Using Data from Multiple Sources – the Reality of Genetic Evaluations

João Dürr, CDCB CEO
ICAR 2016 – Challenges and Opportunities
Puerto Varas, Chile, October 26, 2016
Data management

Case 1: Interbull Centre (2008-2014)

Case 2: CDCB (2014-Present)

Take home messages
Data Policies
Data validation is the process of ensuring that a program operates on clean, correct and useful data. It uses routines, often called "validation rules" "validation constraints" or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system.
CASE 1: INTERBULL CENTRE
Features of the Interbull Data Pipeline

• Data suppliers (April 2016)
  • 391 dairy cattle populations, from 34 countries

• Evaluations calendar
  • 3 Annual official evaluations
  • 2 Test runs
  • 5 different national evaluation validation methods

• Data types
  • National genetic merit data (EBV, PTA)
  • 1825 country-breed-trait combinations
  • Pedigrees
  • Population parameters
  • National evaluation validation tests
  • Genotypes (Intergenomics - BSW)
Interbull Centre - 2008 Opportunities

- No database, only flat files
- Each trait group developed separately
  - Independent file formats
    - Duplication - inconsistencies
  - Separate procedures
    - Different edits/checks
    - Separate processing, different levels of automation
  - Analyst-dependent
- Pedigree re-built from scratch every evaluation
- Limited documentation
- Validation of national evaluations not synchronized with users
The joy of developing a database…

Test if you are ready to start developing a DB by answering these very simple questions:

• Why do you need a database?

• Which are the business rules?
  • Are those effectively using the DB involved in validating the business rules?

• Would a person that knows nothing about your business (the DB developer, for instance) be able to follow the business rules?

• Have you identified a driver for the project?

• Do you have a DB administrator since the beginning of the process?
  • Is your DB Admin happy with the choice of tools?

• Is your budget for the project realistic?

IF YOUR ANSWER FOR ANY OF THE ABOVE IS “NOT SURE”, “NOT YET” OR “ALMOST THERE”

YOU ARE NOT READY TO START!!!
Standardizing data ingestion

- Interbull Centre solution: IDEA
  - Data type and range validation performed locally prior to upload
  - Cross-reference validation performed at the Interbull servers during upload
  - Interactive interface with users to intermediate data acceptance
  - **Golden rule: only data suppliers can modify input data**

- IDEA for pedigrees
  - Principle of “Authoritative Organization”
  - Data flow independent from evaluation deadlines

- IDEA for genetic merit
  - Same file format for all traits
  - “Verify” checks summarized by well established indicators
Interbull Validation of National Evaluation Estimates

• Opportunities
  • Tests applied with subtle differences in implementation yielded different results for users when compared to the Interbull Centre results
  • Much time spent on communication to find out why results were not identical

• Interbull Centre solution
  • Software supplied by the Interbull Centre is run locally
  • Test results and implementation details are recorded
  • Users and the Interbull Centre have access to the same figures
Interbull Centre ISO 9000 Certification

- Write what you do, do what you write
- Good documentation makes your life better
- Comprehensive business rules define your system’s credibility
- Version control is much easier when there is only one shared version of the document (Wiki)
- Quality is not an achievement, it is a life style
Lessons from the Case 1

- Databases: be sure you have a plan
- Standardizing data ingestion improves consistency through the use of efficient validation tools
- Keep comprehensive business rules and consistent documentation to stay in business
- Make sure your data suppliers see the same data quality indicators that you see
- Define clear roles and responsibilities between you and your data suppliers
2014 – Present (Discovery phase)

CASE 2: CDCB
Organization

- 12 voting members (3 from each sector)
- 2 nonvoting industry members
US Genetic Evaluation Process

Data
DRP
DRPC
PDCA
NAAB

R&D
AGIL/USDA

Implementation
AGIL/USDA
CDCB

Services
CDCB

U.S. Genetic & Genomic Evaluations
Quality Certification Services Inc.

Mission Statement
Providing a reliable source of information to people interested in the US dairy records industry.
### Official Evaluation Runs Since December 2015

<table>
<thead>
<tr>
<th>Record type</th>
<th>New records added between December 2015 and April 2016</th>
<th>New records added between April 2016 and August 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>First lactation test day records</td>
<td>3,012,084</td>
<td>3,061,753</td>
</tr>
<tr>
<td>Later lactation test day records</td>
<td>4,578,898</td>
<td>4,752,008</td>
</tr>
<tr>
<td>Heifer breeding records</td>
<td>963,249</td>
<td>918,528</td>
</tr>
<tr>
<td>Cow breeding records</td>
<td>5,164,212</td>
<td>4,833,899</td>
</tr>
<tr>
<td>Calving ease records</td>
<td>401,247</td>
<td>458,785</td>
</tr>
<tr>
<td>Stillbirth records</td>
<td>332,704</td>
<td>381,462</td>
</tr>
</tbody>
</table>
Number of genotypes received by CDCB
Number of genotypes stored in the CDCB database by continent of origin, sex and availability of phenotypic information (September 2016)

<table>
<thead>
<tr>
<th>Continent</th>
<th>Predictor Females</th>
<th>Predictor Males</th>
<th>Predicted Females</th>
<th>Predicted Males</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>6</td>
<td>-</td>
<td>374</td>
<td>48</td>
<td>428</td>
</tr>
<tr>
<td>Asia</td>
<td>15</td>
<td>1,826</td>
<td>2,101</td>
<td>883</td>
<td>4,825</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>24</td>
<td>425</td>
<td>2,120</td>
<td>591</td>
<td>3,160</td>
</tr>
<tr>
<td>West and Central Europe</td>
<td>226</td>
<td>15,250</td>
<td>57,113</td>
<td>45,886</td>
<td>118,475</td>
</tr>
<tr>
<td>Latin America</td>
<td>343</td>
<td>2</td>
<td>11,983</td>
<td>752</td>
<td>13,080</td>
</tr>
<tr>
<td>North America</td>
<td>324,437</td>
<td>29,240</td>
<td>772,096</td>
<td>133,902</td>
<td>1,259,675</td>
</tr>
<tr>
<td>Oceania</td>
<td>96</td>
<td>439</td>
<td></td>
<td></td>
<td>8,785</td>
</tr>
</tbody>
</table>
Pedigree and conformation records → CDCB

Pedigree, performance and management records → CDCB

Dairy Records Processing Centers

Genomic Nominators

Pedigree and genomic records → CDCB

Genomic Laboratories
Genomic data flow

DNA laboratory

Genotypes

Genotype quality reports

DNA samples

Genomic Nominator

Genotypes

Nominations, pedigrees

Genomic evaluations

Council on Dairy Cattle Breeding (CDCB)

Dairy Record Provider

(farmer or controller)

DNA samples

Genomic evaluations
# CDCB Fee Schedule
(Updated March 2, 2015)

<table>
<thead>
<tr>
<th>Rate Code</th>
<th>Participation type</th>
<th>Female fee ($)</th>
<th>Initial male fee ($)</th>
<th>AI service fee for males ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total program</td>
<td>0.00</td>
<td>15.00</td>
<td>575.00</td>
</tr>
<tr>
<td>2</td>
<td>Member</td>
<td>1.00</td>
<td>22.00</td>
<td>575.00</td>
</tr>
<tr>
<td>3</td>
<td>Non-member</td>
<td>3.00</td>
<td>150.00</td>
<td>575.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;15 mo</td>
</tr>
<tr>
<td>4</td>
<td>Canada</td>
<td>6.00</td>
<td>150.00</td>
<td>575.00</td>
</tr>
<tr>
<td>5</td>
<td>Approved partners</td>
<td>7.00</td>
<td>15.00</td>
<td>575.00</td>
</tr>
<tr>
<td>6</td>
<td>All others</td>
<td>7.00</td>
<td>150.00</td>
<td>1200.00</td>
</tr>
</tbody>
</table>
CDCB: Input Record Formats

- Pedigree
- Herd Info
- Lactation
- Reproduction
- Health
- Type
- Calving Ease
- NAAB AI codes
- Recessive codes
- Genotype Identification
- Genotypes
Bovine SNP chips processed by the CDCB
## Error-Codes for CDCB Data Checks (832)

### Example:

#### Gender Change Errors

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Action</th>
<th>Returned Data</th>
<th>Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0Pa</td>
<td>Format 4 can not change gender of animal.</td>
<td>Reject</td>
<td></td>
<td>09/26/2000</td>
</tr>
<tr>
<td>0Pb</td>
<td>Animal not found under opposite gender. Record type code is changed to 'P'.</td>
<td>Change</td>
<td></td>
<td>11/03/2000</td>
</tr>
<tr>
<td>0Pc</td>
<td>Change of gender for animal with different master file pedigree.</td>
<td>Reject</td>
<td></td>
<td>09/26/2000</td>
</tr>
<tr>
<td>0Pd</td>
<td>Change of gender for animal with master file lactations.</td>
<td>Reject</td>
<td></td>
<td>09/26/2000</td>
</tr>
<tr>
<td>0Pe</td>
<td>Change of gender for animal with master file progeny.</td>
<td>Reject</td>
<td></td>
<td>09/26/2000</td>
</tr>
<tr>
<td>0Pf</td>
<td>Change of gender for animal with multiple identifications.</td>
<td>Notify</td>
<td>Cross-reference Identification and pedigree source</td>
<td>09/26/2000</td>
</tr>
<tr>
<td>0Pg</td>
<td>Change of gender for animal with homozygous row.</td>
<td>Reject</td>
<td></td>
<td>04/08/2009</td>
</tr>
<tr>
<td>0Ph</td>
<td>Change of gender for animal with confirmed genotype.</td>
<td>Reject</td>
<td></td>
<td>09/09/2010</td>
</tr>
</tbody>
</table>
CDCB Evaluation Calendar

- 3 Annual Official Evaluations
  - Conventional
  - Genomic
  - Interbull files
- Monthly Genomic Evaluations
- Weekly Genomic Predictions
- 3 Annual Interim Evaluations
CDCB - Opportunities

- Transition from USDA to CDCB
  - Recruiting
  - Transfer DB, web applications, directory/files structures, programs
  - Knowledge transfer
  - Roles & responsibilities between AGIL and CDCB
  - Communication

- Multiple file formats
- Web applications developed in several platforms
- Heavy use of SAS in data processing
- Documentation
  - Not consolidated into a unique platform
  - More oriented to operations
  - Limited on business rules
Agents involved in the data pipeline

- Technical advisor
- Dairy Farmer
- Genetic Evaluation Specialist
- Field Technician
- Data Processor
- Laboratory Technician
Standardization of New Data Types

Example: MAST = 132 unique 4-letter acronyms used!
CDCB – First steps

- No changes to the legacy before transition was complete
- Keeping the “old pals” around
- Documenting the legacy
- Strengthening AGIL
- Establishing a policy to compensate phenotypic data suppliers
- Reviewing data access policy
- Developing a new web portal
- Standardizing file formats
- Refining genomic data flow
Lessons from Case 2

- Dairy data awareness has changed the business
  - Control, roles and responsibilities need to be redefined
  - Business rules need to adapt
  - Data access needs to be adjusted
  - Data flow needs to be renegotiated
- Data quality
  - Every link in the chain has to participate
  - Acquiring and validating new data types requires a new mentality
Take Home Message

• Dairy data recording services need to remain relevant for dairy farmers in this fast changing industry.
• Data for genetic evaluations are a by product, not the main goal.
• Making data ingestion more efficient is an effort that involves all agents in the dairy chain.
• Access to dairy data will define the future of dairy genetics.
• Increasing awareness about data quality is the best protection against opportunistic new products in the market.
Thank You!
www.cdcb.us

Acknowledgements:
• Interbull Centre
• AGIL
• CDCB staff
• Data Providers