



Data ownership, privacy, use, sharing and stewardship - The U.S. experience

J.W. Dürr

*Council on Dairy Cattle Breeding, 4201 Northview Dr. One Town Centre, Suite 302,
Bowie, MD 20716, USA
Corresponding Author: joaodurr@uscddb.com*

The Council on Dairy Cattle Breeding (CDCB) is a non-profit collaboration between four sectors of the U.S. dairy industry: dairy records providers, dairy records processing centers, Purebred Dairy Cattle Association (PDCA) and National Association of Animal Breeders (NAAB). The purpose of CDCB is to host the national cooperator database (NCD) on behalf of the dairy community and use data analytics to provide value back to dairy producers through genetic evaluations and management information. The NCD is composed of three basic data types: animal relationships (pedigrees), management and performance records (phenotypes) and single nucleotide polymorphisms (SNP) markers (genotypes).

Phenotypes and pedigrees have been collected over a century by dairy herd improvement (DHI) services contracted by dairy producers to enhance decision-making process at the herd. Dairy producers own and control the use of data generated from their herds, and access to data is regulated by agreements signed with the DHI service providers. Genomic nominators provide tissue sampling and data collection services and serve as conduits between farmers, genotyping laboratories and CDCB for genomic predictions. Control over the use of genomic data is also exerted by animal owners and regulated by commercial agreements between service providers and users. The NCD receives data from a multitude of sources, and CDCB has material license agreements with each one establishing data access and use limits. A material transfer research data exchange agreement between the Agricultural Research Service (ARS) of U.S. Department of Agriculture (USDA) and CDCB allows ARS researchers access to the NCD for research purposes only. Other research organizations can only access data from the NCD if formally authorized by the data controllers. International data sharing initiatives mainly involve genotypes and are negotiated by data controllers and executed by CDCB accordingly.

CDCB is also investing in novel phenotypes data generation projects such as feed efficiency, which require a new business model to secure sustainable data flow. Dairy herds data from sensors is growing rapidly in the industry but in a disordered process that lacks standards, quality assurance, and means to be properly integrated to the existing data systems. Dairy herd data belong to dairy producers who invest in technology and pay for all services. Therefore, dairy data must serve primarily to improve the decision-making process at the herd level and the role of farmer-based organizations such as CDCB is to facilitate an effective integration of the existing and emerging data streams.

Keywords: Data stewardship, CDCB, U.S. dairy.

Abstract

Introduction

Data ownership has become a common discussion in the dairy industry since dairy herds started the transition from small-medium technology-empirical management to larger-hi-tech-professional management kind of operations. Nowadays most herds need to be managed as an enterprise that optimizes economic outputs by controlling costs and using the right technologies. This optimal management can only be achieved by using reliable data to monitor all aspects of the operation and therefore have a sound decision making process in place. The sector has accompanied the trends of precision agriculture and an abundance of new technologies to generate data at the farm have been incorporated into the system.

Besides the traditional milk recording data collected periodically (milk yield, components, somatic cell counts, reproductive events, culling and dry off events), now on farm sensors generate data 24 hours a day, a large proportion of calves are genotyped right after birth and genomic predictions are used as management tools, and a myriad of other service providers offer specialized data to be added to the farm hard disks (vets, nutritionists, artificial insemination planners, feedstuff suppliers, crop specialists, milk buyers, etc.). One can easily assume that dairy farmers are overwhelmed with data and integration of all these pieces of information is still far from adequate. The question to be addressed here is who owns these data, who has access to it and what can be done with it. This paper will focus on the specific situation of the U.S. dairy industry and how data ownership, privacy, use, sharing and stewardship are currently managed.

Dairy data in the U.S.

Dairy data in the U.S. has a long history and usually it is traced back to 1908, when the dairy herd improvement associations (DHIA) started collecting regular milk recording data. From the start, the U.S. Department of Agriculture (USDA) provided logistic support and research that evolved into the modern services supporting decision making and genetic evaluations. For most of this history, the Agricultural Research Service (ARS) maintained the dairy national cooperator database (NCD) and provided genetic evaluation services. In 1999 the CDCB was formed as a nonprofit corporation with the purposes of providing a forum in which to share information and coordinate activities that improve dairy cattle genetics and maintaining the integrity of data used in the genetic evaluation of dairy cattle. CDCB served as a communication channel between ARS, the dairy data providers and the organizations using and promoting U.S. dairy genetics. The three sectors represented in the CDCB board were the DHIA sector, the breed associations and the artificial insemination (AI) industry.

When the genomics era initiated in 2008, demand for data management and genetic evaluation services increased dramatically and both ARS and the dairy industry stakeholders realized that a new business model was necessary to fulfil the industry needs and continue at the edge of the scientific knowledge. A significant part of the research that allowed the genomic technologies to be adopted in dairy cattle was carried out by ARS researchers using data from the NCD. It became evident that it was time for the industry to take responsibility over the services and allow ARS to focus on the research. As a consequence, the NCD stewardship and the genetic evaluation services started to migrate from ARS to CDCB in 2013 and the process was completed in December 2015.

A material transfer research data exchange agreement between the two organizations establishes that ARS continues to have full access to the NCD for research purposes only and CDCB, besides receiving ARS legacy programs and expertise, continues to receive scientific support from ARS researchers and is committed to maintain quality certification programs to ensure the NCD data quality.

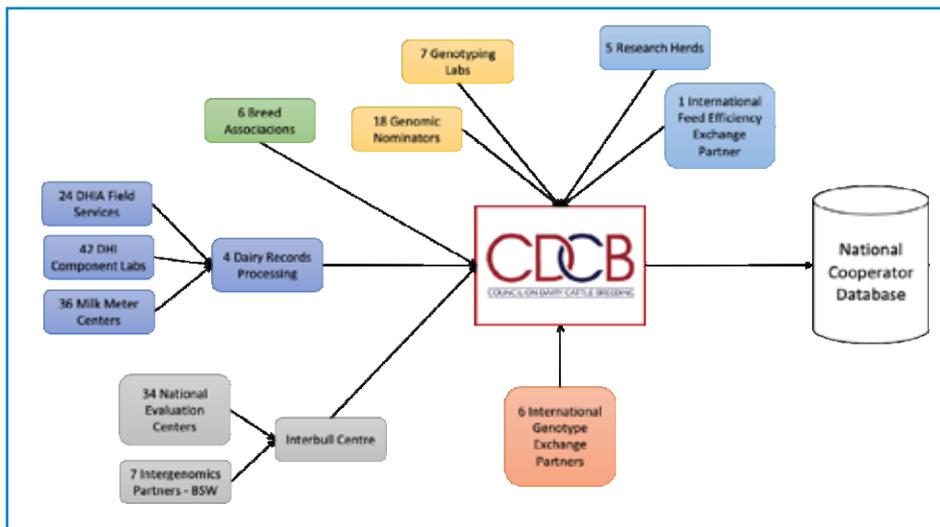


Figure 1 - Number of organizations providing data to the National Cooperator Database hosted by the Council on Dairy Cattle Breeding (April 2021).

Figure 1 shows the number of organizations providing data to the NCD. Data types flowing into the NCD can be described as animal identification, genealogy, performance, management, health, conformation appraisals, dry matter intake, breeding values and genomic markers. Quality certification for animal identification, performance, health and management data is provided by Quality Certification Services (QCS), for international genealogy and breeding values by the Interbull Centre and for genomic and genealogy data by CDCB. Quality control for novel data types (e.g., dry matter intake) is overseen by CDCB.

Dairy data management

The CDCB members position is that dairy data belongs to the owner or controller of the animal from which data was generated. Dairy producers pay for all services and technologies that generate dairy data. All the other agents included in Figure 1 are either data processors, stewards or users. This is a fundamental principle irrespectively of the data type. As a consequence, only dairy producers or animal controllers can authorize access and use of data stored in the NCD. CDCB is the NCD steward and follows the access and use policies established by data providers.

Data ownership, access and use

DHI data access and use is regulated by agreements between dairy producers with the DHI service providers. Genomic nominators provide tissue sampling and data collection services and serve as conduits between farmers, genotyping laboratories and CDCB for genomic predictions. Control over the use of genomic data is also exerted by animal owners and regulated by commercial agreements between service providers and users. Since the NCD receives data from a multitude of sources (Figure 1), CDCB has material license agreements with each data provider establishing data access and use limits. Research organizations can only access data from the NCD if formally authorized by the data controllers. International data sharing initiatives mainly involve genotypes and are negotiated by data controllers and executed by CDCB accordingly. CDCB is also investing in novel phenotypes data generation projects such as feed efficiency, which require a new business model to secure sustainable data flow. Dairy herds data from sensors is growing rapidly in the industry but in a disordered process



that lacks standards, quality assurance, and means to be properly integrated to the existing data systems.

Take home message

Dairy herd data belong to dairy producers who invest in technology and pay for all services. Therefore, dairy data must serve primarily to improve the decision-making process at the herd level and the role of farmer-based organizations such as CDCB is to facilitate an effective integration of the existing and emerging data streams and follow the access and use options of dairy producers as the national data steward.