

Katmandoo

a biosciences database

David Butler¹ Brian Cullis² Avishesh Shrestha² David Rodgers¹

¹Queensland Primary Industries & Fisheries

²NSW Department of Primary Industries



NSW DEPARTMENT OF
PRIMARY INDUSTRIES



Queensland Government
Department of **Primary Industries and Fisheries**

Introduction

Plant genetic improvement is no different to many fields of endeavour:

- the frontiers of investigation push on, while
- practitioners grapple with implementation.

Information ~~management~~ **integration** has been a constraint.

Eg: Lessons from the national winter cereals molecular marker program there to be learnt!

- **Accurate phenotyping is a cornerstone of genetic improvement**
 - Not just experimental methods, data must be re-useable
- **The logical data model reflect the statistical foundations.**

Introduction

Crop improvement programs generate a range of data:

- *passport* information on accessions
- Contacts: cooperators, collaborators, managers, legal, IP, ...
- Environmental
- Genetic resources
- Phenotypic
 - qualitative
 - quantitative univariate/multivariate/temporal responses
 - multi-spectral

from single or multiphase experiments

- Ancestral records
- Genotypic
 - biparental crosses
 - diallel designs
 - routine screening of breeding populations

Not just crop genetic improvement:

- Horticulture
- Forestry
- Agrostology
- Sheep
- Beef
- ...

Information management: Does one size fit all?

Power comes from data integration

- Marker-trait associations
- Meta-analyses across environments
- Simulation tools for breeding strategies
Eg: QuGene

The goal of the *Katmandoo* project is to provide this integration.

Yet Another Mousetrap ?

Public and commercial options: (~ circa 2004)

- User hostile
- Flawed design
- Inadequate to meet current (statistical) needs
- Context dependent
- Response to emerging research needs?

Design brief

- Data model independent of the application
- Key data structures model based
- Multiple crops/species/studies
- Modular (not monolithic) software application
 - phenotypic
 - genealogy
 - genetic markers
 - cross development / breeding strategies
 - genetic resources management
- Extensible
- Scaleable
 - desktop to enterprise
- Affordable
 - run on common computing platforms
 - use common/free tools

Implementation

- **Application**

C# using the Microsoft .NET framework.

- **Database**

MySQL or SQLserver

- Currently migrating the data access layer to platform independent code.
- *Katmandoo* is free under the terms of a binary code license.

Phenotypic data model:

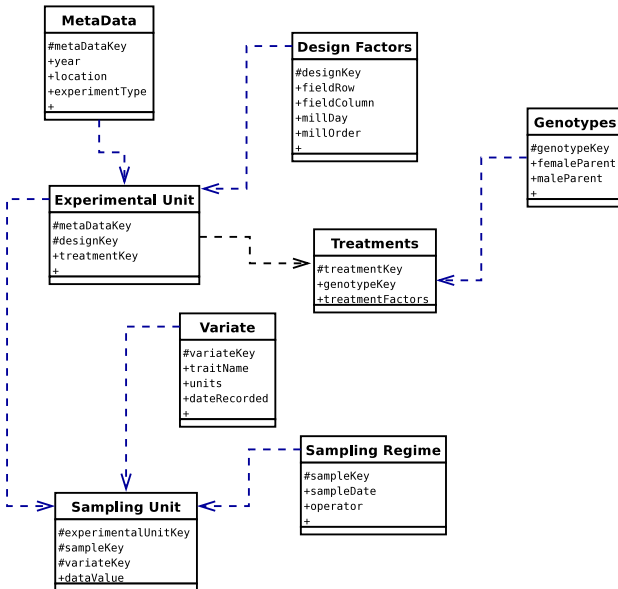
- Based on statistical notions of the:

Experimental unit: The smallest division of the experimental material such that any two units receive different treatments

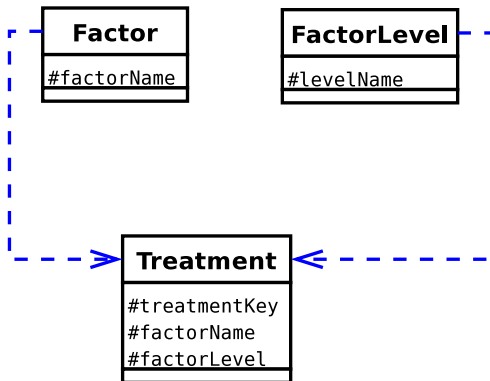
Sampling unit: The unit of experimental material on which observations are made.

- Experimental units identified by **design factors**
row+column or *batch+day+time-of-day*
- Sampling units identified by
 - experimental unit
 - sampling layers
 - trait

Data Structures



Extending the Data Model



Pedigrees in *Katmandoo*

- Recursive structure

me	dad	mum	fgen
1	0	0	0.8
2	0	0	0
3	1	0	2
4	1	2	0
5	1	2	1
6	1	2	2
7	1	2	3

- *Katmandoo* can build this from
 - historical records
 - cross generation
- Export to derive numerator relationship matrices.
- Companion application for pedigree based marker assisted selection.

Genetic Markers in *Katmandoo*

Typically:

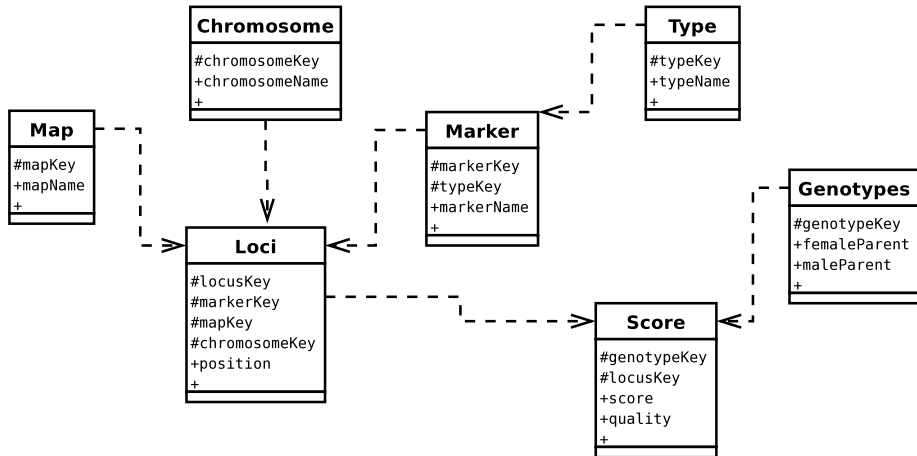
- multiple characteristics per marker position
- 10^x marker positions ($x = 3, \dots?$)
- $> 10^3$ genotypes per year

A first cut:

- Normalized relational model
- Consequences:
 - every data point addressed
 - indexing overheads
 - performance ?

Genetic Markers in *Katmandoo*

A first cut:



High Dimensional Data

The marker problem is a window to wider considerations:

- Gene expression
- High throughput phenotyping
- NIR
- Hyper-spectral

looming in applied breeding programs.

Genetic Markers in *Katmandoo*

A second go:

- Array based methods
- Hybrid system
 - map to vectors
 - store in array structures (NetCDF)
 - address the *vectors* from relational structure

Alternative:

- Novel indexing/storage technologies

Constraints:

- Corporate servers and foreign processes
- DBMS implementation

Summary

Katmandoo is a practical example of a generic approach in a defined application zone.

- Model based
 - relational algebra
 - statistical principles
- Scalable within the confines of common architectures.
- Respects current statistical practice.
- Working towards exploiting emerging data-rich technologies.