Farm-to-table science: dairy data mining for future resilience

J.I. Ohlsson, T. Klingström, D.-J. de Koning

Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden
Corresponding Author: jan.ingemar.ohlsson@slu.se

Abstract

Adoption of automated milking systems (AMS) and other precision livestock farming (PLF) systems offers access to large, multidimensional data that allow exploration of resilience traits and sustainable farming strategies in real-world scenarios. Since their inception in the Netherlands in the 1990s, AMS have seen increased adoption Adoption of automated milking systems (AMS) and other precision livestock farming (PLF) systems offers access to large, multidimensional data that allow exploration of resilience traits and sustainable farming strategies in real-world scenarios. Since their inception in the Netherlands in the 1990s, AMS have seen increased adoption in the Nordic countries, with around a third of the total milk production collected by robots. The major brands of farm management systems (FMS) in the Nordic region are only configured to report data as a current overview, discarding older information that is vital to studies of the herd's genetics, behaviour, and environment.

In this work, we present the infrastructure for dairy cattle data at the Swedish University of Agricultural Sciences, Gigacow (SLU Gigacow) that collects data from a set of Swedish dairy farms. Each farm's FMS sends nightly reports to SLU Gigacow, where records are harmonised and collected in a central database. Collected records include milking statistics, health events, traffic data, and SNP genotypes for thousands of cows, and are made accessible to researchers through SQL or R queries. SLU Gigacow also integrates data from the Swedish national cow registry, including pedigrees and herd transfers for cows resident at participating farms.

SLU Gigacow's longitudinal observations (first data collected in 2020) link genotype to phenotype and animal welfare with the goal of accelerating pilot studies in dairy science, as well as providing a big dataset from cows in active, commercial settings. The data collection software written in Python 3 (Beaverton, USA) has modules that enable collection from several versions of DeLaval DelPro (Tumba, Sweden), and can be extended to any FMS with a graphical user interface running on most consumer operating systems. After harmonisation to resolve differences in language and FMS versions, data are stored in a database maintained at SLU with SQL Server Integration Services (SSIS) (Microsoft, Redmond, USA). By agreement with Växa Sverige AB (Uppsala, Sweden), participating farmers also get a large number of animals genotyped using the 45k EuroG MD beadchip (Amsterdam, The Netherlands). Currently, the database includes information on over 17,000 cattle, over 3,000,000 milkings, and 2,969 SNP genotypes. The cross-referenced data can be mined for various purposes, including stress responses and resilience traits.

While SLU Gigacow is intended to collect from Swedish farms and support Swedish researchers, it serves as a proof-of-concept that data from diverse sources and systems at dairy farms can be automatically gathered and collated in a researcher-friendly format. We believe that this shows the great utility of farm-to-table statistics

and increased FMS interoperability. SLU Gigacow was constructed essentially without standardised interfaces for dairy data communication. Establishment of such data standards is ongoing within the industry but the development and adoption of standards take time and rely on active participation of multiple actors. The research-driven approach of SLU Gigacow enables more rapid and integrated measurements of many facets of the dairy farm environment, creates new niches for PLF equipment, and opens great new vistas of information to explore for adaptation to changing climates. .

Keywords: ingemar ohlsson, tomas klingström, dirk-jan de koning, dairy science, big data, resilience.

Introduction

Changing climates make resilience a highly desirable target in livestock breeding programs, including for dairy cattle production. Some resilience indicators specific Changing climates make resilience a highly desirable target in livestock breeding programs, including for dairy cattle production. Some resilience indicators specific to dairy cattle have been identified (Bengtsson, C., 2022; Kašná, E., 2022), and increasingly sophisticated methods are being applied to find genetic factors implicated in e.g. heat tolerance (Carabaño M.J., 2017; Chen S., 2023). Automated milking systems (AMS) and farm management systems (FMS) integrating a variety of sensors around the cow and the farm provide a great amount of data that can be leveraged to refine resilience studies. In a simple example, daily temperature on a farm and daily milk yield from its resident cattle can be correlated to explore the impact of temperature on productivity. If those cattle are also genotyped, which is done as a routine measure in genomic breeding, varying responses to heat stress can be correlated with genomic features, and novel heat stress tolerance traits can be identified.

Access to data can be a challenge to new data-driven dairy science projects, not least in the case of resilience studies, where location and timing can greatly affect the stresses animals experience. Having an existing database of recent and historical data alleviates the problem of timing, and can reduce the threshold investment necessary for a pilot study. The SLU infrastructure for dairy cattle data at the Swedish University of Agricultural Sciences, Gigacow (SLU Gigacow) aims to provide such a data source. .

Framework and implementation

Legal framework

Data collection has been organised based on the EU Code of conduct on agricultural data sharing by contractual agreement and Swedish animal protection law. The farmer is the data originator for all data collected in SLU Gigacow and provides a broad consent for research using data originating from the farm. The farmer also authorises SLU Gigacow to request data from service providers such as Växa Sverige and the Nordic Cattle Genetic Evaluation (NAV; collaboration including Växa Sverige) who provide data to SLU Gigacow under separate contracts regulating immaterial property rights and requiring each researchers using SLU Gigacow to sign a researcher consent to comply with the contracts set between SLU Gigacow, the data originator and the data providers.

Data collected directly from the farm or different data providers are kept separately to ensure that researchers do not accidentally use data for which they are unable to fulfil their obligations to a specific data provider. Specifically, in the current implementation, data may originate from participating farms (anything extracted from FMS), from Växa Sverige (Kokontrollen), or from NAV (SNP genotypes), and research users may be granted specific access to data from any combination of these sources. All unique

identifiers for a farm are pseudonymised and public example datasets are scrambled to provide examples of Gigacow content without the risk of a viewer using third party data to identify a farm. Farmers can request access to data collected from their farms and when data collected for SLU Gigacow is useful for the farmer, such as in the case of genotyping for genomic selection, the data is available for the farmer through their normal service provider.

Data collection and storage

Currently, all farms providing data to Gigacow use DeLaval DelPro AMS or milking parlors, monitored by DelPro Farm Manager software (versions 5.1-6.13).

The central Gigacow data collection pipeline consists of three blocs of software:

Client scripts — Python 3.10 script running on the FMS client computer at each farm. This script is executed nightly, simulating a user via the PyAutoGUI Python package, to output the past day's milking reports from the FMS. It then attempts to upload these milking reports, plus others scheduled to output from DelPro Farm Manager, to the SLU Gigacow harmonisation server by SFTP connection.

Server scripts — Python 3.10 script running nightly on the SLU collection server. The script processes all data in the farm upload file area. Herd identifiers (in the Swedish system, an integer of max 6 digits; occurring independently, or as part of animal ID) are pseudonymised to an 8-character alphanumeric string. A farm ID-to-pseudonym key is retained for future reference on the collection server. Thus, the same pseudonym can be used downstream to cross-reference animals belonging to the same herd, but the herds and farms are not directly identifiable by third-party users of the data. Data are also harmonised to CSV files with structured file and field names for each data type (milking records, culling records, feed data, etc.).

Database — SQL Server Integration Services storage platform managed by SLU. A set of import scripts takes in the pseudonymised and harmonised CSV intermediates and processes them for storage in SQL tables. End users with SLU intranet credentials can then access this database, either by direct SQL queries or intermediary applications such as the R DBI (R-SIG-DB, 2022) and ODBC (Hester *et al.*, 2023) packages.

Separate Python scripts also exist for maintenance and updates, as well as pseudonymisation integration of corresponding data from the Swedish national cow registry VÄXA Kokontrollen.

A repository of auxiliary scripts is under active development (*https://github.com/TKlingstrom/Gigacow-tools*), which includes various tools for accessing and manipulating data from the SLU Gigacow database.

Genotyping

All cattle born on the farm since joining SLU Gigacow are genotyped, and as many older animals as possible are genotyped when a farm joins SLU Gigacow. Genotyping is done using the normal commercial process for genomic selection where ear tissue removed when punching and tagging animals' ears is collected for genotyping by chip sequencing. SLU Gigacow covers the cost for the farmer to order genotyping from

Växa Sverige, which outsources the sequencing work to Eurofins, and sends the data to the Nordic Cattle Genetic Evaluation (NAV) for breeding evaluation. NAV shares the SNP genotype files with SLU Gigacow, where the files are pseudonymised and stored in separate tables in the SQL database.

Data access

Once stored in the SQL database, collected data can be extracted and viewed with SQL queries, either directly or through intermediary helper programs. Access requires registration with the SLU Department of Animal Breeding and Genetics, Quantitative Genetics division, which maintains records of farmer and researcher consent forms, ensuring that research users reach only the subsets of SLU Gigacow data which they are allowed to use.

Data are primarily organised in views corresponding to the originating FMS export file.

Results

Data collection from participating farms commenced in 2020. The SLU Gigacow software was iteratively developed as more farms were connected, and currently collects data from DelPro Farm Manager versions 5.1-6.13 controlling fully automated milking systems as well as parlor milking systems. The current client-side scripts will continue to support versions past 5.10.

Support for Lely Time4Cows was planned but put on hold, as this software is supposed to be replaced by Lely Horizon, with major changes expected. As of the time of writing, the SLU Gigacow team has left Lely integration on indefinite hiatus until the participating Lely farms (N = 2) receive software updates and FMS operation can be tested.

	Milkings		Traffic		Feed	
Farm Pseudonym	Aggregate	Unique animals[1]	Events	Unique animals[1]	Records	Unique animals[1]
169e580a	-[2]	-	27401	355	706	93
540275a1	315778	232	1008	168	-	-
5b581702	12912	105	-	-	5235	83
5c06d92d	600110	469	-	-	-	-
5f7f33d6	1684	1[3]	-	-	-	-
752efd72	566127	474	74725	487	925444	1170
a624fb9a	316963	194	128945	276	423617	212
a756bc39	161326	193	505	1[3]	863452	999
ab18b151	308466	201	25658	161	269083	203
ad0a39f5	923850	482	135299	485	-	-
afdd9a78	46301	68	-	-	25741	69
f454e660	336456	287	38036	232	394661	341
Total	3589973	2706	431577	2165	2907939	3170

[1]Unique animal ID found associated with the given data type.
[2]Available data varies extremely from farm to farm.
[3]Client scripts will attempt to find and upload historical data that is up to one year old, 10 files at a tim
Recently added farms take some time to catch up to current records.

culling, feeding, milking, traffic, and health related events. Genotyping has returned SNP data for 2,969 individual cows to date, and updates will be ordered every 3 months.

Some cow identities in the collected data are incomplete. Animal ID numbers used by DelPro may occur without any associated national animal ID, though every animal should have complete data in its birth record. While the number of incomplete identities is small, some patterns of missing data can be found.

An example dataset with scrambled and pseudonymised data is available at *https://tklingstrom.github.io/gigacow_exampledata/*.

Discussion

Since connecting the first farms in 2020, the volume and diversity of data collected by SLU Gigacow has grown considerably. Farms using DelPro now upload, at minimum: daily milking reports, culling events, reproduction events (calving, heat, dry-off, etc.), and cow identity information. Additional reports can be attached depending on the data that farmers choose to store in their FMS, such as health events and gate traffic. Progress with Lely has been delayed due to implementation of Lely Horizon coinciding with the planned rollout, highlighting the difficulty of working with the rapid pace of development in the industry.

It should be noted that some forms of data are often absent or unreliable, for reasons that were common between farms. Health events, for example, are frequently recorded in hardcopy and kept in binders, likely in the same office as the FMS client computer. This can, for example, simplify handling of veterinarians' signatures on certain procedures, and may involve forms and record sheets that have been in use much longer than the digital FMS. The main challenge in redirecting hardcopy records to FMS would seem to be making the interfaces simple yet competent enough to match pen and paper.

In addition to information being physically stored, fragmentation of data between different digital ecosystems of data present a challenge for farmers and researchers. In Sweden feeding systems and activity monitors are frequently chosen from other manufacturers than the provider of the milking equipment and the farm management system. Farmers are therefore often forced to consult multiple different applications to inform themselves of the current status of the farm. Identifying key data sources in this digital milieu is therefore an important consideration for further development of SLU Gigacow. Some companies like Nedap make data sharing a competitive advantage and provide an easily accessible API to which farmers can generate and share 'tokens' facilitating data access for advisors or other actors such as SLU Gigacow (*https://api.nedap-bi.com/api/redoc/*). In other cases the fragmentation of data between equipment manufacturers create new business opportunities such as Feedlync (formerly Cowconnect, Asperup, Denmark), which can be installed on feed mixer wagons and supply live feed data to a cloud storage with further integration with advisory systems and other equipment. Mapping these resources and identifying the best way to collect research data from them is therefore a continuous task to expand the capabilities of a dairy data infrastructure like SLU Gigacow.Incomplete cow identities are not unique to DelPro, although some error classes we have identified are shaped by the way data is input and linked in DelPro Farm Manager. With another FMS, you might for example not encounter an integer "animal number" as the primary identifier, and thus the records would be broken in a slightly different way. In the end, broken records derive either from human error during input into the FMS, or due to problems with automated detection of animal tags. It is difficult to conclusively prevent or repair every possible error, but we continue to investigate ways to identify and highlight errors. So far, a part of the design philosophy behind SLU Gigacow has been to present the

data unaltered whenever possible, so we want to avoid editing or removing records unless they specifically damage data consistency. However, near-future versions of SLU Gigacow may implement methods similar to those presented by Hermans *et al*. (2017), where key cattle life events are codified and examined to make sure they occur in logical order, e.g. heat-insemination-calving.

An important component of SLU Gigacow has been the focus on agile development. Each bloc of development has been done in collaboration with researchers able to use the data collected as a "minimum viable product". This has provided the development team with rapid input from researchers helping to prioritise efforts and make design decisions such as relying on Rstudio (Posit Software, Boston, USA) and the dplyr package (Wickham *et al*., 2023) as a primary data collection method from the database rather than a programmatically more complex solution relying on OLAP cubes or similar business information management solutions.

SLU Gigacow operates from a farmer-centric view where data sources and sensors useful for farming operations are evaluated and technical solutions for data collection identified. Most commercial sensor developers operating in the region recognise the rights of the farmer as the data originator and owner of data which makes data collection possible even if not always easy due to technical barriers. An ongoing trend with new systems such as Nedap and CowConnect providing open Application Programming Interfaces only requiring a token enables farmers, advisors and researchers to maximise the value of sensor investments by integrating data from multiple sources. In combination with the development of iDDEN as a standard widely supported by major equipment manufacturers not yet providing full APIs this is likely to lead to greatly enhanced data access for livestock researchers. The structure of SLU Gigacow, with data harmonised into a unified format and stored in a versatile SQL database, makes it well adapted to follow data standards that overlap with its available information, including iDDEN. We hope it will convincingly show the promise of open data exchange to enable and empower future livestock research..

Conclusions

Even with the great data generation potential of modern digitised milk collection systems and other PLF technologies, researchers and developers must be mindful of the most basic errors, like mistyped input and falsely identified cows. It is not possible to ask perfection of either the farmers or their technology, but aggregating diverse data sources, such as with SLU Gigacow, can help detect and correct animal identity errors.

Merging data from the wealth of available sources on a modern PLF-enabled farm has uses beyond simply verifying identities. Connected data sources allow farmers, equipment developers, and researchers to find and observe complex patterns in livestock management. Such connections benefit from, or outright demand that data standards be in place to enable communication system-to-system and system-to-user. Global standards projects like iDDEN represent a unifying force, while the expanding and diversifying market of PLF devices may drive systems apart. Representatives of farming, scientific, and industrial interests should maintain communication to encourage a future PLF market that allows both creativity and diverse niches for system developers, and a good range of systems that work together for the farmers. In that way, we can ensure the resilience of PLF technologies going forward.

The process of developing SLU Gigacow has repeatedly shown that farms are individual, and both FMS and systems like ours that extract data from them must carefully consider

the needs and idiosyncracies of each customer farm. Active cooperation with farmers is necessary to advance agricultural tech and research, and open data standards can only support it. Research can here motivate early use cases for data integration and provide the experience necessary for optimizing new standards and define which data to be shared between systems. We hope SLU Gigacow will serve as a springboard for new researchers in the intersection of PLF and resilience, and illustrate what data sharing can accomplish with a modest effort..

References

Bengtsson, C., J.R. Thomasen, M. Kargo, A. Bouquet, and M. Slagboom, 2022. Emphasis on Resilience in Dairy Cattle Breeding: Possibilities and Consequences. J Dairy Sci 105 (9): 7588–99. https://doi.org/10.3168/jds.2021-21049.

Carabaño, M.J., M. Ramón, C. Díaz, A. Molina, M.D. Pérez-Guzmán, and J.M. Serradilla, 2017. BREEDING AND GENETICS SYMPOSIUM: Breeding for Resilience to Heat Stress Effects in Dairy Ruminants. A Comprehensive Review. J Animal Sci 95 (4): 1813–26. https://doi.org/10.2527/jas.2016.1114.

Chen, S., J.P. Boerman, L.S. Gloria, V.B. Pedrosa, J. Doucette, and L.F. Brito, 2023. Genomic-Based Genetic Parameters for Resilience across Lactations in North American Holstein Cattle Based on Variability in Daily Milk Yield Records. J Dairy Sci, April. https://doi.org/10.3168/jds.2022-22754.

Hermans, K., W. Waegeman, G. Opsomer, B. Van Ranst, J. De Koster, M. Van Eetvelde, and M. Hostens, 2017. Novel Approaches to Assess the Quality of Fertility Data Stored in Dairy Herd Management Software. J Dairy Sci 100 (5): 4078–89. https://doi.org/10.3168/jds.2016-11896.

Hester, J., H. Wickham, O. Gjoneski, 2023. odbc: Connect to ODBC Compatible Databases (using the DBI Interface). R package version 1.3.5, https://CRAN.R-project.org/package=odbc.

Kašná, E., L. Zavadilová, J. Vařeka, and J. Kyselová, 2022. General Resilience in Dairy Cows: A Review. Czech J Animal Sci Science 67 (2022) (No. 12): 475–82. https://doi.org/10.17221/149/2022-CJAS.

Lokhorst, C., R.M. de Mol, and C. Kamphuis, 2019. Invited Review: Big Data in Precision Dairy Farming. Animal 13 (7): 1519–28. https://doi.org/10.1017/S1751731118003439.

R Special Interest Group on Databases (R-SIG-DB), H. Wickham, K. Müller, 2022. DBI: R Database Interface. R package version 1.1.3, https://CRAN.R-project.org/package=DBI.

Silpa, M.V., S. König, V. Sejian, P.K. Malik, M.R.R. Nair, V.F.C. Fonseca, A.S. Campos Maia, and R. Bhatta, 2021. Climate-Resilient Dairy Cattle Production: Applications of Genomic Tools and Statistical Models. Front Vet Sci 8. https://www.frontiersin.org/articles/10.3389/fvets.2021.625189.

Wickham, H., R. François, L. Henry, K. Müller, D. Vaughan, 2023. dplyr: A Grammar of Data Manipulation. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr.