeneous graphs to enrich analytical pipelines. To enhance the capabilities of Gradoop in the data integration domain two frameworks working on top of Gradoop are under research and development. The frameworks are FAMER (FAst Matching and Entity Resolution) and GrETL (Graph ETL). FAMER supports the matching and clustering of entities from many data sources and utilizes Apache Flink for a distributed execution. It supports different blocking and matching strategies to first determine a so-called similarity graph. Furthermore, several clustering schemes can be applied to determine clusters of matching entities from similarity graphs. GrETL supports the holistic process of fusing the structure of multiple sources and there properties by providing different strategies for the information fusion. It uses FAMER for the clustering and identification of similar vertices.

Determining Annotations in the Life Sciences

V. Christen, Y.-C. Lin, A. Groß, E. Rahm

The annotation of real-world objects with concepts from an ontology is a common approach in the life sciences, e.g. to facilitate the semantic analysis of data such as electronic health records and to support data exchange. In our previous work, we investigated the annotation of medical forms used in clinical studies, e.g., forms asking for eligibility criteria (e.g. specific disease symptoms). Often there are many heterogeneous forms for similar topics impeding the integration of study results. To overcome such issues, it is a crucial aim to annotate medical forms

with standardized vocabularies such as the Unified Medical Language System (UMLS). Therefore, we developed novel methods to (semi-) automatically annotate medical forms. However, automatic matching of form questions (items) is a complex task since questions are written in free text, use different synonyms for the same semantics and can cover several different medical concepts. Our annotation workflow includes several preprocessing steps, different linguistic match approaches and a novel group-based strategy to select the most promising concepts for annotating a question in the medical form.

We extended the initial solution by developing a novel reuse-based approach that uses existing verified annotations to identify new annotations for so far not annotated medical documents. Moreover, we improved the group-based selection strategy by including term co-occurrences in verified annotation sets as well as ontological relations between the concepts.

Moreover, we evaluated existing annotation tools such as Ctakes and MetaMap and investigated how to combine their results with the results of our annotation approach. In particular, we proposed the use of machine learning to classify each annotation based on the determined scores that each tool provides. In general, an annotation corresponds to a vector where each entry represents a score of a tool. To generate a classification model, we utilize a sample of labeled annotations with given scores. The resulting model is used to classify unclassified annotations that were determined by a set of tools. The results show, that a machine-learning based approach significantly improves the annotation quality.

For future work, we plan to utilize neural networks for representing concepts and documents as embeddings. We plan to use the determined embeddings for computing the similarities between document fragments and concepts and also for disambiguation and annotation.

ELISA - Evolution of Semantic Annotations

Y.-C. Lin, V. Christen, A. Groß , E. Rahm

Ontologies have been used frequently in biomedical fields to annotate genes, electronic health records (EHRs) or literatures. The changes of the ontologies through consistently released new versions can have influence on the linked annotations and therefore impact the functionality of their applications. The research project ELISA (Evolution of Semantic Annotations) investigate the impact of these ontology changes on their annotations as well as determining methods to keep annotations up-to-date despite changes in the ontology.

Together with the Luxembourg Institute of Science and Technology (LIST), the University of Paris-Sud, we quantified how many annotations are affected by ontology evolution and investigated the types of impacts on the annotations. We also proposed a model that supports annotation evolution. Based on these observations, we designed a framework for the (semi)-automatic maintenance of semantic annotations (MAISA).

In addition to developing this framework, we also focus on improving (semi)-automatic annotation tools. For this purpose, we evaluated two well-established annotators, MetaMap and cTAKES, and the one previously developed by our research group, AnnoMap. Furthermore, we investigated how to improve the annotation quality of the tools by post-filtering computed annotations as well as by combining several annotation approaches by applying set operations. Our approaches can significantly improve the annotation quality of cTAKES and MetaMap.

Another focus on the ELISA project is to investigate the evolution of the documents being annotated. In this case, we target the multi-lingual evolution of the Case Report Forms (CRFs) used in epidemiological studies. Many of these forms are derived from English versions and have been translated into other languages. We are currently developing annotation strategies for annotating these translated forms and to improve the annotation quality.

LHA - Leipzig Health Atlas (LHA)

Y.-C. Lin, A. Groß, E. Rahm

Data sharing in medicine and clinical trials has gained more and more attention. It has the potential to enhance research efficiency and quality significantly, for instance, by enabling data integration of different studies for further meta-analyses or to allow reuse of tools or models that developed by previous studies. The aim of the project Leipzig Health Atlas (LHA) is to facilitate data sharing between medical and scientific communities. In this interdisciplinary project, researchers from different institutions of University of Leipzig cooperate, in particular from the Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig Research Center for Civilization Diseases (LIFE), Interdisciplinary Centre for Bioinformatics (IZBI), and our research group. We will provide models, tools, data and metadata to enhance the research in systems medicine, clinical trials and medical ontology.

One of the approaches to assist data sharing in LHA is to semantically enrich the data collected by applying semi-automatic annotation tools. To enhance the interoperability in epidemiological research, we aim to enrich the metadata in the Case Report Forms (CRFs) used in LIFE using standardized vocabularies from the Unified Medical Language System (UMLS). We have investigated different annotation tools using real-world medical forms from the Medical Data Models portal to give us indication, which components of different tools are useful for such corpus and can improve the tools to suit our needs. A further challenge is that our task is a multilingual biomedical annotation problem, i. e. the CRFs are in German while the biomedical ontologies are mostly available only in English. To determine the best annotation strategies, we manually gathered the English versions of the CRFs from which the German forms are deribed and built a parallel corpus. We have adapted our AnnoMap tool to annotate the German CRFs using UMLS German version and compared the results of the English CRFs using UMLS English version. We concluded that the German CRFs shall be translated into English due to limited amount of entries available in UMLS German version. We further determined the most relevant subset of UMLS English ontologies for our CRFs using cTAKES. Currently, we are building a silver standard manually to enable the quality evaluation of the annotation results.

The LHA platform for data sharing can be reached under https://www.health-atlas.de/

Privacy Preserving Record Linkage

Z. Sehili, M. Franke, M. Gladbach, E. Rahm

Record linkage aims at linking records that refer to the same real-world entity, such as persons. Typically, there is a lack of global identifiers, therefore the linkage can only be achieved by comparing available quasi-identifiers, such as name, address or date of birth. However, in many cases, data owners are only willing or allowed to provide their data for such data integration if there is sufficient protection of sensitive information to ensure the privacy of persons, such as patients or customers. *Privacy Preserving Record Linkage (PPRL)* addresses this problem by providing techniques to match records while preserving their privacy allowing the combination of data from different sources for improved data analysis and research. For this purpose, the linkage of person-related records is based on encoded values of the quasi-identifiers and the data needed for analysis (e. g., health data) is separated from these quasi-identifiers. PPRL is confronted with several challenges needing to be solved to ensure its practical applicability. In particular, a high degree of privacy has to be ensured by suitable encoding of sensitive data and organizational structures, such as the use of a trusted linkage unit. PPRL must achieve a high linkage quality by avoiding false or missing matches. Furthermore, a high efficiency with fast linkage time and scalability to large data volumes are needed. A main problem for performance is the inherent quadratic complexity of the linkage problem when every record of the first source is compared with every record of the second source. For better efficiency, the number of comparisons can be reduced by adopting blocking or filtering approaches. Furthermore, the matching can be performed in parallel on multiple processing nodes.

Our research concentrated on improving the performance and scalability of our PPRL methods. As the records are represented as bit vectors (using Bloom filter for encoding), we used metric space similarity measures for filtering. In particular, the pivot-based approach for metric spaces utilizing the triangle inequality to reduce the search space showed significant improvement of performance compared to previous filter techniques. One data source is in-

dexed by determining some records as pivots and assigning the leftover records to them. We can save many similarity computations by comparing the records of the second source with only the pivots first and exclude most records as possible matches. In 2016 and 2017, we parallelized this approach utilizing the modern framework Apache Flink, which supports efficient distributed in-memory processing. We developed parallel algorithms for both determining the pivots and the pivot-based matching process for different strategies to select the pivots. The distributed implementation outperforms the centralized version for large dataset and improves the scalability of this approach.

Additionally, we investigated locality sensitive hashing (LSH) as blocking method and developed a distributed algorithm using Apache Flink. LSH enables probabilistic blocking by applying hash functions on bit vectors (Bloom filters). LSH-based blocking supports a flexible configuration and can be applied on encoded data. An extensive evaluation has shown that LSH-based blocking can achieve high quality and high scalability up to millions of records. In contrast, phonetic blocking approaches based on Soundex are susceptible to cryptanalysis and more sensitive with respect to data quality problems. Moreover, phonetic blocking based on a single attribute is not feasible for large datasets due to the low number of blocks and data skew effect introduced by frequent names.

Furthermore, we analyzed methods for multiparty PPRL considering applications, where more than two data holder are involved in the process of privacy preserving record linkage. The most challenging case of multiparty PPRL is to find subset matches, i. e. sets of matching records that are not in all sources, but in subsets of them. The intuitive way is to link the sources sequentially using for example the metric space as a growing index structure. In each iteration a new source is linked with the other already matched sources, then indexed. At the end of this process resulting clusters need to be cleaned if we consider that the sources are deduplicated (only one records from each source). Solutions to this problem are still a research topic.

For future work we plan to bring those PPRL approaches into practical applications, e.g., for the SMITH consortium within the *Medical Informatics Initiative*. We will also develop a toolbox to bring together our developed methods and make them available for broader application.

DAAD project: Advancing Data Integration: Privacy and Semantics for Record Linkage

E. Rahm, Z. Sehili, V. Christen, A. Groß

This project is a joint DAAD research project between the Australian National University and the University Leipzig. The project focuses on privacy-preserving record linkage and the semantic matching of historical census data. The focus on privacy-preserving record linkage (PPRL) has been on the support for more than two parties (data owners). One approach utilizes the idea of so-called Counting Bloom Filters (CBF). CBF support a higher degree of privacy compared to regular bloom filters while still allowing the linking of encrypted records. A detailed evaluation showed that the proposed multiparty protocols also outperform previous approaches.

We also developed effective semantic matching approaches, in particular, group-based linkage methods. We initially focused on temporal semantic matching approaches for census data including person records and their familial relationships. The temporal linkage of census data allows the detailed analysis of population-related changes in an area of interest. It should not only link records about the same person but also support the linkage of groups or clusters of related persons such as households. Our approach is based on semantic features such relationships between different persons e. g. living in one household. The use of semantic features and a group-based linking approach could improve the effectiveness of the linkage compared to standard record linkage approaches. We further built a semantic evolution graph to enable insightful analysis of household changes determined by so-called evolution patterns. We evaluated the semantic matching approach using data from UK.

Exploids – Explicit Privacy-Preserving Host Intrusion Detection System

M. Grimmer, E. Peukert, E. Rahm

The research project Explicit Privacy-Preserving Host Intrusion Detection System (EXPLOIDS, www.exploids.de) aims to increase the security of virtual machines in data centers and cloud environments. The idea is to monitor a Linux guest system for attacks as well as to ensure that any traces of an attack are detected for legal clarification at a later date and to protect data privacy.

The database group is researching anomaly detection techniques for this application. With such methods, it is possible to detect previous unknown attacks. We focus on graph-based approaches in combination with existing methods to increase recognition rates and reduce false alarm rates. By taking the inherent structure of the underlying data into account, it is possible to gain more insights compared to other known methods.

LOD: Link Discovery and entity clustering

M. Nentwig, A. Groß , E. Rahm

Linked (Open) Data (LOD) uses web standards like RDF and HTTP for storage, publication and linkage of structured, heterogeneous data. Finding semantic relations (links) within and between LOD sources is a great challenge since many data is published, but links between different sources do not always comply with the given requirements: Given the size of sources manual linking techniques are time-consuming such that typically (semi-)automatic techniques are used to discover new links. However, previous techniques produce incomplete or imprecise mappings. To provide a survey on existing approaches we analyzed weak and strong points of (semi-) automatic link discovery systems and derived a generic architecture of link discovery frameworks. In particular, we compare the general approaches as well as functional aspects like utilized matching strategies, runtime optimizations, availability for other researchers, and support for parallel processing. The survey further analyzes the reported performance evaluations.

In the period under review, we developed a holistic clustering approach to increase the quality of links within LOD. Link sets between multiple LOD data sources are used to create clusters of equivalent resources as a basis to derive missing or new links to data sources that are not yet linked. Furthermore, incorrect links can be identified and assigned to the correct equivalence cluster. Based on different semantic type as well as insufficiently high transitive similarity between resources cluster elements are separated. Within a final cluster merge step the approach combines clusters abiding high similarity and data source restrictions. To evaluate the quality gained by the clustering approach we created a manually curated multi source reference data set for the geographic domain and made it available for other researchers. Due to the quantity and size of published LOD sources, this process is laborious. We therefore also implemented the clustering approach on top of the distributed dataflow framework Apache Flink and showed scalability for data sets with several million resources on different domains. Additionally to the current Apache Flink clustering approach we plan to add an incremental clustering to enable the addition of data sources to already existing clusters without reprocessing all contained resources.

KOBRA: learning-based deduplication of customer data

G. Alkhouri, E. Peukert, E. Rahm

For businesses, well-maintained customer and partner data is often the most valuable asset they own. A key challenge to achieve a high quality of this data is to identify and eliminate duplicates. Uniserv GmbH, Pforzheim., is offering a sophisticated rule-based system ti find such duplicates in customer data. A key goal n the joint project KOBRA (Configuration of Business Rules for Users of Duplicate Detection Systems) is to simplify the configuration of this tool by applying learning-based methods. The project should help Data Stewards, Data (Quality) Analysts and Citizen Data Scientists to perform data preparation and tool configuration to a large extent automatically via easy-to-understand and agile self-service tools. In doing so, task and company-specific rules will be adapted to the specific problem by adding positive and negative samples given by the users. This, in turn shall be achieved by a combination of different machine learning techniques with training data selection, historization, reinforcement learning, and a simulation environment.

4. Publications and theses

Books

- Mitschang, B.; Ritter, N.; Schwarz, H.; Klettke, M.; **Thor, A**.; Kopp, A.; Wieland, M. (eds.): Datenbanksysteme für Business, Technologie und Web (BTW) – Workshopband, Lecture Notes in Informatics (LNI), P-266, 2017
- Christen, P.; Kemme, B.; **Rahm, E.**; (eds): Proc. of the VLDB 2017 Ph. D. Workshop. CEUR proceedings Vol 1882, Munich, Aug. 2017

Journal publications

- Arnold, P.; Wartner, C.; Rahm, E.: *Semi-Automatic Identification of Counterfeit Offers in Online Shopping Platforms.* Journal of Internet Commerce 15(1), 59-75, 2016
- Groß, A.; Pruski, C.; Rahm, E.: *Evolution of Biomedical Ontologies and Mappings: Overview of Recent Approaches*. Computational and Structural Biotechnology Journal. Vol. 14, 333-340, 2016
- Hirmer, P.; Waizenegger, T.; Falazi, G.; Abdo, M.; Volga, Y.; Askinadze, A.; Liebeck, M.; Conrad, S.; Hildebrandt, T.; Indiono, C.; Rinderle-Ma, S.; Grimmer, M.; Kricke, M.; Peukert, E.: *The First Data Science Challenge at BTW 2017*. Datenbank-Spektrum, 2017
- Junghanns, M., Petermann, A.: *Verteilte Graphanalyse mit Gradoop*. JavaSPEKTRUM 05/2016
- Kricke, M.; Grimmer, M.; Schmeißer, M.: *Preserving Recomputability of Results from Big Data Transformation Workflows Depending on External Systems and Human Interaction*. Datenbank-Spektrum, 2017
- Lin, Y.-C.; Groß, A.; Kirsten, T.: *Integration and visualization of spatial data in LIFE.* it - Information Technology 59(4), 2017
- Nentwig, M., Hartung M., Ngonga Ngomo, A.-C., Rahm, E.: *A survey of current Link Discovery frameworks*. Semantic Web 8(3): 419-436, 2017
- Petermann, A., Junghanns, M.: *Scalable Business Intelligence with Graph Collections.* it - Information Technology, Special Issue: Big Data Analytics 58 (4), 2016.
- Rahm, E.: *Big Data Analytics* (Editorial). it - Information Technology, Special Issue: Big Data Analytics. Vol. 58 (4), 2016, pp. 155–156
- Sehili, Z.; Rahm, E.: *Speeding up Privacy Preserving Record Linkage for Metric Space Similarity Measures*. Datenbank-Spektrum, 2016

Book chapters and conference/workshop publications

- Amann, W.: *Vergleich und Evaluation von RDF-on-Hadoop-Lösungen*. Proc. BTW Workshopband, GI Lecture Notes in Informatics (LNI) P266, 2017
- Cardoso, S.D.; Pruski, Cédric; Da Silveira, M.; Lin, Y.-C.; Groß, A.; Rahm, E.; Reynaud-Delaître, C.: *Leveraging the Impact of Ontology Evolution on Semantic Annotations.* Knowledge Engineering and Knowledge Management, Proc. EKAW conf., Springer LNCS 10024, pp. 68-82, 2016
- Cardoso, S.D.; Reynaud-Delaître, C.; Da Silveira, M.; Lin, Y-C; Groß, A.; Rahm, E.; Pruski, C.: *Towards a Multi-level Approach for the Maintenance of Semantic Annotations.* Proc. 10th Int. Joint Conf. on Biomedical Engineering Systems and Technologies (BIOSTEC 2017), HEALTHINF, 2017
- Christen, V.; Groß, A.; Rahm, E.: *Approaches for Annotating Medical Documents.* Proc. Lernen. Wissen. Daten. Analysen. (LWDA), 2016
- Christen, V.; Groß, A.; Rahm, E.: *A Reuse-based Annotation Approach for Medical Documents*. Proc. 15th International Semantic Web Conference (ISWC), Springer LNCS 9981, pp. 135-150, 2016

- Christen, V.; Groß, A.; Fisher, J.; Wang, Q.; Christen, P.; Rahm, E.: *Temporal group linkage and evolution analysis for census data*. Proc. 19th Int. Conf. on Extending Database Technology (EDBT), 2017
- Junghanns, M.; Petermann, A.; Teichmann, N.; Gomez, K.; Rahm, E.: *Analyzing Extended Property Graphs with Apache Flink*. Proc. Int. SIGMOD workshop on Network Data Analytics (NDA), 2016.
- Junghanns, M.; Kießling, M.; Averbuch, A.; Petermann, A.; Rahm, E.: *Cypher-based Graph Pattern Matching in Gradoop.* Proc. ACM SIGMOD workshop on Graph Data Management Experiences and Systems (GRADES), 2017.
- Junghanns, M.; Petermann, A.; Neumann, M.; Rahm, E.: *Management and Analysis of Big Graph Data: Current Systems and Open Challenges*. Handbook of Big Data Technologies. Springer 2017.
- Junghanns, M.; Petermann, A.; Rahm, E.: *Distributed Grouping of Property Graphs with GRADOOP*. Proc. Datenbanksysteme für Business, Technologie und Web (BTW), GI Lecture Notes in Informatics (LNI) P265, 2017
- Junghanns, M.; Petermann, A.; Teichmann, N.; Rahm, E.: *The Big Picture: Understanding large-scale graphs using Graph Grouping with GRADOOP.* Proc. Datenbanksysteme für Business, Technologie und Web (BTW), GI Lecture Notes in Informatics (LNI) P265, 2017 (Demo paper)
- Kemper, S.; Petermann, A.; Junghanns, M.: *Distributed FoodBroker: Skalierbare Generierung graphbasierter Geschäftsprozessdaten.* Proc. BTW Workshopband, GI Lecture Notes in Informatics (LNI) P266, 2017
- Kricke, M.; Grimmer, M.; Schmeißer, M.: *Preserving Recomputability of Results from Big Data Transformation Workflows*. Proc. BTW Workshopband, GI Lecture Notes in Informatics (LNI) P266, 2017
- Lin, Y.-C.; Christen, V.; Groß, A.; Cardoso, S. D.; Pruski, C.; Da Silveira, M.; Rahm, E.: *Evaluating and improving annotation tools for medical forms.* Proc. Data Integration in the Life Science (DILS), Springer LNCS 10649, 2017
- Nentwig, M.; Groß, A.; Rahm, E.: *A Clustering Approach for Holistic Link Discovery* (Project overview). Proc. Lernen. Wissen. Daten. Analysen. (LWDA), 200-205, 2016
- Nentwig, M.; Groß, A.; Rahm, E.: *Holistic Entity Clustering for Linked Data*. ICDM Workshops, 194-201, 2016
- Nentwig, M.; Groß, A.; Möller, M.; Rahm, E.: *Distributed Holistic Clustering on Linked Data*. OTM Conferences (2), Springer LNCS 10574, pp. 371-382, 2017
- Petermann, A.: *Graph Pattern Mining for Business Decision Support.* Proc. VLDB PhD Workshop 2017
- Petermann, A.; Junghanns, M.; Kemper, S.; Gomez, K.; Teichmann, N.; Rahm, E.: *Graph Mining for Complex Data Analytics.* Proc. IEEE Int. Conf. on Data Mining, 2016 (demo paper)
- Petermann, A.; Micale, G.; Bergami, G.; Pulvirenti, A.; Rahm, E.: *Mining and Ranking of Generalized Multi-Dimensional Frequent Subgraphs*. Proc. Int. Conf. on Digital Information Management (ICDIM) 2017
- Petermann, A.; Junghanns, M.; Rahm, E.: *DIMSpan - Transactional Frequent Subgraph Mining with Distributed In-Memory Dataflow Systems.* Proc. Int. Conf. on Big Data Computing, Applications and Technologies (BDCAT), 2017
- Peukert, E; Wartner, C.: *LEAP Data and Knowledge Integration Infrastructure*. In: Taking the LEAP - Methods and Tools of the Linked Engineering and Manufacturing Platform (LEAP), Academic Press 2016
- Rahm, E.: *The Case for Holistic Data Integration*. Proc. ADBIS, Invited keynote paper, Springer LNCS 9809, 2016
- Saeedi, A.; Peukert, E.; Rahm, E.: *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution.* Proc. ADBIS, LNCS 10509, pp 278-293, 2017

- Vatsalan, D.; Sehili, Z.; Christen, P.; Rahm, E.: *Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges*. Handbook of Big Data Technologies, Springer 2017
- Vatsalan, D.; Christen, P.; Rahm, E.: *Scalable privacy-preserving linking of multiple databases using Counting Bloom filters*. Proc. ICDM workshop on Privacy and Discrimination in Data Mining (PDDM), 2016

Bachelor and Master Theses

1. Amman, W.: *Vergleich und Evaluation von RDF-on-Hadoop-Lösungen*, M. Sc., Univ. Leipzig, 2016
2. Alkhouri, G.: *Deduplizierung durch künstliche neuronale Netze*, M. Sc., Univ. Leipzig, 2017
3. Baumberg, C.: *Evaluierung und Benchmark von verteilten Graph Processing Systemen*, Dipl., Univ. Leipzig, 2016
4. Bugge, Y.: *Performanzanalyse von Alluxio unter Verwendung Java basierter Cluster Computing Frameworks*, M. Sc., Univ. Leipzig, 2017
5. Döring, T.: *Asynchrones Frequent Subgraph Mining mit Apache Storm*, M. Sc., Univ. Leipzig, 2016
6. Franke, M.: *Implementierung und Evaluation von parallelen Verfahren mit Flink zum Schutz personenbezogener Daten beim Record-Linkage*, M. Sc., Univ. Leipzig, 2017
7. Fuß, C.: *Graph-based similarity measures*, B. Sc., Univ. Leipzig, 2017
8. Gladbach, M.: *Verteilte Verfahren für Privacy Preserving Record Linkage unter Verwendung von metrischen Räumen*, M. Sc., HTWK Leipzig, 2017
9. Gomez, K.: *Large-Scale Graph Analyse mittels Apache Giraph und Apache Flink*, B. Sc., Univ. Leipzig, 2016
10. Günther, S.: *Migration des Dokumentenservers*, B. Sc., Univ. Leipzig, 2017
11. Jakob, K.: *Aufbau einer Infrastruktur zur Analyse von Massenspektrometrie-Daten am UFZ*, M. Sc, Univ. Leipzig, 2016
12. Kießling, M.: *Towards Cypher on Apache Flink – Implementing a Graph Query Language on a Data Flow System*, M. Sc., Univ. Leipzig, 2017
13. Lin, Y.-C..: *Algorithms for Map Generation and Spatial Data Visualization in LIFE*. M. Sc., Univ. Leipzig, 2016
14. Hieu, N. D.: *Anpassung des Gaston Algorithmus an das Distributed Dataflow Programmiermodell,* B. Sc., Univ. Leipzig, 2016
15. Pretzsch, F.: *Entwicklung von Techniken zur Datenintegration und Datenqualitätsverbesserung für die Graph-Processing-Platform GRADOOP* (Joint supervision with Prof. Thor from HFTL), M. Sc., 2016
16. Rost, C.: *Skalierbare bildbasierte Deduplikation*, M. Sc., Univ. Leipzig, 2017
17. Saalmann, E.: *Fallstudie zu graphbasierter Business Intelligence mit Gradoop am Beispiel der Immobilienwirtschaft*, B. Sc., Univ. Leipzig, 2017
18. Swoboda, O.: *Serverseitige Aggregation von Zeitreihendaten in verteilten NoSQL-Datenbanken*, M. Sc., Univ. Leipzig, 2017
19. Tran, N. H.: *Visualisierung von Graphdaten für Geschäftsprozesse*, M. Sc., Univ. Leipzig, 2017

5. Talks

- Christen, V.: *Approaches for Annotating Medical Documents*. Lernen. Wissen. Daten. Analysen. (LWDA), Potsdam (Germany), 2016
- Christen, V.: *A Reuse-based Annotation Approach for Medical Documents*. 15th International Semantic Web Conference (ISWC), Kobe (Japan), 2016

- Christen, V.: *Evaluating and improving annotation tools for medical forms*. Data Integration in the Life Science (DILS) conf., Luxembourg, (Luxembourg), 2017
- Grimmer, M.; Kricke, M.; Schmeißer, M.: *Building a real time Tweet map with Flink in six weeks*. Flink Forward 2016, 2016
- Groß, A.: *Reuse of Ontology Mappings*, Australian National University (ANU), Canberra, Australia, March 2016
- Groß, A.: *NoSQL-Data Stores for Big Data*. 2nd International ScaDS Summer School on Big Data, Leipzig, July 2016
- Junghanns, M.*: Gradoop: Scalable Graph Analytics with Apache Flink.* FOSDEM, Bruessel, Feb. 2016
- Junghanns, M.: *Gradoop: Scalable Graph Analytics with Apache Flink,* Graph Fun with Apache Flink & Neo4j, Berlin, Mar. 2016
- Junghanns*: Analyzing Extended Property Graphs with Apache Flink. SIGMOD* NDA workshop, San Francisco, June 2016
- Junghanns, M.: *Distributed Graph Analytics with Gradoop.* Let's talk about Graph Databases, Munich, July 2016
- Junghanns, M.: *Scalable Graph Data Analytics with Gradoop.* BBDC Symposium, Dresden, Nov. 2016
- Junghanns, M.; Kießling M.: *(Cypher)-[:ON]->(ApacheFlink)<-[:USING]-(Gradoop).* FOSDEM 2017, Bruessel, Feb. 2017
- Junghanns, M.: *Distributed Graph Flows: Cypher on Flink and Gradoop*. openCypher Implementers Meeting, Walldorf, Feb. 2017
- Junghanns, M.: *Extended Property Graphs and Cypher on Gradoop*. openCypher Implementers Meeting, Walldorf, Feb. 2017
- Junghanns, M.: *Distributed Graph Analytics with Gradoop*. LDBC TUC Meeting, Walldorf, Feb. 2017
- Junghanns, M.: *Distributed Grouping of Property Graphs with GRADOOP*. BTW conference, Stuttgart March 2017
- Junghanns, M.; Kießling, M.: *Cypher-based Graph Pattern Matching in Apache Flink*. Flink Forward 2017, Berlin, Sep. 2017
- Lin, Y.-C.: *Integration and visualization of spatial data in LIFE*. Health - Exploring Complexity: an interdisciplinary Systems (HEK), Munich, 2016
- Nentwig, M.: *A Clustering Approach for Holistic Link Discovery*. Lernen. Wissen. Daten. Analysen. (LWDA), Potsdam, 2016
- Nentwig, M.: *Holistic Entity Clustering for Linked Data*. ICDM Workshops, Barcelona, Spain, Dec. 2016
- Nentwig, M.: *Distributed Holistic Clustering on Linked Data*. OTM Conferences (ODBASE), Rhodes, Greece, June 2017
- Petermann, A.; Junghanns, M.*: Gut vernetzt: Skalierbares Graph Mining für Business Intelligence,* data2day, Karlsruhe, Oct. 2016
- Petermann, A.: *Graph Mining for Complex Data Analytics*, International Conference on Data Mining (ICDM) 2016, Barcelona, Spain, Dec. 2016
- Petermann, A.: *From Shopping Baskets to Structural Patterns,* FOSDEM 2017, Bruessel, Feb. 2017
- Petermann, A.: *Skalierbare Graph-basierte Analyse und Business Intelligence*, bitkom Bit Data Summit 2017, Hanau, Feb. 2017
- Petermann, A.: *Graph Pattern Mining for Business Decision Support*, VLDB PhD Workshop 2017, Munic, Aug. 2017
- Petermann, A.: *Mining and Ranking of Generalized Multi-Dimensional Frequent Subgraphs*, Int. Conf. on Digital Information Management (ICDIM), Fukuoka, Japan, Sept. 2017

- Petermann, A.: *DIMSpan - Transactional Frequent Subgraph Mining with Distributed In-Memory Dataflow Systems,* Int. Conf. on Big Data Computing, Applications and Technologies (BDCAT), Austin, Texas, Dec. 2017
- Peukert, E.: *Big Data und ScaDS @SEPT*, Leipzig, Nov. 2016
- Peukert, E.: *ScaDS Overview,* Sachsentag der angewandten Informatik, Nov. 2016
- Peukert, E.: *Graph-basierte Business Intelligence mit Gradoop*, Big Data All-Hands, Karlsruhe, Nov. 2017
- Rahm, E.: *Scalable and privacy-preserving data integration* (3 lectures). Big Data Winter School, Bilbao, Feb. 2016
- Rahm, E.: *Überblick ScaDS.* Vortrag vor SMWA-Beirat „Digitale Wertschöpfung", Leipzig, Mai 2016
- Rahm, E.. *Scalable Management and Analysis of Graph Data*. Keynote Computer Science conf. (CSCUBS), Uni Bonn, Mai 2016
- Rahm, E.: *Big Data Integration.* ScaDS summer school, Leipzig, July 2016
- Rahm, E.: *The case for holistic data integration*. ADBIS keynote, Prague, Aug. 2016
- Rahm, E.: *Big Data Integration: challenges and new approaches.* Data integration workshop, Isaac Newton Institute, Cambridge, Sep. 2016
- Rahm, E.: *Scalable graph analytics with Gradoop.* BBDC conference, Berlin, Nov. 2016
- Rahm, E.: *Graph-Based Data Integration and Analysis with Gradoop*. Colloquium talk, TU Dresden, Dec. 2016
- Rahm, E.: *Graph-Based Data Integration and Analysis with Gradoop*. ScaDS Ringvorlesung, Uni Leipzig, April 2017
- Rahm, E.: *Scalable graph analytics.* ScaDS/BBDC summer school, Munich, Aug. 2017
- Rahm, E.: *Ontologies and ontology matching*. Univ. Linkoping, Sep. 2017
- Rahm, E.: *Privacy-preserving data integration for Big Data.* Big Data All-Hands, Karlsruhe, Nov. 2017
- Rahm, E.: *Datenschutzerhaltende Verknüpfung von Patientendaten*. Workshop AG Datenschutz, TMF, Berlin, Nov,. 2017
- Rahm, E.: *Scalable graph analytics.* Distinguished Lecture series, Uni Jena, Dec. 2017
- Saeedi, A: *Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution*. ADBIS conf., Nicosia, Sep. 2017

6. Service / Memberships in Committees and Boards

Christen, V.:

- Journal Reviewer: *Bioinformatics*

Groß, A.:

- Member of the executive committee of the GI group Database Systems
- Co-Chair: BTW Workshop BigDS 2017
- PC Member: DILS 2017, VOILA ISWC Workshop 2016-2017, JOWO 2017, DINA ICDM Workshop 2016, BTW 2017 student program, GI-Workshop ODLS 2016
- Journal Reviewer: a.o. Bioinformatics, Journal of Biomedical Semantics (JBMS), Journal of Web Semantics (JWS), Knowledge and Information Systems (KAIS)

Lin, Y.-C.:

- Additional reviewer for Data Integration in the Life Science (DILS) 2017

Rahm, E.:

- Scientific co-coordinator of the Big Data Center ScaDS Dresden/Leipzig
- Speaker of the executive committee of the GI (German Informatics Society) group Databases and Information Systems
- Elected member of the DFG Review Board for Computer Science
- Steering Committee DILS conference series (Data Integration in the Life Sciences)
- Advisory Board Europar conference series
- Board Member IZBI (Interdisciplinary Centre for Bioinformatics, Leipzig)
- PC co-chair VLDB PhD workshop 2017
- PC group leader SIGMOD 2018
- Tutorial chair EDBT 2018
- Chair ScaDS Big Data summer school, Leipzig, 2016
- Chair ScaDS Workshop Big Data in Business (BIDIB), Leipzig, June 2017
- PC Member of several conferences and workshops (BTW 2017, DILS 2017)
- Reviewer for different journals, research associations etc.