Tameus

Project meeting
Friday December 16, 2011, 10:00 - 17:00

Minutes

Participants: Böhlen, Boltshauser, Bracher, Cafagna, Taliun

Agenda

- 11:00 - 12:00 status of the feed database;
- 12:00 - 12:30 SNF Project and action points;
- 12:30 - 14:00 lunch;
- 14:00 - 15:00 workshop;
- 15:00 - 16:00 fundings, environmental modeling, mobile application.

The System

- The Goals:
  - detailed data and research challenges:
    highly relevant for UZH;
  - the product that is demanded by the feed industry:
    highly relevant for the Agroscope?.
- The Outline:
  - data;
  - user interface;
  - maintenance and development.

The System: The Data

- We had imported over 2.4 millions of nutrient measurements:
  - collected from 10^5 of feed samples;
  - 2% of measurements are from Agridea and 92% are from Lims;
- The database contains over 900 nutrients:
  - 41% of nutrients are classified;
- There are more than 1100 feeds in the database:
  - 56% of feeds are classified.
- TODO's:
  - queries are expensive and especially for computation of derived nutrients that can take minutes (UZH);
  - for more than 400 feeds assign *groupid* in the *dfeed* table (Agroscope);
  - assign $2 \cdot 10^5$ samples to correct feed categories by changing *idfeedfkey* value in *facttable* table (Agroscope).

The System: The Data

- Geographical information for individual feed samples:
  - location is specified with geographical coordinates and region names;
  - altitude is specified with a range, i.e., min and max;
- Properties of the individual samples:
  - a sample can have up to 14 biological properties;
  - a sample can have up to 7 technical properties;
- TODO's:
  - currently, this information is available only for the Agridea data;
  - import this information for the Lims data (Agroscope):
    - more than 80 columns in 2 tables must be modified for about $2 \cdot 10^5$ records.

The System: The User Interface

- **Compact** and, at the same time, **rich search interface**:
  - the search options are organized into **6 categories**:

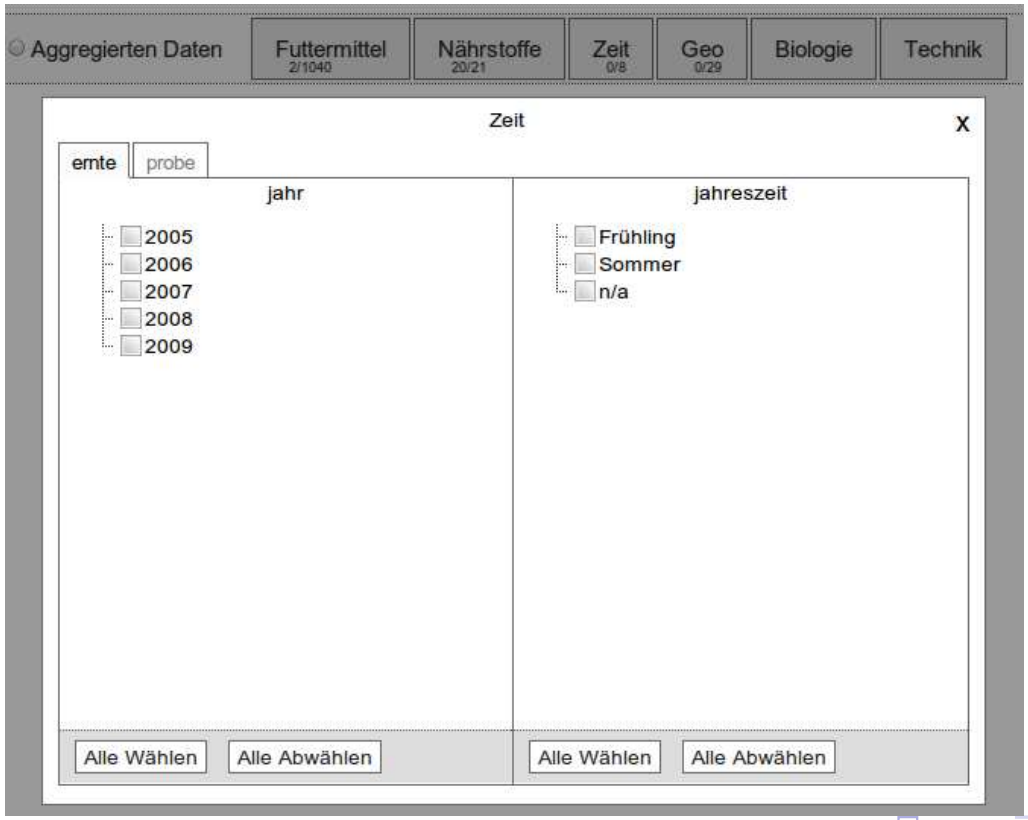
    i.e., feeds, nutrients, geo, time, bio and technical categories;
  - each search options has **sub-categories**:

    i.e., feeds are sub-categorized into agridea, classified and unclassified feeds:
  - search options are **loaded dynamically** based on the current user selection:

    i.e., only those nutrients which are contained in the selected feeds are loaded.

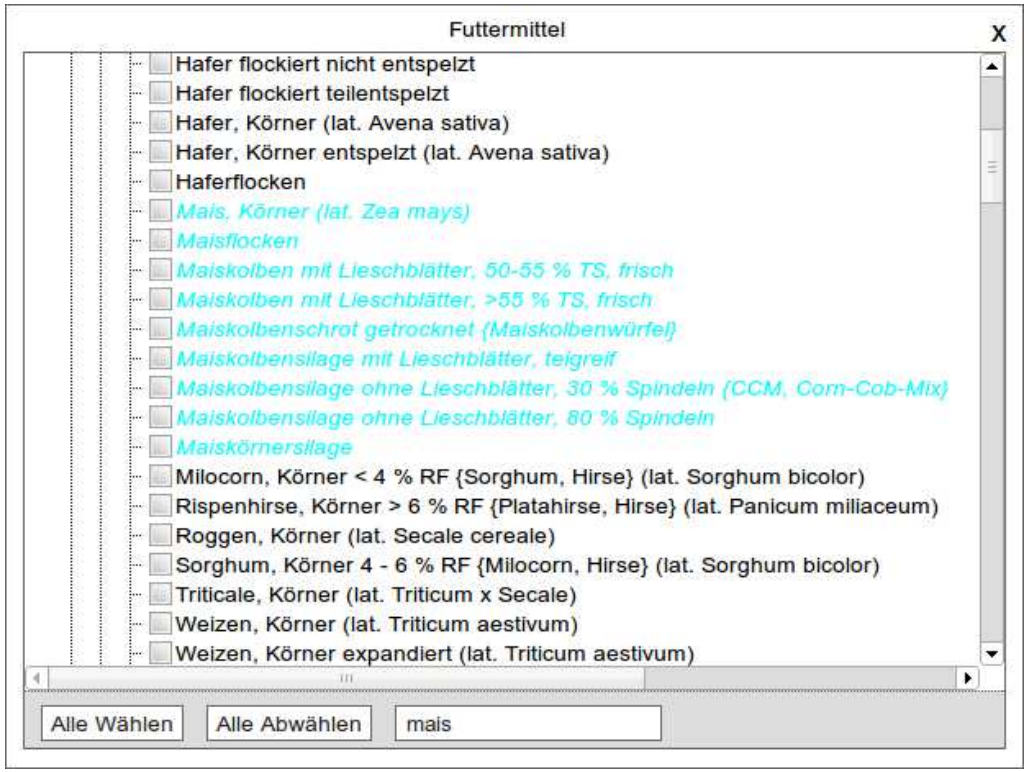
The System: The User Interface

- **Compact** and, at the same time, **rich search interface**:
  - each search category corresponds to a button;
  - one click activates a window for the chosen search category;
  - sub-categories are organized into tabs.



The System: The User Interface

- **Compact** and, at the same time, **rich search interface**:
  - search options are organized into a tree of check boxes;
  - easy search and selection.



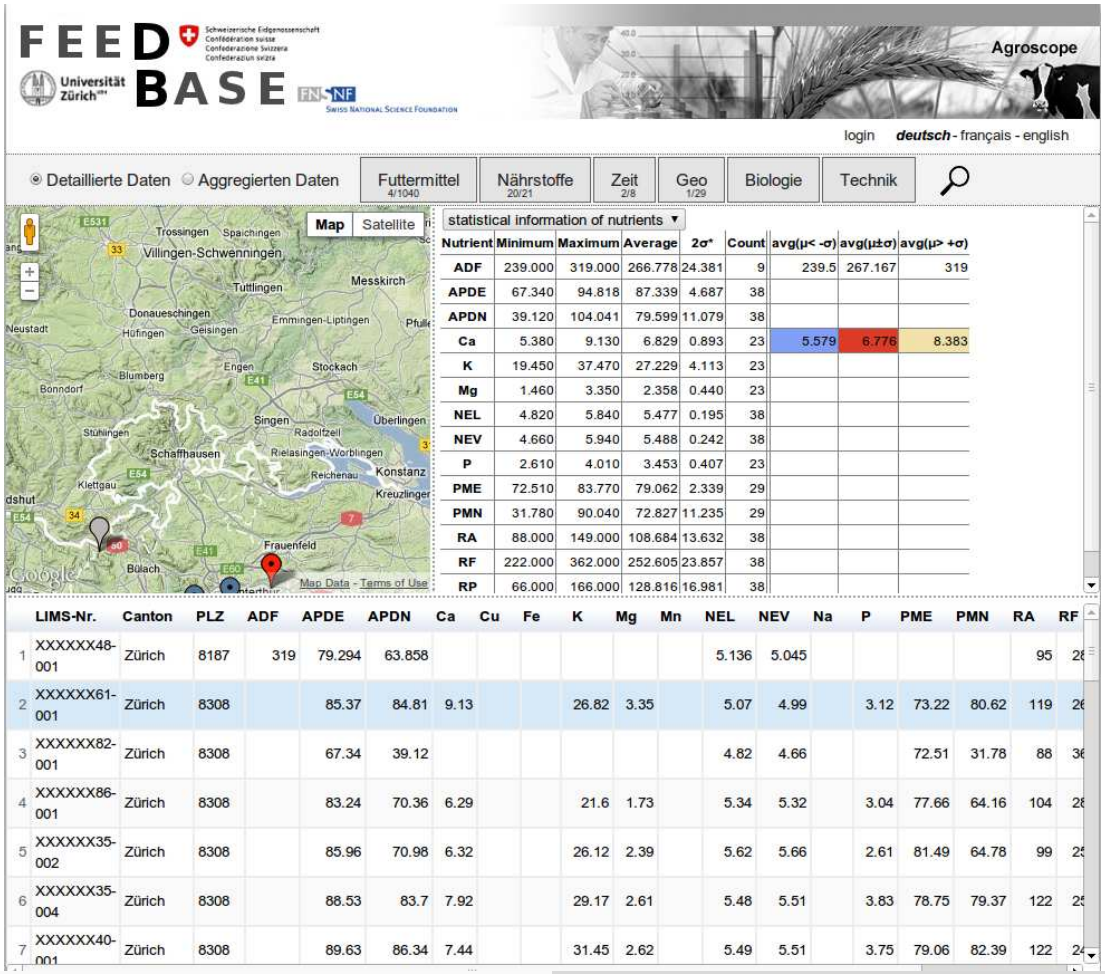
The System: The User Interface

- **Interactive** visualization of the query results:
  - 3 frames are used to visualize statistics, raw and geo data
  - is possible for the user to change the size of each frame;
  - the user can select individual records;
  - visualization in the geographical frame is adjusted to highlight the user selection:
    - a flag is activated showing all associated measurements;
    - flags are colored based on the deviation of the measurements;

The System: The User Interface

- **Interactive** visualization of the query results:

0



The System: The User Interface

- For the successful product the following missing futures are critical:
  - translation and proper description:
    - most of the data is in German;
    - description of the feed types, nutrients for Agridea and Lims data;
  - feed catalog:
    - currently is available only in old system;
  - help for the users;
  - integration of formulas;
  - collection of the user feedback;
- Only a competent person from Agroscope can accomplish these tasks.

The System: Development and Maintenance

- Development:
  - we organize the work into projects: 10 projects has been defined and 6 projects are completed by the students;
  - the major focus of the project is the enrichment of the user interface with missing functionality on temporal and geographical information:

    as radius search and computation of derived nutrients from the temporal data
  - A. Bracher significantly contributed in describing the projects and providing the necessary feedback;
  - for the successful development we need to define new projects and a competent person to evaluate the results.

The System: Development and Maintenance

- Maintenance:
  - Our main tool to update the data is PgAdmin:
    - similarly to Excel, allows to modify the data by clicking in a cell of the table;
  - For large amounts of data we provide the tools to automatically import the data whenever it is possible:
    - import of the Lims files;
    - import of the data from the excel files (in progress);
    - updates on the individual samples via web interface.
  - a person from the Agroscope is required to perform these task on a daily basis.

The Research

- Linear Parsimonious Temporal Aggregation (on-going).
- Efficient Nearest Neighbor Join among Fact Tables (to be started.)
- Paper at IEMSS 2012 (abstract submitted).

The Research: Clustering

- Goals:
  - detect changes in the feed quality and new feed types;
  - classification if the feed samples.
- The Data:
  - amino-acid profiles;
  - historical nutrient measurements;

The Research: Clustering

- clustering aims to group records based on the similarity/dissimilarity between them.
- it is common to define similarity as the Euclidian distance in n-dimensional space, where n-is the number of attributes.

The Research: Clustering

- dimensionality of the data:
  - as dimensionality grows the distances between records increase and, thus, the clusters become vague or disappear;
  - the feed data is high-dimensional, i.e, more than 600 nutrients and 30 amino acids.
- parameters:
  - parameters help to tune the clustering for the desired data, i.e., we can assign different weight for dimensions, limit to find only spherical clusters or eliminate outliers;
  - to find necessary parameters and their values is one of the most time consuming tasks.

The Research: Clustering

- analyses of the results:
  - the goal is to provide interpretation of the clustering results;
  - the most crucial and challenging task;
  - requires:
    - experience with clustering techniques and knowledge of the data domains;
    - time... to recompute clustering with different parameters and exploration of the raw data.

Project Structure

WP0 Project Coordination

WP1 Modeling of Time-Varying Measurement Sets

WP2 Parsimonious Aggregation of Multi-Granular Measurement Sets

WP3 Time-Varying Correlations of Measurement Sets

WP4 The Swiss Feed Data Warehouse

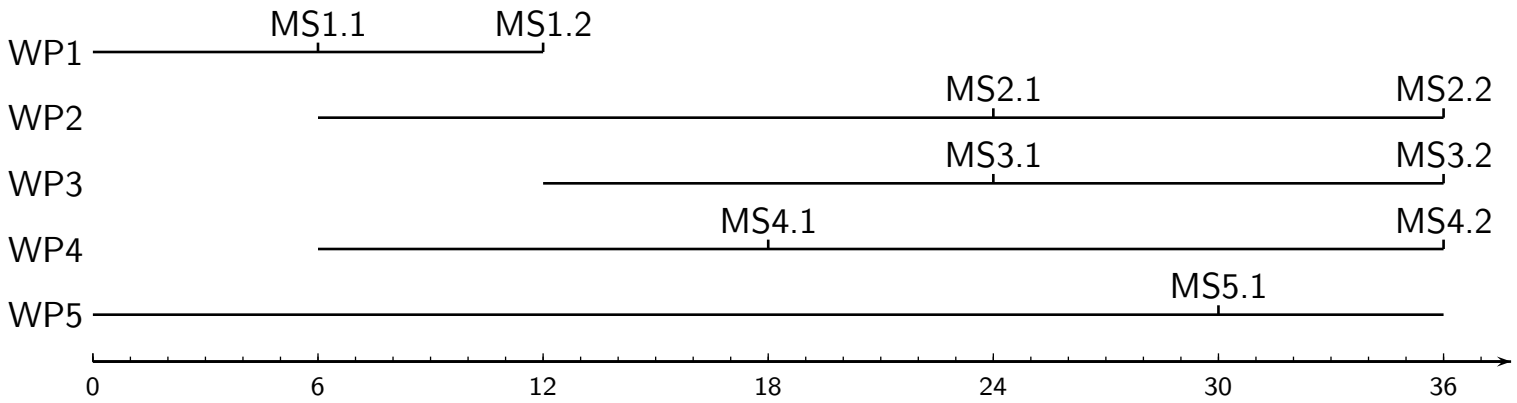
WP5 Curated Swiss Feed Data

Work Package 0

- regular meetings (Agroscope and ifi meet bimonthly)
- organization of seminars
- definition and supervision of student projects
- presentation of student projects during meetings
- scientists in charge meet once a year; presentation of ongoing research; discussion of planned activities; demonstration of feed database

Schedule

Project start: September 1, 2011
Project end: August 31, 2014



Work Package 1

Modeling of Time-Varying Measurement Sets

MS1.1 (month 6) design and initial build of a temporal model of the Swiss Feed Database.

MS1.2 (month 12) populate the feed database with measurement sets collected over the last years; online computation of extended up-to-date summaries.

D1 temporal model for the Swiss Feed Database;

D2 implementation of the temporal Swiss Feed Database with online computation of extended up-to-date summaries.

Work Package 2

Parsimonious Aggregation of Multi-Granular Measurement Sets

MS2.1 (month 24) publication of a paper on parsimonious temporal aggregation of multi-granular data.

MS2.2 (month 36) publication of a paper on parsimonious temporal aggregation across categorical attributes.

D3 publication about parsimonious aggregation of multi-granular measurement sets

D4 publication about parsimonious temporal aggregation over categorical attributes.

Work Package 3

Time-Varying Correlations of Measurement Sets

MS3.1 (month 24) an algorithm for automatic detection of correlations; increased number of available correlations in the feed database.

MS3.2 (month 36) publication of a paper on the imputation of missing values in scientific databases with multiple correlations.

D5 algorithm to determine time-varying correlations

D6 publication and implementation of online imputation of missing measurement sets via time-varying correlations.

Work Package 4

The Swiss Feed Data Warehouse

MS4.1 (month 18) first version of a data warehouse with reporting of data quality and confidence, and traceability of legacy query results.

MS4.2 (month 36) publication of a paper on the detection of relevant projections in scientific data.

D7 Swiss Feed Data Warehouse with reporting that includes data quality and confidence, and possibility to trace previous query results

D8 publication on detection of relevant projections in multi-dimensional time-varying measurement sets.

Work Package 5

Curated Swiss Feed Data

- common format and extraction procedure for the raw data with varying level of detail
- visualization and pre-processing of the raw data for detection of shifts and outliers
- increased availability of spatial, biological and technical properties of the feed data
- more quality feed data from new sources
- surveys on data quality