

Open Source Datenintegration für die Krebsforschung

Berlin Open 09 - Berlin, Deutschland

Bernhard Pfeifer, Michael Netzer

`{bernhard.pfeifer|michael.netzer}@umit.at`

Institute of Biomedical Engineering / UMIT

Juni 22-23, 2009

Table of Contents

Motivation

Principal Configuration

Research Groups & Data Provider
System Configuration

DWH: Back & Front Room

Back Room

Front Room

Turning data into information

Discussion & Conclusion

Acknowledgements

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room
Turning data into
information

Discussion &
Conclusion

Acknowledgements

Motivation



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room
Turning data into
information

Discussion &
Conclusion

Acknowledgements

Complex high-dimensional systems represent an important area of interdisciplinary research in biomedical analysis and systems biology. To afford this, a dedicated, platform independent system is required.

- ▶ Biomedical research describes an organism to be an integrated and interacting network of genes, proteins and reactions.
- ▶ To get a deeper understanding in diseases, for discovering potential biomarker candidates, for advancing biomedical analysis, a platform with all relevant information integrated is needed.
 - ▶ *clinical data, literature data for verification and validation*
 - ▶ *highthrouput data (-omics data) and external biosources*
 - ▶ *semantic knowledge and ontologies*

Research Groups & Data Provider

The prostate cancer project research groups are:

1. Department of Urology (Innsbruck, AUT): *phenomic data, patient samples*
2. Biocrates Life Science GmbH (Innsbruck, AUT): *metabolomics*
3. Institute of Analytical Chemistry and Radiochemistry (Innsbruck, AUT): *proteomics*
4. German Cancer Research Centre (Heidelberg, GER): *genomics*
5. Max Planck Institute of Molecular Genomics (Berlin, GER): *systems biology*
6. University for Health Sciences, Medical Informatics and Technology (Hall i.T., AUT): *data warehouse, infrastructure*

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

[Research Groups &
Data Provider](#)

[System Configuration](#)

[DWH: Back &
Front Room](#)

[Back Room](#)

[Front Room](#)

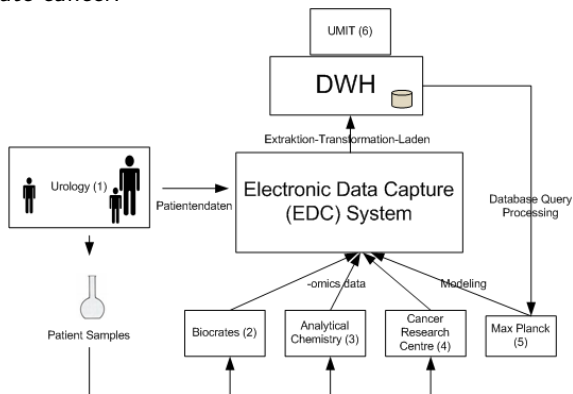
[Turning data into
information](#)

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Project configuration

The project focuses on the integration of high-throughput technologies to identify molecular signatures allowing the stratification of patients who are susceptible to curative treatment of prostate cancer.



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

Research Groups &
Data Provider

[System Configuration](#)

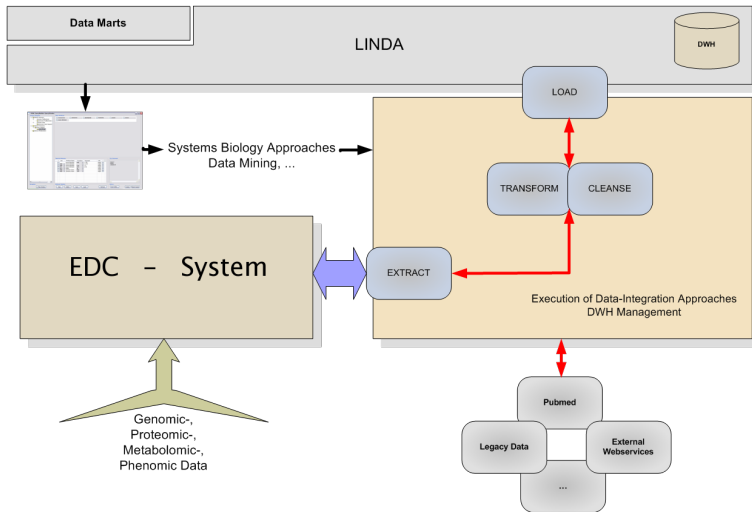
[DWH: Back &
Front Room](#)

Back Room
Front Room
Turning data into
information

[Discussion &
Conclusion](#)

[Acknowledgements](#)

DWH configuration



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

[Research Groups &
Data Provider](#)

[System Configuration](#)

[DWH: Back &
Front Room](#)

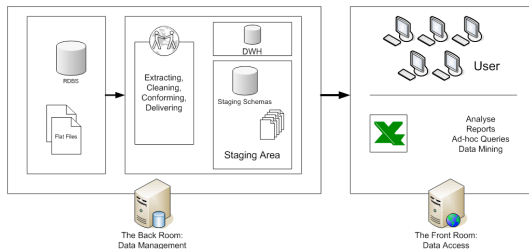
[Back Room](#)
[Front Room](#)
[Turning data into
information](#)

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Back & Front Room

It is common practice to separate between back room and front room entities when talking about data warehouses. These two parts are in most cases separated physically and logically. While the back room is holding and managing the data, the front room enables data access methods.



- ▶ In the back-room four different steps have to be performed: *extraction, cleaning, conforming, delivering*
- ▶ The schema used for building the DWH is a bio-star schema
- ▶ In the front-room the KD³ toolbox with integrated ad-hoc query builder and self developed as well as integrated *Functional Objects* for data processing (statistical framework-JavaStat and data mining toolbox Weka is located).

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

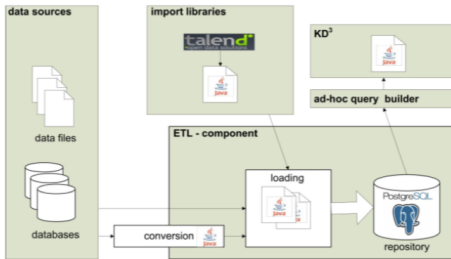
Back Room
Front Room
Turning data into
information

Discussion &
Conclusion

Acknowledgements

Back Room

The back room is often described as data management or data preparation component. It contains the data, prepares and delivers data retrieved by queries, but it does not support any user queries from the outside since this is a task of the front room. A back room, in this context, may be regarded as permanently storing the information to a physical entity.



The RDBM System used is the Open Source product PostgreSQL.

The data integration framework is the Open Source product Talend Open Studio.

The servers are HP Proliant DL380 running Suse Linux operating system.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room

Front Room

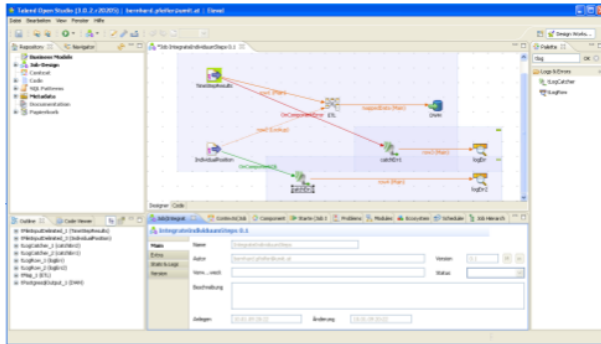
Turning data into
information

Discussion &
Conclusion

Acknowledgements

Back Room

With the Talend designer the integration tasks are designed graphically. The workflow and the data sources are assembled together and parametrized using the Talend Open Studio GUI components.



Special situations and errors are treated via the integrated exception handler. In case of failure an automatic message is send to the back room administrator and a rollback is performed in the underlying database system.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room

Front Room

Turning data into
information

Discussion &
Conclusion

Acknowledgements

Front Room

- ▶ The front room component enables a user or client application to access the data held in the warehouse.
- ▶ The main task of the front room is mapping the large amount of low-level data, usually stored in a data warehouse, to another more valuable form.
- ▶ The front room manages the queries performed at the outside and schedules and plans them, to achieve the results defined for performance issues.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room

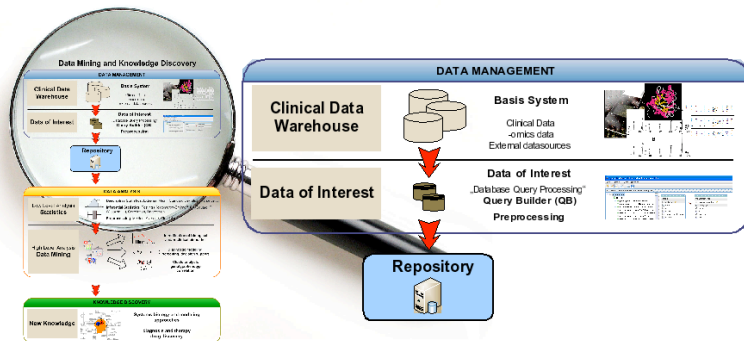
Front Room

Turning data into
information

Discussion &
Conclusion

Acknowledgements

Turning data into information



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room

Turning data into
information

Discussion &
Conclusion

Acknowledgements

Extraction of the data of interest

The Ad-Hoc Query Builder (QB) is the central component in the data management module.

- ▶ Registration of any JDBC driver and data sources
- ▶ Delete and edit data sources
- ▶ Visual design of complex queries
- ▶ Automatic detection of relations between the data objects
- ▶ dialog driven creation of conditional (WHERE) clauses
- ▶ Load / save of generated queries
- ▶ Presentation of resultsets in tabular form and export possibility in different formats (CSV, TXT, ARFF, INSERT, Repository, ...)
- ▶ Design of abstract queries, based on metadata
- ▶ possibility of generation and adaption of abstract layers to hide the complete data warehouse schema from the front door users
- ▶ fully integration of the QB as external Functional Object into the KD³ designer

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

Research Groups &
Data Provider
System Configuration

[DWH: Back &
Front Room](#)

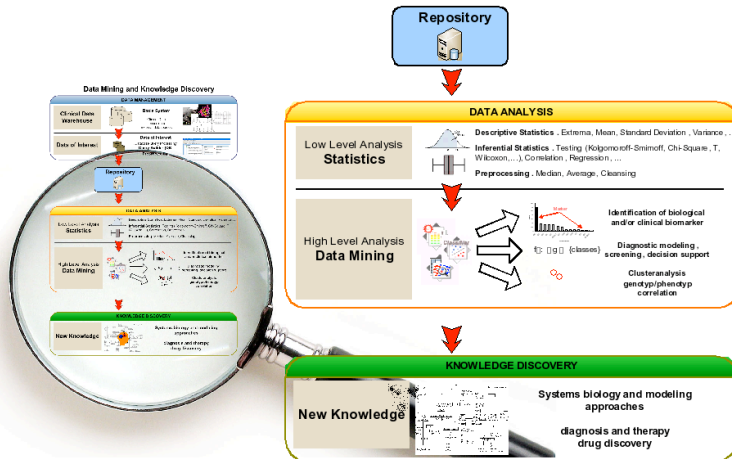
Back Room
Front Room

[Turning data into
information](#)

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Turning data into information



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room

Turning data into
information

Discussion &
Conclusion

Acknowledgements

KD³ - Knowledge Discovery in Databases Designer

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

The complete software package is written using the programming language Java.

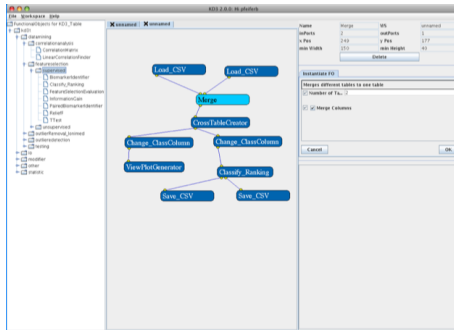


Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room

Turning data into
information

Discussion &
Conclusion

Acknowledgements

External components can be integrated as Functional Objects if the defined interface is implemented using an adaptor class. The classes are loaded into the system using the reflection API. The parametrizing component is designed using annotations.

Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases".

- ▶ Creation of a dynamic mining-workflow by assembling Functional Objects
- ▶ Parametrizing of these objects and enabling user defined inputs
- ▶ Execution of the workflow (and parallel execution of autonomous Functional Objects)

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

Research Groups &
Data Provider
System Configuration

[DWH: Back &
Front Room](#)

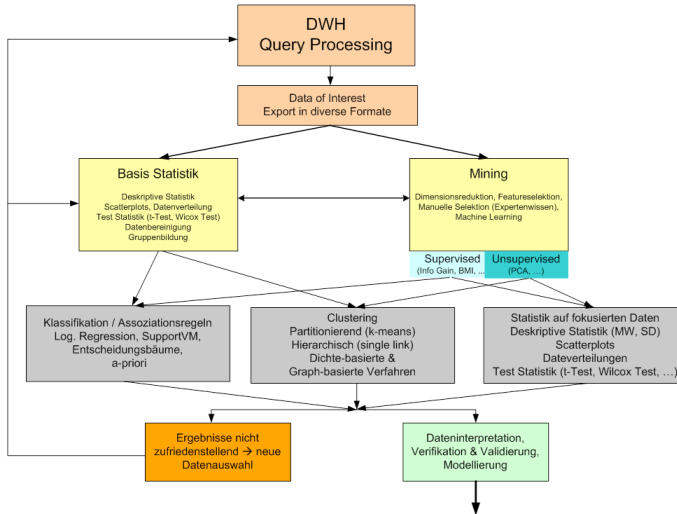
Back Room
Front Room

[Turning data into
information](#)

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Data Mining & Statistics



Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

[Research Groups &
Data Provider
System Configuration](#)

[DWH: Back &
Front Room](#)

[Back Room
Front Room](#)

[Turning data into
information](#)

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Discussion & Conclusion

- ▶ The integrative system enables project collaborators to process steps in a (pre)defined workflow.
- ▶ The integrative, adaptive and extensible system focuses on the specific requirements existing in biomedical research projects.
- ▶ The quality assured and integrated data can be accessed with the ad-hoc query builder tool. Neither internal schema representation nor SQL knowledge is needed in order to use the query tool.
- ▶ The system focuses towards bridging the clinical and the molecularbiological domain in order to offer a more patient individual medicine in the face of an integrated treatment of genotype and phenotype.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

[Table of Contents](#)

[Motivation](#)

[Principal
Configuration](#)

Research Groups &
Data Provider
System Configuration

[DWH: Back &
Front Room](#)

Back Room
Front Room
Turning data into
information

[Discussion &
Conclusion](#)

[Acknowledgements](#)

Discussion & Conclusion

- ▶ Talend Open Studio helps to enable all data integration tasks in a quick, easy and accurate way.
- ▶ The generated code (from Talend Open Studio) can be directly used in our KD³ tasks (cleaning and/or preparation tasks for further processing).
- ▶ Due to the fact that Talend Open Studio is Open Source new components can be developed and integrated in order to develop a biomarker cancer research platform.
- ▶ System is based on several Open Source Tools like, Talend Open Studio, PostgreSQL, Linux, Eclipse, Java
 - ▶ The development of such a platform would have been impossible if only using Closed Source Software, because the costs of the infrastructure would exceed the funding of such academic projects.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room
Turning data into
information

Discussion &
Conclusion

Acknowledgements

Acknowledgements

This work is part of the Project "Systems Biology of Prostate Cancer" and was supported by the National Foundation for Research and Development.

UMIT Bernhard Pfeifer, Michael Netzer, Melanie Osl,
Michael Seger, Christian Baumgartner

ARC Robert Modre, Günter Schreier



A special thanks to the Talend Team for their assistance. Furthermore, we want to thank Roland Kienast, Thomas Schwarzmayr, Andreas Dander, Michael Handler and Leonhard Helminger for their efforts in developing the entire system.

Open Source
Datenintegration
für die
Krebsforschung

Bernhard Pfeifer,
Michael Netzer

Table of Contents

Motivation

Principal
Configuration

Research Groups &
Data Provider
System Configuration

DWH: Back &
Front Room

Back Room
Front Room
Turning data into
information

Discussion &
Conclusion

Acknowledgements