Research Report 2014/2015

Database Group, Univ. of Leipzig

Web: http://dbs.uni-leipzig.de, www.scads.de

Overview

Database group in May 2015. *f.l.t.r:* Ziad Sehili, Eric Peukert, Patrick Arnold, Victor Christen, Martin Junghanns, Simon Chill (student), Anika Groß, Markus Nentwig, Kevin Gomez (student), Prof. Dr. Erhard Rahm, André Petermann

1. Staff

Professor	Prof. Dr. Rahm, Erhard
Research associate (EU)	Dr. Arnold, Patrick
Research associate	Christen, Victor (since Oct. 2014)
Research associate	Dr. Groß, Anika
Secretary	Hesse, Andrea
Research associate	Junghanns, Martin (since April 2014)
Programmer	Jusek, Stefan (until June 2015)
Research associate	Dr. Kolb, Lars (until Sep. 2014)
Research associate (DFG)	Nentwig, Markus
Research associate (ESF, BMBF)	Petermann, André
Research associate (BMBF)	Dr. Peukert, Eric (since Dec. 2014)
Research associate (BMBF)	Sehili, Ziad
Associated team member	Prof. Dr. Thor, Andreas (HfTL Leipzig)
Scientific employee (EU)	Wartner, Christian (until Apr. 2015)

2. Highlights

There have been several highlights in 2014 and 2015:

1. The new BMBF-funded research project „ScaDS (Competence Center for Scalable Data Services and Solutions) Dresden/Leipzig " was officially launched in October 2014 as one of two national competence centers for Big Data. Prof. Nagel (TU Dresden) and Prof. Rahm are the scientific coordinators of ScaDS Dresden/Leipzig.
2. In Oct. 2014, the new research project "ELISA - Evolution of Semantic Annotations" has been granted by the German Research Foundation (DFG) and the National Research Fund Luxembourg (FNR). ELISA is a collaborative project between the Luxembourg Institute of Science and Technology (LIST), the University of Paris-Sud, and the Database Group at Leipzig University to develop and evaluate new methods for the creation and maintenance of semantic annotations.
3. In 2015, a second funding period of another two years has been granted for the DFG research project "LOD Link Discovery - learning-based scalable link-discovery for the Web of Data". It is a research co-operation together with the AKSW group at Leipzig University (Dr. Ngonga).
4. In Dec. 2015 the collaborative project "Leipzig Health Atlas – LHA" has been granted by the Federal Ministry of Education and Research (BMBF). LHA will provide an interoperable semantic platform to share highly annotated data, usable models and working software tools for novel and selective medical decision models. It is an interdisciplinary project together with researchers from the Institute for Medical Informatics, Statistics and Epidemiology (IMISE), the LIFE-Research Center for Civilization Diseases (LIFE), the Interdisciplinary Centre for Bioinformatics (IZBI) and other partners.
5. The EU project LinkedDesign, a cooperation with SAP Research and other research and industry partners, has been successfully finished in April 2015. The consortium has developed an advanced, innovative and integrated software platform to support product designers, engineers and manufacturers.
6. In the beginning of 2015, Prof. Rahm stayed at the Australian National University (ANU) to start a close research collaboration with Prof. Peter Christen, especially on privacy-preserving data integration.

7. The German Academic Exchange Service (DAAD) has granted a collaborative project on advanced data integration topics in 2015. The project supports mutual research visits between the database group and the computer science department of the Australian National University (ANU).
8. Prof. Rahm has been PC Chair for the 10th International Conference on Data Integration in the Life Sciences (DILS 2014) in Lisbon, Portugal.
9. The workshop "Big Data in Business" has been organized by ScaDS Dresden/Leipzig. The workshop took place at Leipzig University in October 2015 and provided a new forum of exchange for different business users of big data technologies and researchers.
10. Prof. Rahm and his colleagues Prof. Saake (Otto von Guericke University Magdeburg) and Prof. Sattler (Ilmenau University of Technology) published a new textbook on distributed and parallel data management („Verteiltes und Paralleles Datenmanagement: Von verteilten Datenbanken zu Big Data und Cloud", Springer, 2015).
11. In 2014, Dr. Eric Peukert received the GFFT dissertation award for his PhD thesis "Process-based Schema Matching: From Manual Design To Adaptive Process Construction".
12. During the period under report four Ph.D. theses could be successfully defended, namely by Anika Groß, Hanna Köpcke, Lars Kolb and Patrick Arnold.
13. The 13th and 14th research seminar of the Database Group took place at the Leipzig University branch in Zingst / Baltic Sea in June 2014 and May 2015.

Workshop "Big Data in Business", November 2015, Leipzig

Excursion during research seminar 2015 in Zingst

3. Projects

ScaDS

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

The „Competence Center for Scalable Data Services and Solutions Dresden/Leipzig (ScaDS)" lead by Prof. Nagel from the TU Dresden and Professor Rahm from the University of Leipzig is a nationwide competence center for Big Data in Germany. The center is funded by the German Federal Ministry of Education and Research and organized by the project partners: Dresden University of Technology, University of Leipzig, Leibniz Institute of Ecological Urban and Regional Development and the Max Planck Institute of Molecular Cell Biology and Genetics.

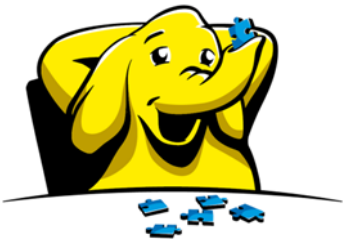
Research activities in the area of Big Data and existing know-how in academia will be focused to support a broad range of different scientific domains to work on challenges in data-intensive computing. The Competence Center for Scalable Data Services and Solutions Dresden/Leipzig will address Big Data challenges holistically and application oriented. It combines the methodical expertise of the universities in Dresden and Leipzig to a virtual organization and brings together leading experts of the environment of Big Data. The initial research activities include "Efficient Big Data Architectures", "Data Quality and Integration", "Knowledge Extraction", "Visual Analysis" as well as "Data Life Cycle Management and Workflows". The major project is set onto data integration, knowledge extraction and visual analysis. Furthermore the competence center also tackles a broad range of application domains that are Life Sciences, Material Sciences, Environmental and Transport Sciences, Digital Humanities and Business Data.

The competence center will develop a comprehensive concept for Big Data Services in an iterative process and will provide them as application-oriented and interdisciplinary solutions for different economic and scientific users.

The Database group of Prof. Rahm is managing the Leipzig part of the Big Data Competence Center with 10 researchers from different computer science groups at the University of Leipzig. Members of the database group are directly involved in the ScaDS competence center working on Big Data Integration as well as Privacy Preserving Record Linkage techniques. Moreover the ScaDS and the Graph Data Management team of the database group collaborate in the development of the highly distributed graph data management and analytics framework Gradoop.

Big Data Integration

E. Rahm, L. Kolb, Z. Sehili, E. Peukert



In the last years, we developed Dedoop, an advanced Hadoop-based framework for parallel object matching. It allows a high-level specification of matching and data integration workflows and utilized automatically generated MapReduce jobs for a highly scalable and parallel execution. In the context of Dedoop several optimizations and functionalities were developed like the execution on Amazon S3, dynamic load balancing, iterative workflows, MapReduce-based parallelization, computation of the transitive closure of a match result, the identification of connected components in huge graphs and the execution on massively parallel graphic processors (GPUs). The Dedoop work has been the main subject of the PhD thesis by L. Kolb who successfully defended it in December 2014.

Big Data applications often involve the processing person-related data which are often subject to strict privacy policies regarding the exchange of information and processing by different parties. For example, it would be desirable that hospitals match their patient-related data in a privacy-preserving manner to investigate possible correlations between some diseases of the same patients but without disclosing the entire databases? Hence the linkage of such data faces the tradeoff between finding useful results efficiently while preserving a high degree of privacy. *Privacy Preserving Record Linkage (PPRL)* addresses this problem by providing different techniques to match records while preserving their privacy. PPRL methods tackle three main challenges: privacy, quality and scalability. Our initially proposed solution evolved from a cooperation with the University of Duisburg. To grant the privacy of the data to be matched, each record is encrypted by using a novel method based on bloom filter. Furthermore two methods were developed to address the scalability problem of the matching process. The first one applies different filters to exclude dissimilar records from comparison. The second method uses dedicated hardware like graphics cards to parallelize the matching process. Our experiments showed that using a single graphic card already leads to significant performance improvements. However, the proposed filter techniques could not further improve the performance. This observation motivated us to investigate other ways like metric spaces to resolve the scalability problem. Metric spaces and their distance functions permit to exclude dissimilar records from comparison in a simple way by applying the property of triangle inequality. We are analyzing the possible application of metric spaces for our records which are represented as bit vectors. The main goal is to reduce the search space by applying the triangle inequality for the two metrics Jaccard and Hamming distances. Furthermore, we investigate different possibilities to index and partition the records by using so-called pivots (reference points). Initial evaluations show that the new schemes are highly effective and outperform previous filter techniques.

Distributed Large-Scale Graph Analytics

E. Rahm M. Junghanns, A. Petermann

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 but also collections of heterogeneous graphs and includes a wide range of combinable 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 build 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 is currently under review. In our ongoing work, we focus on optimization of our existing implementation and the addition of more operators, e.g. for graph pattern matching.

Graph-based Business Intelligence

E. Rahm, A. Petermann, M. Junghanns

Using graph data models for business intelligence applications is a novel and promising approach. In contrast to traditional data warehouse models, graph models enable the mining of relationship patterns. We introduced an approach to graph-based data integration and analytics called BIIIG (Business Intelligence with Integrated Instance Graphs). We implemented an initial prototype based on a productive graph database. Its evaluation with our dedicated data generator "Foodbroker" has shown functional limitations and poor scalability for big data scenarios. Thus, we decided to integrate the BIIIG approach into Gradoop, our novel framework for distributed graph analytics. With this change, we focus on the development of distributed techniques for graph-based business intelligence. Currently, we are investigating a data mining techniques to retrieve meaningful graph patters which correlate with certain business measure values. The major challenge of our ongoing work is the efficient implementation of highly complex mining algorithms based on distributed computing.

LinkedDesign - Next Generation Design, Engineering & Manufacturing

E. Rahm, C. Wartner, P. Arnold, E. Peukert

In cooperation with the Institute for Applied Computer Science Leipzig (InfAI e.V.), the Database Group contributed to the EU-funded project LinkedDesign. Its goal was to develop the Linked Engineering and mAnufacturing Platform (LEAP) as an integrated information system for manufacturing design. LEAP federates relevant information to drive engineering and manufacturing processes, independent of its format, location, originator, and time of creation and provides a holistic view on enterprise data in the product lifecycle process. In 2015 the project had its final review in Brussels with very positive outcome. Within the project the database group developed a service-based architecture to perform schema and object matching. Based on these tools an integrated system for integrating and analyzing process data was developed. A prototype was tested and evaluated in cooperation with the LinkedDesign industrial partners in particular with SAP. The results could be published on the BTW-Conference a PhD could be finished which partially contributed to the project outcome. Moreover, the LinkedDesign partners are close to finishing a Book that will summarize the research results.

Semantic Enrichment of Ontology Mappings and Ontology Merging

E. Rahm, P. Arnold, S. Raunich

Semantic enrichment of ontology mappings plays a key role in data integration, like schema and ontology mapping, ontology merging or ontology evolution. Having knowledge about the semantic relation type of correspondences between schemas or ontologies can foster such data integration tasks, may considerably augment mapping quality and accuracy, and results in more expressive (semantically enriched) mappings. During the last two years, the research prototype STROMA has been developed to carry out such relation type determination. Given an initial mapping created with any arbitrary schema or ontology matcher, STROMA automatically calculates the semantic relation type for each correspondence within the given mapping.

Determining those relation types, like equality ("equal" or "same-as"), subsumption ("is-a" or "less-general"), aggregation ("part-of") or relatedness ("related") is largely based upon linguistics strategies and insights. In many cases, analyzing and comparing the morphological structure of concept names facilitates

the type determination to a high degree. Additionally, large amounts of background knowledge have been integrated in a so-called semantic repository, which is applied as a supplementary strategy for relation type determination. This repository, dubbed SemRep, is an independent background knowledge tool that can be used together with STROMA or standalone. Exploiting background knowledge always incurs a performance cost, yet SemRep has been particularly designed for semantic mapping enrichment and is apt to carry out common mappings within a few seconds. It is a well-scaling, highly configurable, multilingual and extendable solution suitable for general mapping scenarios and domain-specific mappings (e.g. biomedical mappings) alike. STROMA and SemRep are described in detail in the PhD thesis of Dr. Patrick Arnold, who successfully defended his thesis in December 2015.

Semantically enriched mappings are useful to merge two or more ontologies into one unified ontology. Thereby, it is important to retain all information from the initial ontologies. The ontology merging tool ATOM (Asymmetric Target-driven Ontology Merging) can make use of semantic relations (e.g. *equal*, *is-a*) between two different ontologies to generate an integrated ontology. Instead of using a common symmetric approach, ATOM uses a target-driven approach where one of the input ontologies is a fixed target system. The ATOM approach could be successfully applied to large real-life taxonomies from different domains. ATOM generates a default solution in a fully automatic way that may interactively be adapted by users if needed.

LOD Link Discovery

E. Rahm, M. Nentwig, M. Hartung, H. Köpcke, A. Groß

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 a lot of 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. A related topic to link discovery is the matching of real world objects such as in the e-commerce domain. The Ph.D. thesis of Dr. Hanna Köpcke discusses object matching for real-world problems in detail; she defended successfully in May 2014.

In the period under review, we further implemented the new open-access and open-source portal LinkLion in close cooperation with the AKSW group (Department of Computer Science, Universität Leipzig). LinkLion is a central repository to store existing as well as newly computed links from arbitrary data sources and different domains. One of the key intentions of the portal is to supply metadata for links like creator, date and algorithm information enabling the user to understand the provenance and versioning of the mappings. Users can easily upload new links and specify how these were created. Moreover, they are able to select and download link sets as dumps or based on SPARQL queries. Currently, our portal contains nearly 15 million links distributed over more than 3200 mappings. For future work, we plan to extend LinkLion to enable collaboration with external link discovery frameworks and applications.

To increase the quality of links within LOD we are working on a holistic clustering approach based on link sets between several LOD sources. Existing links should be 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. Due to the quantity and size of published LOD sources, this process is laborious. Therefore we make use of the distributed dataflow framework Apache Flink. Future results will be made publicly available on the LinkLion repository.

Determination of ontology and annotation mappings in the life sciences

E. Rahm, V. Christen, A. Groß

The automatic detection of mappings between different ontologies as well as ontology-based annotation of real world objects is an active field of research in the life sciences. Often there are several ontologies for the same sub domain containing overlapping information. It is a crucial aim to identify the interrelations between different ontologies. Moreover, it is useful to annotate biomedical objects with ontology concepts to enable their integration in comprehensive analysis scenarios. Our system GOMMA (Generic Ontology Matching and Mapping Management) already allows for efficient and effective matching of especially large ontologies. GOMMA has been extended by a novel multi-part matching strategy to map clinical vocabularies used by hospitals to the standardized vocabulary LOINC. Beyond this, there is still a need for (semi-) automatic *annotation methods* for real-world biomedical objects such as electronic health records or clinical data.

In the period under review, we investigated the annotation of medical forms typically used in clinical studies. For instance, such forms ask for eligibility criteria (e.g. specific disease symptoms) to include or exclude probands of a clinical trial. 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). Due to the size of UMLS a fully manual annotation process is not feasible or even impossible. 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.

We developed an initial annotation workflow that annotates a set of questions from medical forms with UMLS concepts by using several preprocessing steps, different sophisticated, linguistic approaches and a novel group-based selection strategy. The selection strategy enables the identification of several concepts that are semantically different. We evaluate the efficiency and effectiveness of our approaches based on real-world medical forms from the Medical Data Models portal. Our results have been verified by medical experts. 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. For future work, we plan to develop an annotation-based search that enables the creation of new forms based on clustered questions in a reuse-repository. Moreover, we plan to semantically partition the UMLS to improve the effectiveness and the efficiency of the annotation method.

Ontology and mapping evolution

E. Rahm, A. Groß, V. Christen, M. Hartung

Ontologies are heavily used to semantically describe (annotate) data, especially in the life sciences. Due to new research findings or changed requirements the ontologies underlie continuous changes. These changes can have significant effects on dependent data sources, mappings and software systems. It is

necessary to update invalid ontology-based mappings and applications accordingly. The evolution and adaptation of ontology-based mappings in the life sciences are discussed in the PhD thesis of Dr. Groß who successfully defended in March 2014.

Since life science ontologies are usually very large it is useful to discover change-intensive and stable regions within an ontology. In the period under review we developed the web application Region Evolution Explorer (REX) to detect interesting ontology regions based on their evolution. REX is also available as web service. Users can get a compact overview on differently evolving ontology parts and can discover evolutionary trends over longer time periods. In new research project "ELISA - Evolution of Semantic Annotations" we started to develop novel approaches enabling the adaptation of outdated, ontology-based annotations especially in the biomedical domain. In particular, we plan to automatically adapt annotations between medical forms and ontology concepts (e.g., SNOMED CT) when the underlying ontologies change. To allow for a correct adaptation we will utilize and extend our algorithm COnto-Diff to compute an expressive diff evolution mapping between different ontology versions. In particular, we will enhance change operations by further semantic relationship types like is-a or part-of.

Bibliometrics Analysis

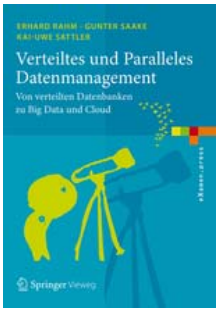
E. Rahm, A. Thor

Bibliometrics quantitatively analyzes scientific publications w.r.t. citation counts and bibliographic metadata. For example, citation analysis determines the impact of publications, venues, and authors as well as their research institutes by measuring the number of citing publications. To this end we developed several easy-to-use applications that can be used for both in-depth analysis of single publications and comparative evaluation of large-scale publication databases. For example, we recently introduced the CitedReferencesExplorer (CRExplorer, www.crexplorer.net) which can be used to disambiguate and analyze the cited references of a publication set downloaded from the Web of Science (WoS). The tool is especially suitable to identify those publications which have been frequently cited by the researchers in a field and thereby to study for example the historical roots of a research field or topic. One feature of the CRExplorer is its ability to perform duplicate detection semi-automatically for efficient and effective clustering of several variants of the same cited references. With the help of our applications we were able to conduct several analyses.

In another study, we evaluated a research institute in the domain of humanities and social sciences using citation data from Google Scholar (GS). We developed a normalization method (citation impact) following existing bibliometric techniques. To evaluate the expressiveness of normalized citation values based on GS, we computed corresponding values for publications available in WoS and Scopus. Although absolute values differ for the considered data sources, we observed a similar assessment for relative values. For instance, the percentage of citations for one publication w.r.t. all publications for the same conference / journal differed similarly. This indicates that GS is useful for bibliographic evaluations, in particular when publications are not included in WoS or Scopus as it is often the case in the domain of humanities and social sciences.

4. Publications / Theses

Books



- Galhardas, Helena; Rahm, Erhard (eds.): *Data Integration in the Life Sciences*. Springer Lecture Notes in Bioinformatics (LNBI) 8754, 2014
- Galhardas, H.; Rahm, E. (eds): DILS 2014: Poster and demo papers. 2014
- Rahm, E.; Saake, G.; Sattler, K.: *Verteiltes und Paralleles Datenmanagement*. Text book, Springer-Verlag, 2015
- Ritter, N.; Henrich, A.; Lehner, W.; Thor, A.; Friedrich, S.; Wingerath, W.: *Datenbanksysteme für Business, Technologie und Web (BTW 2015) – Workshop proceedings*, 2015

Journals

- Arnold P.; Rahm, E.: *Enriching Ontology Mappings with Semantic Relations. Data and Knowledge Engineering*, 93:1–18, 2014.
- Arnold P.; Rahm, E.: *Automatic Extraction of Semantic Relations from Wikipedia*. International Journal on Artificial Intelligence Tools (IJAIT), 24(2), 2015.
- Bornmann, L.; Thor, A.; Marx, W.; Schier, H.: *The application of bibliometrics to research evaluation in the humanities and social sciences*. Journal of the Association for Information Science and Technology (JAIST), 2015
- Christen, V.; Hartung, M.; Groß, A.: *Region Evolution eXplorer - A tool for discovering evolution trends in ontology regions*. Journal of Biomedical Semantics 2015, 6(26), 2015
- Härder, T.; Rahm, E.: *30 Jahre "Datenbanksysteme für Business, Technologie und Web" - Die BTW im Wandel der Datenbank-Zeiten*. Datenbank-Spektrum, 15(1) Springer, 2015
- Kolb, L.; Sehili, Z.; Rahm, E.: *Iterative Computation of Connected Graph Components with MapReduce*. Datenbank-Spektrum 14 (2), 2014
- Lee, L. H.; Groß, A.; Hartung, M.; Liou, D. M.; Rahm, E.: *A Multi-Part Matching Strategy for Mapping LOINC with Laboratory Terminologies*. Journal of the American Medical Informatics Association (JAMIA), 21(5), 792-800, 2014
- Nentwig, M.; Hartung, M.; Ngonga Ngomo, A.-C.; Rahm, E.: *A Survey of Current Link Discovery Frameworks*. Semantic Web Preprint: 1-18, 2015
- Palma, G.; Vidal, M.-E.; Haag, E.; Raschid, L.; Thor, A.: *Determining similarity of scientific entities in annotation datasets*. Database: The Journal of Biological Databases and Curation, 2015
- Pfeiffer, K.; Peukert, E.: *Integration of Text Mining Taxonomies*. Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2015
- Rahm, E.: *Discovering product counterfeits in online shops: a big data integration challenge*. ACM Journal Data and Information Quality, Vol. 5, 2014
- Raunich, S.; Rahm, E.: *Target-driven Merging of Taxonomies with ATOM*. Information Systems, Vol. 42, 2014
- Scherzinger, S.; Thor, A.: *Cloud-Technologien in der Hochschullehre - Pflicht oder Kür? - Eine Standortbestimmung innerhalb der GI-Fachgruppe DBS*. Datenbank-Spektrum 14(2), 2014
- Thor, A.; Scherzinger, S.; Specht, G.: *Editorial*. Datenbank-Spektrum 14(2), 2014

Proceedings

- Arnold P.; Rahm, E.: *Extracting Semantic Concept Relations from Wikipedia*. Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantic (WIMS), 2014.
- Arnold P.; Rahm, E.: *SemRep: A Repository for Semantic Mapping*. Proc. of 16. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), 2015.
- Christen, V.: *Annotation und Management heterogener medizinischer Studienformulare*. Proc. 27th GI-Workshop Grundlagen von Datenbanken (GvDB), 2015
- Christen, V.; Groß, A.; Hartung, M.: *REX - a tool for discovering evolution trends in ontology regions*. Proc. 10th Intl. Conference on Data Integration in the Life Sciences (DILS), 2014
- Christen, V; Groß, A.; Varghese, J.; Dugas, M.; Rahm, E.: *Annotating Medical Forms using UMLS*. Proc. 11th Intl. Conference on Data Integration in the Life Sciences (DILS), 2015

- Dorok, S., König-Ries, B., Lange, M., Rahm, E.; Saake, G.; Seeger, B.: Joint workshop on data management for science (DMS). Proc. BTW Workshops 2015
- Fisher, J.; Christen, P.; Wang, Q.; Rahm, E.: *A Clustering-Based Framework to Control Block Sizes for Entity Resolution*. Proc. 21st ACM SIGKDD Conf. on Knowledge Discovery and Mining (KDD), 279-288, 2015
- Markl, V.; Rahm, E.; Lehner, W.; Beigl, M.; Seidl,T.: *Big Data-Zentren: Vorstellung und Panel*. Proc. of 16. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), 2015
- Nentwig, M.; Soru, T.; Ngonga Ngomo, A.-C.; Rahm, E.: *LinkLion: A Link Repository for the Web of Data*. The Semantic Web: ESWC 2014 Satellite Events, 2014
- Palma, G.; Vidal, M.E.; Raschid, L.; Thor, A.: *Exploiting Semantics from Ontologies and Shared Annotations to Partition Linked Data*. Proc. 10th Intl. Conf. on Data Integration in the Life Sciences (DILS), 2014
- Petermann, A.; Junghanns, M.; Müller, R.; Rahm, E.: *BIIIG : Enabling Business Intelligence with Integrated Instance Graphs*. Proc. 5th International Workshop on Graph Data Management (GDM 2014), 2014
- Petermann, A.; Junghanns, M.; Müller, R.; Rahm, E.: *FoodBroker - Generating Synthetic Datasets for Graph-Based Business Analytics*. Proc. 5th Workshop on Big Data Benchmarking (WBDB 2014), LNCS 8991, 2015
- Petermann, A.; Junghanns, M.; Müller, R.; Rahm, E.: *Graph-based Data Integration and Business Intelligence with BIIIG*. Proc. of the VLDB Endowment, 7(13), 1577-1580, 2014
- Peukert, E.; Wartner, C.; Rahm, E.: *Smart Link Infrastructure for Integrating and Analyzing Process Data*. Proc. 16th GI conf. on Database systems for Business, Technology and Web (BTW), 2015
- Rahm, E.: *Scalable graph analytics with GRADOOP*. Proc. of the 27th GI-Workshop Grundlagen von Datenbanksystemen (GvDB), 2015
- Rahm, E.; Nagel, W. E.: *ScaDS Dresden/Leipzig: Ein serviceorientiertes Kompetenzzentrum für Big Data*. Proc. GI-Jahrestagung, LNI, 2014
- Scherzinger, S.; Thor, A.: *AutoShard - A Java Object Mapper (not only) for Hot Spot Data Objects in NoSQL Data Stores*. Proc. 16th GI conf. on Database systems for Business, Technology and Web (BTW), 2015
- Scherzinger, S.; Thor, A.: *AutoShard - Declaratively Managing Hot Spot Data Objects in NoSQL Data Stores*. Proc. 17th Int. Workshop on the Web and Databases (WebDB@SIGMOD)
- Sehili, Z.; Kolb, L.; Borgs, C.; Schnell, R.; Rahm, E.: *Privacy Preserving Record Linkage with PPJoin*. Proc. 16th GI conf. on Database systems for Business, Technology and Web (BTW), 2015

Technical Reports and Posters

- Junghanns, M.; Petermann, A.; Gomez, K.; Rahm, E.: *GRADOOP: Scalable Graph Data Management and Analytics with Hadoop*. Techn. Report, Univ. Leipzig, 2015
- ScaDS Dresden/Leipzig Posters (eds: W. Nagel, E. Rahm):
    - Cebit, Hannover, March 2015
    - BMWI-Smart Data Day, Berlin, April 2015
    - Fachtagung Big Data, Berlin, May 2015

PhD Theses

- Anika Groß: *Evolution von ontologiebasierten Mappings in den Lebenswissenschaften*, Univ. Leipzig, 2014
- Hanna Köpcke: *Object Matching on Real-world Problems*, Univ. Leipzig, 2014
- Lars Kolb: *Effiziente MapReduce-Parallelisierung von Entity Resolution-Workflows*, Univ. Leipzig, 2014
- Patrick Arnold: *Semantic Enrichment of Ontology Mappings*, Univ. Leipzig, 2015


PhD defenses of Anika Groß and Hanna Köpcke in 2014


PhD defenses of Lars Kolb (2014) and Patrick Arnold (2015)

Bachelor and Master Theses

1. Bulka, S.: *OLAP-basierte Fallstatistikanalysen von epidemiologischen Daten am LIFE-Forschungsinstitut.* MSc, Univ. Leipzig, 2015
2. Christen, V.: Verfahren für die Reparatur von Ontologie-Mappings in den Lebenswissenschaften. MSc, Univ. Leipzig, 2014
3. Chyhir, A.: Evaluierung von Blocking-Verfahren zum effizienten Matching großer Ontologien in den Lebenswissenschaften. MSc, Univ. Leipzig, 2014
4. Förster, K.: *Eignung von Workflow-Management-Tools für BigData- Aufgabenstellungen.* BSc, 2015
5. Fröschke, T.: Aufbau eines Online-Dienstes zur webbasierten Versionierung von Ontologien. MSc, Univ. Leipzig, 2014
6. Gassner, M.: Adaptierung von Ontologie-Mappings in den Lebenswissenschaften. MSc, Univ. Leipzig, 2014
7. Junghanns, M.: Untersuchung zur Eignung von Graphdatenbanksystemen für die Analyse von Informationsnetzwerken. MSc, Univ. Leipzig, 2014
8. Koch, H.-H.: Evaluation of Backends for the Use in a Horizontally Scalable Version of ipoque's Net Reporter. MSc, Univ. Leipzig, 2014
9. Merzdorf, A.: *Formulierung und Umsetzung einer generischen Schnittstelle zur Extraktion von Daten aus Datenbanken im Rahmen von Softwaretests.* MSc, Univ. Leipzig, 2014
10. Möller, M.: Connecting GOMMA with STROMA: Semantic Ontology Mapping in the Biomedical Domain. BSc, Univ. Leipzig, 2015
11. Preßler, B.: *Erstellung eines Data-Warehouses zur Erfassung von Online-Angeboten zur Unterstützung der Aufdeckung von Produktfälschungen.* BSc, Univ. Leipzig, 2014
12. Swoboda, O.: Realisierung des COnto-Diff Algorithmus innerhalb eines Protégé-Plugins. BSc, 2015
13. Trabandt, J.: Matching von Medikamenten im Rahmen von LIFE. MSc , Univ. Leipzig, 2015

5. Talks

- Arnold, P.: Extracting Semantic Concept Relations from Wikipedia. The 4th International Conference on Web Intelligence, Mining and Semantics (WIMS), Thessaloniki, 2014
- Arnold, P.: SemRep: A Repository for Semantic Mapping. 16. Fachtagung „Datenbankensysteme für Business, Technologie und Web" (BTW), Hamburg, 2015
- Arnold, P.: Semantic Enrichment of Ontology Mappings. Dissertation defense, Leipzig, 2015
- Christen, V.: Annotating Medical Forms using UMLS. 11th Intl. Conf on Data Integration in the Life Sciences (DILS), Los Angeles (USA), 2015
- Christen, V.: Annotation und Management heterogener medizinischer Studienformulare. 27th GI-Workshop Grundlagen von Datenbanken, Magdeburg/Gommern (Germany), 2015
- Groß A.: Evolution von ontologiebasierten Mappings in den Lebenswissenschaften. Dissertation defense, Leipzig, 2014
- Groß, A.: Schema- und Ontologiematching in den Lebenswissenschaften. Institute of Medical Informatics, Westfälische Wilhelms-Universität Münster, 2014
- Groß A.: On2Vers - Online Ontology Versioning. Intl. Conf. on Infrastructures and Cooperation in E-Science and E-Humanities, Leipzig, 2014
- Groß A.: REX - A Tool for Discovering Evolution Trends in Ontology Regions. 10th Intl. Conf. on Data Integration in the Life Sciences (DILS), Lisbon, Portugal, 2014
- Groß, A.: Evolution of Ontology-based Mappings. Invited talk at ALIGNED project meeting, Leipzig, 2015
- Junghanns, M.: Gradoop: Scalable Graph Analytics with Apache Flink. Flink Forward, Berlin, 2015
- Junghanns, M.: Big Graph Processing. Big Data Meetup, Big Data User Group Dresden, Dresden, 2015
- Kolb L.: Effiziente MapReduce-Parallelisierung von Entity Resolution-Workflows. Dissertation defense, Leipzig, 2014
- Köpcke H.: Object Matching on Real-world Problems. Dissertation defense, Leipzig, 2014
- Petermann, A.: BIIIG : Enabling Business Intelligence with Integrated Instance Graphs. 5th International Workshop on Graph Data Management (GDM), Chicago (USA), 2014
- Petermann, A.: FoodBroker - Generating Synthetic Datasets for Graph-Based Business Analytics. 5th Workshop on Big Data Benchmarking (WBDB), Potsdam, 2014
- Peukert, E.: Smart Link Infrastructure for Integrating and Analyzing Process Data. Proc. of 16. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), Hamburg, 2015
- Peukert, E.: „New Trends and Concepts in Big Data (Spark/Flink/GraphProcessing). Big Data in Business Workshop, Leipzig, 2015
- Peukert, E.: GRADOOP: Scalable Graph Data Management and Analytics with Hadoop. 9th. Parallel Tools Workshop Dresden, 2015
- Rahm, E.: Large-Scale Ontology Matching. GMDS-Jahrestagung, Göttingen, Sep. 2014
- Rahm, E.: ScaDS Dresden/Leipzig: ein serviceorientiertes Kompetenzzentrum für Big Data, GI-Jahrestagung, Stuttgart, Sep. 2014
- Rahm, E.: Big Data Zentrum ScaDS Dresden/Leipzig. ScaDS Kickoff-Veranstaltung, Dresden Oct. 2014
- Rahm, E.: Semantic ontology mappings: How to determine and use them. Colloquium Talk, Univ. of Paris-Sud, 2014
- Rahm, E.: Big Data Integration at ScaDS Dresden/Leipzig. Informatik-Kolloquium, RWTH Aachen, Dec.2014
- Rahm, E.: Big Data Integration. Australian National University and Data Science Meetup, Canberra, 2015
- Rahm, E.: Big Data Panel. BTW-Tagung, Hamburg, 2015
- Rahm, E.: Scalable Graph Analytics with GRADOOP. Keynote GvDB-Workshop, Gommern, 2015
- Rahm, E.: Data integration research at the Univ. of Leipzig. SAP, Walldorf, Oct. 2015
- Sehili, Z.: Privacy Preserving Record Linkage with PPJoin. 16. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), Hamburg, 2015

Prof. Rahm at the Kick-off event of ScaDS Dresden/Leipzig, TU Dresden, October 2014

A. Petermann at ICDE workshop (Chicago, April 2014); V. Christen at GI-Workshop (Gommern, May 2015)

6. Memberships in Program/Steering Committees, Editorial/Advisory Boards, ...

Rahm, E.:

- Scientific co-coordinator of the Big Data Centre ScaDS Dresden/Leipzig
- Vice speaker of the executive committee of the GI (German Informatics Society) group Databases and Information Systems
- Steering Committee DILS conference series (Data Integration in the Life Sciences)
- PC Co-Chair DILS 2014, 10th Int. Conf. on Data Integration in the Life Sciences, Lisbon, July 2014
- Co-Chair PopInfo 2015, 1st Int. KDD workshop on Population Informatics for Big Data, Sydney, Aug. 2015
- Co-Chair BTW Workshop on Big Data in Science, Hamburg 2015
- Chair ScaDS Workshop Big Data in Business (BIDIB), Leipzig, Nov. 2015
- PC Member of several conferences and workshops (VLDB 2015, ICDE 2015, BTW 2015, DILS 2014, DINA 2015, ISC Conf. on Cloud & BigData 2015)
- Reviewer for different journals and research associations
- Advisory Board Europar conference series
- Board Member IZBI (Interdisciplinary Centre for Bioinformatics, Leipzig)

Groß, A.:

- PC Member: VOILA 2015, Bio-Ontologies 2014
- Journal Reviewer: PLOS ONE, Semantic Web Journal (SWJ), International Journal on Semantic Web and Information Systems (IJSWIS)

Thor, A.:

- Member of the executive committee of the GI group Databases and Information Systems
- Speaker of the GI (German Informatics Society) working group "Data Management in the Cloud "
- Member of the e-learning working group
- Chair: BTW Studierendenprogramm 2015, Workshop on Data Management in the Cloud (DMC) 2014
- Guest editor Datenbankspektrum 2/2014 (Datenmanagement in der Cloud)
- PC Member: VLDB 2015, SCDM 2014, MR.BDI 2014, ESWC 2014
- Journal Reviewer: a.o. ACM Transcations on the Web, Journal of Data and Information Quality, ACM Computing Surveys, IEEE Transactions on Knowledge & Data Engineering, Data & Knowledge Engineering