Next Generation Data Integration (for the Life Sciences)

[Abstract]

Ulf Leser
Humboldt-Universität zu Berlin
Institute for Computer Science
leser@informatik.hu-berlin.de

ABSTRACT

Ever since the advent of high-throughput biology (e.g., the Human Genome Project), integrating the large number of diverse biological data sets has been considered as one of the most important tasks for advancement in the biological sciences. The life sciences also served as a blueprint for complex integration tasks in the CS community, due to the availability of a large number of highly heterogeneous sources and the urgent integration needs. Whereas the early days of research in this area were dominated by virtual integration, the currently most successful architecture uses materialization. Systems are built using ad-hoc techniques and a large amount of scripting. However, recent years have seen a shift in the understanding of what a "data integration system" actually should do, revitalizing research in this direction. In this tutorial, we review the past and current state of data integration (exemplified by the Life Sciences) and discuss recent trends in detail, which all pose challenges for the database community.

About the Author

Ulf Leser obtained a Diploma in Computer Science at the Technische Universität München in 1995. He then worked as database developer at the Max-Planck-Institute for Molecular Genetics before starting his PhD with the Graduate School for "Distributed Information Systems" in Berlin. Since 2002 he is a professor for Knowledge Management in Bioinformatics at Humboldt-Universität zu Berlin.

In: G. Specht, H. Gamper, F. Klan (eds.): Proceedings of the 26^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 21.10.2014 - 24.10.2014, Bozen, Italy, published at http://ceur-ws.org.