Joint Workshop on Data Management for Science

Sebastian Dorok[1,5], Birgitta König-Ries[2], Matthias Lange[3],
Erhard Rahm[4], Gunter Saake[5], Bernhard Seeger[6]

[1] Bayer Pharma AG
[2] Friedrich Schiller University Jena
[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben
[4] University of Leipzig
[5] Otto von Guericke University Magdeburg
[6] Philipps University Marburg

Message from the chairs

The *Workshop on Data Management for Science* (DMS) is a joint workshop consisting of the two workshops *Data Management for Life Sciences* (DMforLS) and *Big Data in Science* (BigDS). BigDS focuses on addressing big data challenges in various scientific disciplines. In this context, DMforLS focuses especially on life sciences. In the following, we give short excerpts of the call for papers of both workshops:

Data Management for Life Sciences In life sciences, scientists collect an increasing amount of data that must be stored, integrated, processed, and analyzed efficiently to make effective use of them. Thereby, not only the huge volume of available data raises challenges regarding storage space and analysis throughput, but also data quality issues, incomplete semantic annotation, long term preservation, data access, and compliance issues, such as data provenance, make it hard to handle life science data. To address these challenges, advanced data management techniques and standards are required. Otherwise, the use of life science data will be limited. Thereby, one question is whether general purpose techniques and methods for data management are suitable for life science use cases or specialized solutions tailored to life science applications must be developed.

Big Data in Science The volume and diversity of available data has dramatically increased in almost all scientific disciplines over the last decade, e.g. in meteorology, genomics, complex physics simulations and biological and environmental research. This development is due to great advances in data acquisition (e.g. improvements in remote sensing) and data accessibility. On the one hand, the availability of such data masses leads to a rethinking in scientific disciplines on how to extract useful information and on how to foster research. On the other hand, researchers feel lost in the data masses because appropriate data management tools have been not available so far. However, this is starting to change with the recent development of big data technologies that seem to be not only useful in business, but also offer great opportunities in science.

The joint workshop DMS brings together database researchers with scientists from various disciplines especially life sciences to discuss current findings, challenges, and opportunities of applying data management techniques and methods in data-intensive sciences. The joint workshop is held for the first time in conjunction with the 16th Conference on Database Systems, Technology, and Web (BTW 2015) at the University of Hamburg on March 03, 2015.

The contributions were reviewed by three to four members of the respective program committee. Based on the reviews, we selected eight contributions for presentation at the joint workshop. We assigned each contribution to one of three different sessions covering different main topics.

The first session comprises contributions related to *information retrieval*. The contribution *Ontology-based retrieval of scientific data in LIFE* by Uciteli and Kirsten presents an approach that utilizes ontologies to facilitate query formulation. Colmsee et al. make also use of ontologies, but use them for improving search results. In *Improving search results in life science by recommendations based on semantic information*, they describe and evaluate their approach that uses document similarities based on semantic information. To improve performance of sampling analyses using MapReduce, Schäfer et al. present an incremental approach. In *Sampling with incremental MapReduce*, the authors describe a way to limit data processing to updated data.

In the next session, we consolidate contributions dealing with *data provenance*. In his position paper *METIS in PArADISE*, Heuer examines the importance of data provenance in the evaluation of sensor data, especially in assistance systems. In their contribution *Extracting reproducible simulation studies from model repositories using the COMBINE archive toolkit*, Scharm and Waltemath deal with reproducible simulation studies.

The last session covers the topic *data analysis*. In *Genome sequence analysis with MonetDB: a case study on Ebola virus diversity*, Cijvat et al. present a case study on genome analysis using a relational main-memory database system as platform. In *RightInsight: Open source architecture for data science*, Bulut presents an approach based on Apache Spark to conduct general data analyses. In contrast, Authmann et al. focus on challenges in spatial applications and suggest an architecture to address them in their paper *Rethinking spatial processing in data-intensive science*.

We are deeply grateful to everyone who made this workshop possible – the authors, the reviewers, the BTW team, and all participants.

Program chairs

Data Management for Life Sciences
Gunter Saake (Otto von Guericke University Magdeburg)
Uwe Scholz (IPK Gatersleben)

Big Data in Science
Birgitta König-Ries (Friedrich Schiller University Jena)
Erhard Rahm (University of Leipzig)
Bernhard Seeger (Philipps University Marburg)

Program committee

Data Management for Life Sciences

Sebastian Breß (TU Dortmund)
Sebastian Dorok (Otto von Guericke University Magdeburg)
Mourad Elloumi (University of Tunis El Manar, Tunisia)
Ralf Hofestädt (Bielefeld University)
Andreas Keller (Saarland University, University Hospital)
Jacob Köhler (DOW AgroSciences, USA)
Matthias Lange (IPK Gatersleben)
Horstfried Läpple (Bayer HealthCare AG)
Ulf Leser (Humboldt-Universität zu Berlin)
Wolfgang Müller (HITS GmbH)
Erhard Rahm (University of Leipzig)
Can Türker (ETH Zürich, Switzerland)

Big Data in Science

Alsayed Algergawy (Friedrich Schiller University Jena)
Peter Baumann (Jacobs Universität)
Matthias Bräger (CERN)
Thomas Brinkhoff (FH Oldenburg)
Michael Diepenbroeck (Alfred-Wegner-Institut)
Christoph Freytag (Humboldt Universität)
Michael Gertz (Uni Heidelberg)
Frank-Oliver Glöckner (MPI für Marine Mikrobiologie)
Anton Güntsch (Botanischer Garten und Botanisches Museum, Berlin-Dahlem)
Thomas Heinis (Imperial College, London)
Thomas Hickler (Senckenberg)
Jens Kattge (MPI für Biogeochemie)
Alfons Kemper (TU München)
Meike Klettke (Uni Rostock)
Alex Markowetz (Uni Bonn)
Thomas Nauss (Uni Marburg)
Jens Nieschulze (Forschungsreferat für Datenmanagement der Uni Göttingen)
Kai-Uwe Sattler (TU Ilmenau)
Stefanie Scherzinger (OTH Regensburg)
Myra Spiliopoulou (Uni Magdeburg)
Uta Störl (HS Darmstadt)