UNIKASSEL
VERSITÄT

# Database Systems / Data Transfer (WP2)

Joint Workshop „Organic Dairy Health“ and „2-Org-Cows“, 22th-23th of Feb. 2016

Boris Kulig

AGRARTECHNIK
WITZENHAUSEN

# Database vs. Excel / flat files

UNIKASSEL VERSITÄT

- Why should we use a database?
  - Because it is highly likely that we have a large amount of data.
  - A lot of people want to work with the data.
  - For every experimental aim which comes up we need a recombination of data sets and variables.
  - There is a need for easily sharing data between project partners, but we must maintain the ownership and the right of the data sets.
  - The consistency of the data must be ensured.

  - That means: We would have an incalculable and not fool proof manageable number of Excel files in different versions on different computers in different countries.

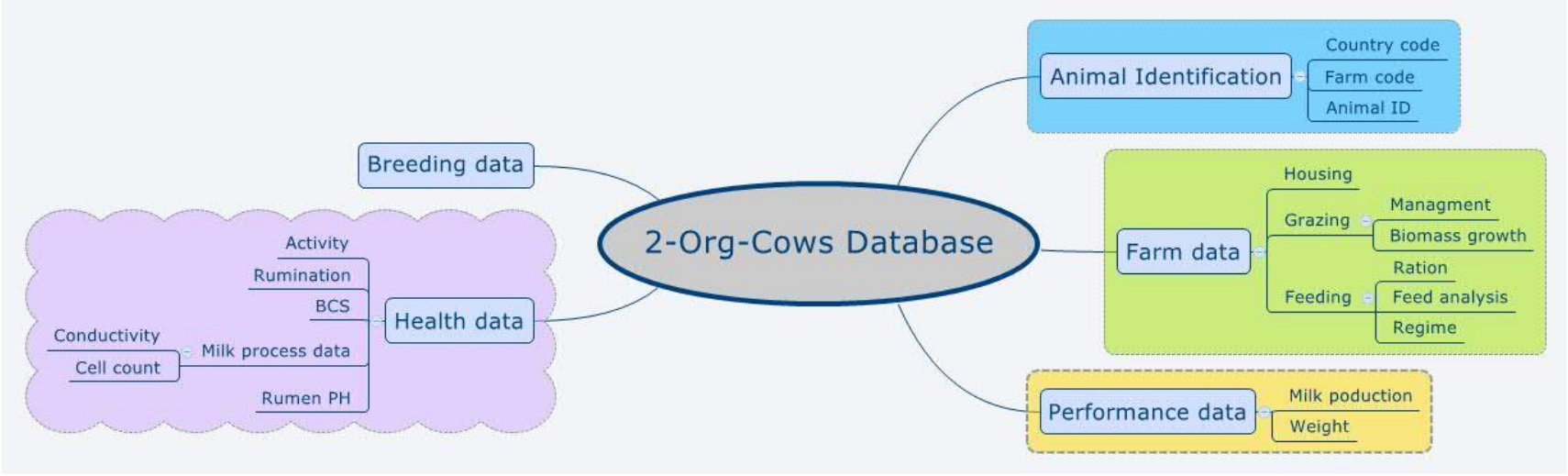
  - In other words: This, we can not efficiently manage with Excel!

# Problem? Yes, we need a database design!

UNIKASSEL VERSITÄT

- There are a lot of issues to be clarified that are vital for the database design, e.g.:
  - Identification of for the project relevant characteristics / traits (environmental, genetic, health, …) is working in progress!
  - What are the entities in our data structure? Animal, farm, field on farm, stable on fram, …
  - How detailed should we handle pedigree data?
  - What is the frequency data / parameters are recorded with?
  - What are the aims in the ongoing project?
  - Will there come up new variables and aims form time to time?
  - …
- Normally this situation is a death blow to a proper database development!

# Yet identified entities were data come from



# Duty and Thinkable Solution

UNIKASSEL VERSITÄT

- **Duty!**
  - A database that:
    - Is maximum flexible against changes in structure of master data
    - Can handle any type and number of parameters / attributes / characteristics
    - Has no problem with different recording frequencies
    - Makes data aggregation and combination easy (OLAP)
    - Has a user and rights management
  - A Database that is easy to use when:
    - Importing data
    - Query or exporting data
  - A Database which comes up maybe with a set of standard data analysis routines
- **Thinkable Solution**
  1. Star scheme database structure
  2. Web interface for importing and query data and optional for standard analyses

# 1. Database -> Why Star Scheme?

UNIKASSEL VERSITÄT

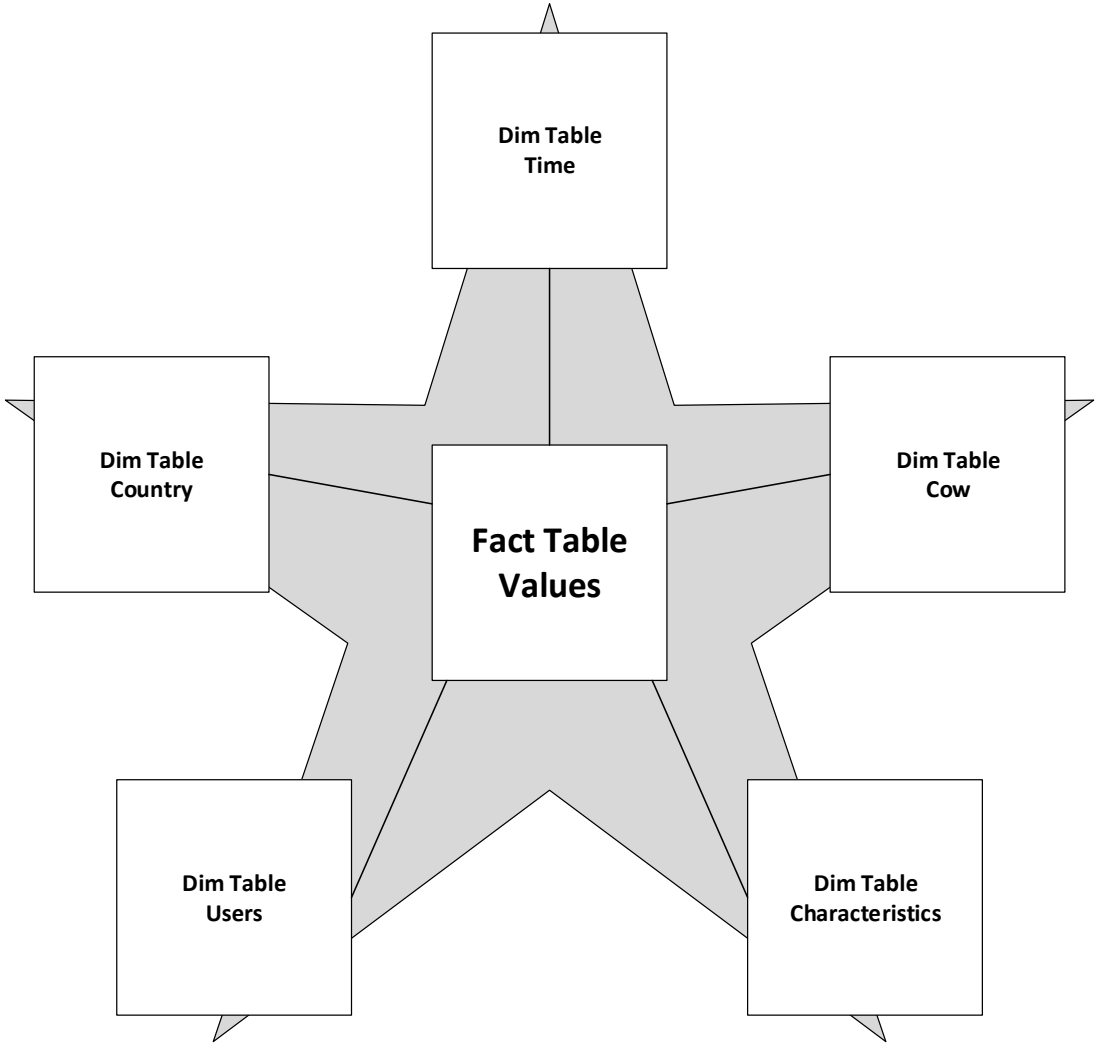
- It is the solution for data warehouse system for example at Amazon and Google.
- They use that, because it is simple, fast and flexible against structure changes.
- Easy data aggregation and OLAP is possible
- Easy manageable table structure with only two hierarchy levels
- Best for setting up web interfaces

# What is a Star Scheme Database about?

UNIKASSEL VERSITÄT

- It gets its name from the pictogram of a star.
  - It consists of one or more **fact table** at its center and **dimension tables** surrounding the center as star beams.
    - A **dimension table** holds a small number of records (tuple) but a large number of attributes to describe the fact data. Dimension tables are the container for master data. Every dimension table has a primary key as an unique record identifier.
    - The **fact table** hold a large number of records but a small number of attributes. Most attributes are the foreign keys which are exported through 1:n relations to the dimension tables. All foreign keys are formed to a primary key as a globally unique identifier of a certain record.
  - Only one column in the fact table is necessary for recording any type of variable measure e.g. milk yield or a claw health index.

# What is a Star Scheme Database about?

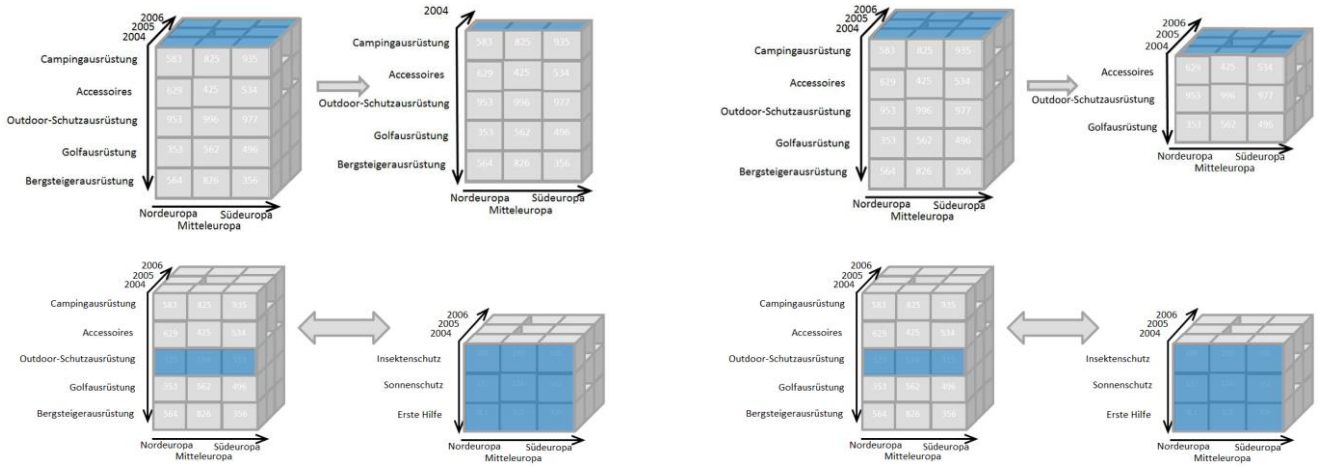



Database Design

# OLAP – Online Analysis Process

UNIKASSEL VERSITÄT

- OLAP – An Integral Part of Star Scheme Databases!
  - OLAP = online analytical processing
  - Flexible technique of data aggregation and analysis
  - Slicing, dicing, drill-up and down and pivoting data




  - With a star scheme join selection and OLAP you get easily aggregated data for deep analysis in statistical software.

# 2. Elements of the Web Interface

UNIKASSEL VERSITÄT

- Login Screen *(basic function)*
  - user and rights management (database driven)
- Data import *(basic function)*
  - Standard import filters
    - on the one hand, scheduled automatically by a program
    - and on the other, triggered by user
  - Free format text file import dialog
- Standard online analysis reports *(optional)*
- Query data for offline data analysis *(basic function)*
- Interface to maintain master data *(optional)*

But we are only at the design stage!

# Standard Import Filters

UNIKASSEL VERSITÄT

- So far identified standard file formats maybe standard for all partners:
  - SensOor Datasets -> Import triggered by user
  - NEDAP Pedometer -> Import triggered by user
  - SMAXTEC pH & Temp -> Import automatically by a program
  - Rumi Watch -> Import triggered by user
- Special data sources in Germany
  - Uniform Herd Management Software (used at DFH)
  - Netrind from VIT and HVL (Web portal for milk recording and herd / cow health data)
  - Genetic values and pedigree data form VIT in Verden
- **What more?**

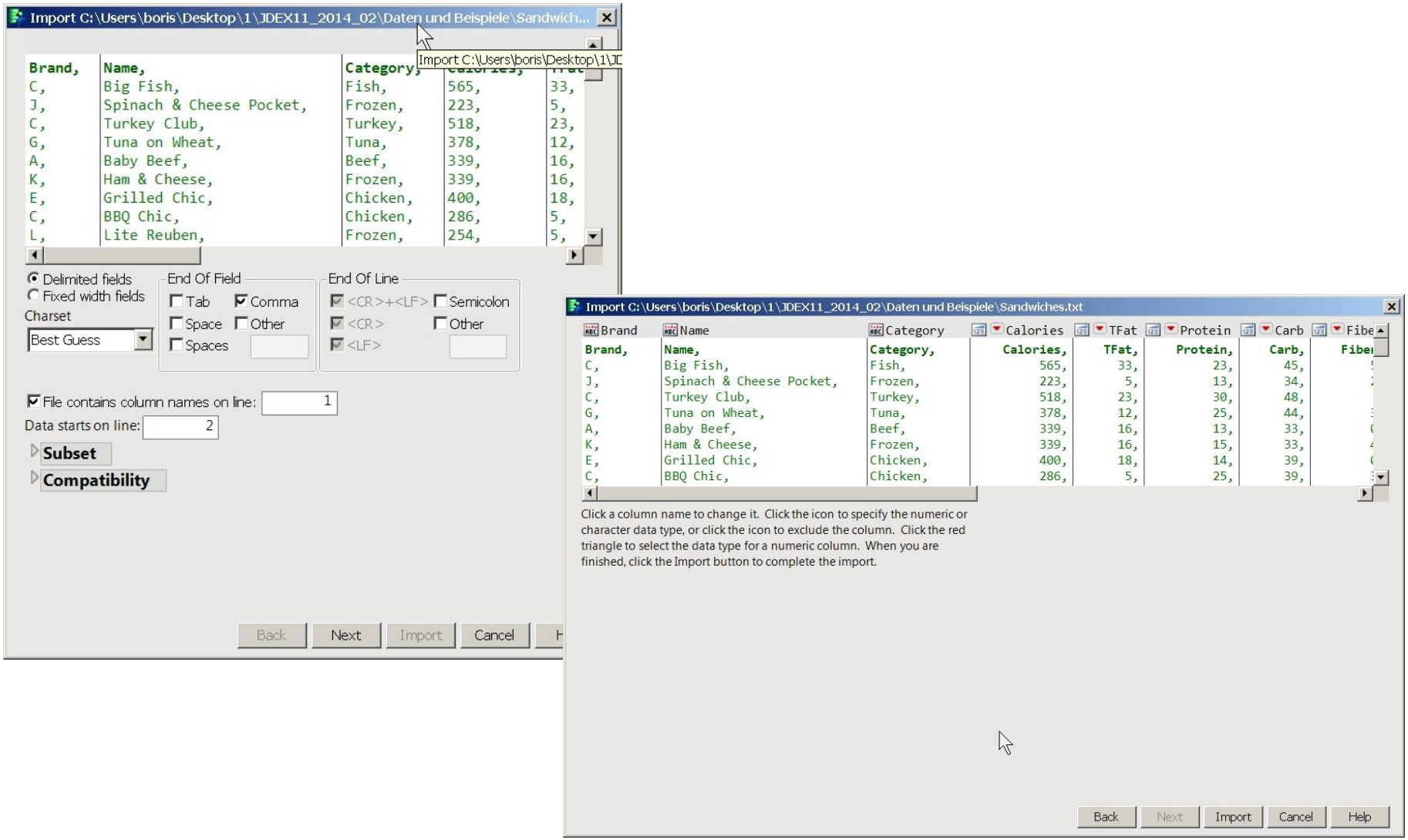
# Free format file import dialog

UNIKASSEL VERSITÄT

- Text files only (csv, txt, …)!
  - A four steps approach:
    1. Login, thus the database can set user rights and pemit or deny certain transactions
    2. Upload data file
    3. Definition of file format, if it is not a standard format (see and ff.)
    4. Mapping of variables in file to attributes in database

# Free format file import dialog

UNIKASSEL VERSITÄT



# Requirements for a valid text file

UNIKASSEL VERSITÄT

- Definition of the minimum requirements for a valid text files
  - Atomic values -> only one information in one column

ok:

| ID Cow | Name Cow | Date | Milk yield | Unit |
|---|---|---|---|---|
| DE 06 640 73373 | Gertrud | 2016-02-19 08:00:00 | 10 | l/day |

not ok:

| ID Cow | Date | Milk yield |
|---|---|---|
| DE 06 640 73373 Gertrud | 2016-02-19 08:00:00 | 10 l/day |

  - Record measurement replicates in separate lines, because otherwise you have to match a database attribute to more than one column it the import file

ok:

| ID Cow | Name Cow | Date | Replicate | Milk yield | Unit |
|---|---|---|---|---|---|
| DE 06 640 73373 | Gertrud | 2016-02-19 08:00:00 | 1 | 10 | l/day |
| DE 06 640 73373 | Gertrud | 2016-02-19 08:00:00 | 2 | 11 | l/day |
| DE 06 640 73373 | Gertrud | 2016-02-19 08:00:00 | 3 | 11.5 | l/day |

not ok:

| ID Cow | Name Cow | Date | Milk yield 1 | Milk yield 2 | Milk yield 3 | Unit |
|---|---|---|---|---|---|---|
| DE 06 640 73373 | Gertrud | 2016-02-19 08:00:00 | 10 | 11 | 11.5 | l/day |

# Requirements for a valid text file

UNIKASSEL VERSITÄT

- Definitions *(continued)*:
  - The safest text format is CSV
    - Character set is Unicode (UTF-8), not ASCII, any proprietary Win/Mac or country specific charset
    - Decimal separator is a dot not a comma!
    - Column separator could be a comma or a semicolon, tab stops and blanks are not as save
  - Time format:
    - Date and time in one column or date and time in two columns
    - Please use the 24 hours format not AM / PM indicators

| Date and Time | | Date | Time |
|---|---|---|---|
| YYYY-MM-DD hh:mm:ss | | YYYY-MM-DD | hh:mm:ss |
| 2016-02-19 08:00:00 | | 2016-02-19 | 08:00:00 |

# Requirements for a valid text file

UNIKASSEL VERSITÄT

- Definitions *(continued)*:
  - The safest text format is CSV
    - Prevent offsets in files, good style is to have the variable definition in the first row
    - The records should start in row two
    - Double quoting of strings is not mandatory, but a good habit
- A good import text file look like this:

```
"ID Cow","Name Cow","Date","Replicate","Milk yield","Unit"
"DE 06 640 73373","Gertrud",2016-02-19 08:00:00,1,10,"l/day"
"DE 06 640 73373","Gertrud",2016-02-19 08:00:00,2,11,"l/day"
"DE 06 640 73373","Gertrud",2016-02-19 08:00:00,3,11.5,"l/day"
```

# Requirements for a valid text file

UNIKASSEL VERSITÄT

- How can I create a CSV file?
  - For instance use the “Save As” dialog in Excel!

# Requirements for a valid record

UNIKASSEL VERSITÄT

- At least we need:
  - Cow's ear tag number
  - The measured values
  - The time stamp of the measurment

Database Design

# Open Questions

UNIKASSEL VERSITÄT

- We still need to clarify the database design.
- We must set up the master data / dimension tables.
- For that we need a list of:
  - Yet identified relevant characteristics / variables / attributes
  - Users, partners, groups for the rights management
  - More lists?
- We need a agreement about rights and data ownership.
- What other standard data import filters?
- Ideas about standard analysis reports?
- What more?

Database Design

# Links

UNIKASSEL VERSITÄT

- https://en.wikipedia.org/wiki/Relational_database
- https://en.wikipedia.org/wiki/SQL
- https://en.wikipedia.org/wiki/Star_schema
- https://en.wikipedia.org/wiki/OLAP_cube
- https://en.wikipedia.org/wiki/Comma-separated_values

AGRARTECHNIK
WITZENHAUSEN