Received: 9 October 2020 | Accepted: 29 January 2021

DOI: 10.1002/agj2.20639

Agronomy Journal

SPECIAL SECTION: BIG DATA PROMISES AND OBSTACLES: AGRICULTURAL DATA OWNERSHIP AND PRIVACY

Agricultural data management and sharing: Best practices and case study

Eli K. Moore[1,2] | **Adam Kriesberg**[1,3] | **Steven Schroeder**[4] | **Kerrie Geil**[1,5] | **Inga Haugen**[6] | **Carol Barford**[7] | **Erica M. Johns**[8] | **Dan Arthur**[9] | **Megan Sheffield**[10] | **Stephanie M. Ritchie**[1,11] | **Carolyn Jackson**[12] | **Cynthia Parr**[1]

[1] National Agricultural Library, USDA Agricultural Research Service, Beltsville, MD 20705, USA

[2] Department of Environmental Science, School of Earth and the Environment, Rowan University, Glassboro, NJ 08028, USA

[3] School of Library and Information Science, Simmons University, Boston, MA 02115, USA

[4] Animal Genomics and Improvement Laboratory, USDA Agricultural Research Service, Beltsville, MD 20705, USA

[5] Big Data Initiative and SCINet Program for Scientific Computing, USDA, Agricultural Research Service, Beltsville, MD 20705, USA

[6] University Libraries, Carol M. Newman Library, Virginia Tech University, Blacksburg, VA 24061, USA

[7] Nelson Institute Center for Sustainability and the Global Environment (SAGE), University of Wisconsin-Madison, Madison, WI 53726, USA

[8] Cornell University Libraries, Cornell University, Ithaca, NY 14850, USA

[9] Pasture Systems and Watershed Management Research Unit, USDA, Agricultural Research Service, University Park, PA 16802, USA

[10] Clemson University Libraries, Clemson University, Clemson, SC 29634, USA

[11] STEM Library, University of Maryland College Park, College Park, MD 20742, USA

[12] Medical Sciences Library, Texas A & M University, College Station, TX 77843, USA

Correspondence
Cynthia Parr, National Agricultural Library, USDA Agricultural Research Service, 10301 Baltimore Avenue, Beltsville, MD 20705, USA.
Email: cynthia.parr@usda.gov

Associate Editor: David E Clay

Funding information
National Institute of Food and Agriculture, Grant/Award Number: 2018-67023-27843; Agricultural Research Service, Grant/Award Numbers: 8042-31000-001-00-D, 8260-88888-001-00-D

Abstract

Agricultural data are crucial to many aspects of production, commerce, and research involved in feeding the global community. However, in most agricultural research disciplines standard best practices for data management and publication do not exist. Here we propose a set of best practices in the areas of peer review, minimal dataset development, data repositories, citizen science initiatives, and support for best data management. We illustrate some of these best practices with a case study in dairy agroecosystems research. While many common, and increasingly disparate data management and publication practices are entrenched in agricultural disciplines, opportunities are readily available for promoting and adopting best practices that better enable and enhance data-intensive agricultural research and production.

Abbreviations: DIDAg, Driving Innovation through Data in Agriculture; USDA-ARS, U.S. Department of Agriculture-Agricultural Research Service

1 | BACKGROUND

With the rise of smart farming technologies in agriculture leading to greater data creation and utilization by producers and researchers, many questions have arisen and still remain regarding data management throughout the agricultural sector (Wolfert et al., 2017). An analysis of 19 federal agency responses to the 2013 Office of Science and Technology Policy Memo (OSTP, 2013) requiring federally funded research agencies to increase and broaden access to research results indicated that data management best practices need further development (Kriesberg et al., 2017).

To improve transparency and efficiency, the Data Management Plans as a Research Tool (DART) project (Whitmire et al., 2017) is developing data management plan guidance in a variety of subject areas to encourage data re-use, to enable meta-analyses across disciplines, and to preserve information for future interpretation. The problem of transparency is amplified by leading discipline-specific repositories being insufficient to meet the needs of data science applications (Assante et al., 2016; Tenopir et al., 2015). The importance of data access to agricultural/natural resources researchers was identified in the 2017–2018 survey by the DataOne project (Tenopir et al., 2020).

In this paper, we address data management common practices in agriculture and describe best practices that will advance the field, while focusing on agricultural economics, dairy agroecosystems, production agriculture, and extension. Two workshops, Driving Innovation through Data in Agriculture (DIDAg), were held in June 2018 and August 2019 to bring together agricultural librarians, researchers, data managers, extension agents, experiment station personnel, university administrators, and other individuals with expertise in agricultural data production and management. As shown in the description of disciplinary best practices, DIDAg participants identified gaps in infrastructure or services needed to support those best practices and the needs of research in the future. Given the desire of the agricultural research community to employ growing data resources and emerging analytics approaches, and the opportunity to capitalize on historical data accumulated by producers and agricultural industry, the adoption of data management best practices is crucial for advancing 21st century agriculture.

Additionally, we provide a case study on current dairy agroecosystem research efforts to reduce greenhouse gas emissions while maximizing production through diet and genetics improvement (Figure 1). The dairy research community is an ideal model to illustrate the importance of integrating of scientific inquiry and historical data. The U.S. Department of Agriculture-Agricultural Research Service (USDA-ARS) researchers have a history of success developing new genetics methods and genomic statistical analyses for biological trait prediction in dairy cows (Van Tassell et al., 2008; VanRaden, 2008; VanRaden et al., 2009). In addition, the recent Dairy Coordinated Agricultural Project (e.g., Lane et al., 2019; Veltman et al., 2018) made substantial progress in integrating data from multiple institutions and disciplines to address key dairy sustainability questions. The USDA-ARS "Dairy Agriculture for People and the Planet" Grand Challenge Synergy Project (Tricarico et al., 2019) represents a unique opportunity to expand the use and integration of data on agricultural–environmental interactions.

Core Ideas

- Peer review can be important for ensuring that the value of agricultural data is maintained.
- Minimal data sets can foster re-use for innovation beyond initial data collection.
- Data repositories should be used and should promote best practices and data transparency.
- Engaging citizens in agricultural research can enhance data and adoption of research results.
- Funders, journals, institutions, librarians, and researchers all support good data management.

2 | BEST PRACTICES VS. COMMON PRACTICES

2.1 | Research data peer review

Peer review is a crucial mechanism for progress in research, including a range of methods (e.g., single blind, double blind, etc. [Blank, 1991]), in the scientific process for the validation

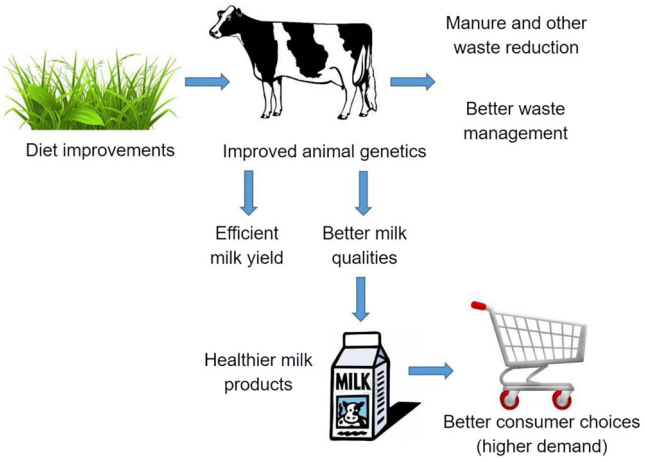

FIGURE 1 Previous work to improve dairy sustainability has focused on only some of the possible areas of impact (CC-BY-NC-ND 4.0 clip art from pixy.org)

and enhancement of research by disciplinary experts. However, there are various pitfalls associated with peer review (Kundzewicz & Koutsoyiannis, 2005; Langfeldt, 2006), and the common practices of peer review can be highly variable, as the amount of information required for peer review in scientific journals differs depending on the editorial board. There is uncertainty about how best to conduct peer review of repository data. There is also concern that the amount of time necessary for peer review of repository data could impede publication, or that additional resources necessary to review data could be needed after the project funding has ceased. The tasks involved in repository data peer review are often time consuming, therefore, clear review criteria will be very useful to the community.

As a best practice, DIDAg participants indicated that it is desirable for data in repositories to have some level of peer review or quality control. Ideally, research data are included in the peer review process of journal article findings based on the data. Currently, many repositories do not themselves offer peer review of research data but they do provide curation to ensure that metadata, methodology, and data processing are well described and consistent with FAIR principles. Following these principles can aid peer review by journals or data consumers, and generally make it easier for the others to use the information. The components of datasets that should be included in this quality review are clear metadata describing different types of data or estimates and limitations for their use, lab work methods or instruments involved, QA/QC for instrument calibration to reduce bias, the presence and circumstances of survey data, clear identification of raw or processed status of data, citations to any source data that was used in compiling the dataset, and references to algorithms used that created results from the raw data. Some workshop participants stated that five times as much commentary compared to actual data analysis code is necessary to fully explain the analysis and provide context to the data. For example, geospatial metadata should include which satellite system was used to collect the geospatial data being analyzed, because different satellite systems give different degrees of accuracy and temporal frequency.

Article peer reviewers could ensure that researchers deposit all their raw data into a repository with no filters or processing to allow a wider range of future analyses. This approach has not yet achieved community acceptance in part because the data collector may not see the value of preparing the metadata. The purpose of metadata is not for peer review, but metadata does allow peer reviewers to identify potential problems and errors in the data set. Peer reviewers could determine, for example, if comprehensive metadata should have been collected from associated meteorological stations with precision agriculture data to provide extensive baseline climate/precision metadata along with researcher-provided metadata. Peer reviewers could also check to be sure that additional site-specific information is included, (e.g., soil structure down to vadose zone with associated soil microbial and eukaryotic communities). Accompanying data, such as associated microbial community data, can serve as an important type of metadata (i.e., one researcher's data is another researcher's metadata and vice versa).

2.2 | Minimal dataset development

A common practice in dataset development is to include only data and metadata that are useful for their own project, without considering if other data would be useful for others. The cost of developing a dataset can be high and this may lead researchers to keep to a narrow scope. However, an appropriate best practice is to adhere to "minimal dataset" standards. A minimal data set is one that includes at least the minimum amount of data and metadata to ensure consistency, utility, and interoperability with other data sets. In many cases analyses of existing data must include replicates to quantify variation, but replicates or multiple observations are often not provided. The minimal requirements for a given category of dataset should be defined by the potential users (a specific research community, e.g. Kuru et al., 2013). For example, nitrogen and nutrient management researchers that provide average estimates of manure quantity and do not provide information on variability are not meeting minimal requirements. Researchers re-using data to understand sustainability of agricultural systems need longitudinal operational information with respect to manure application, soil nutrients, and other parameters. For example, tillage is still poorly understood because the agricultural community hasn't always included that information in datasets over time. The research community has census data that is only collected every 5 yr with variable quality, so adding a spatial component to census data, while protecting privacy (Massey, 2014; Schwartz & Solove, 2011), would be a big step forward.

A proposed minimal dataset is the Nitrogen Recommendation databases for fertility guidelines (Kitchen et al., 2017). Contributors to this project must include the information described in Supplemental Table S1. Conversely, water resource data is very limited and incomplete (Northey et al., 2016). Due to a lack of large-scale research on U.S. aquifers, researcher estimates of aquifer water capacity are limited. Minimal aquifer and livestock datasets would be very useful for understanding water resource management and risks. With large tradeoffs in how water resources are used, adoption of minimal water resource datasets will allow economists to apply different analytical approaches for managing at-risk water resource areas.

Minimal datasets would also make clear the multiple data scales needed to maximize the value of the information. A best practice is for water resource data or nutrient data to be

TABLE 1 Examples of international agriculture-related repositories

Repository	Full name
AGRIS	Agricultural Information Management Standards
	Linked to the Food and Agriculture Organization of the UnitedNations
CABI	Centre for Agriculture and Bioscience International
CAD	Commonwealth Agricultural Database
World Bank Data	
OECD	Organisation for Economic Co-operation and Development
IMF	International Monetary Fund
	Census aggregation IMF products (https://www.exiobase.eu)
UN	United Nations databases
UNFCCC	United Nations Framework Convention on Climate Change
	Greenhouse gas inventories - provides data submission templates for transparency, provenance, consistency, completeness, comparability, and accuracy
FAOSTAT	Food and Agriculture Organization of the United Nations
	Provides food and agriculture data for more than 245 countries and territories from 1961 to the most recent year available.
AgMIP	Agricultural Model Comparison and Improvement Project
Ag GRID	Gridded crop model simulations (https://www.ag-grid.com/)
C3MP	Coordinated Crop Climate Model Project
CMIP	Coupled Model Intercomparison Project
WTO	World Trade Organization

available at point-scale, and also aggregated on larger scales to increase the utility of the data for economic estimates. If researchers want to measure nutrients going into a basin they need to know if a point-based model works better than a gridded model, and what grid size is optimal to answer nutrient transport questions of interest. The finer the resolution, and larger the framework, the more scalability is available to answer different levels of questions. An example of best practices spatial resolution is the National Agricultural Statistics Service (NASS; https://www.nass.usda.gov/) aim to have a representative agricultural production sample in most areas of the United States, which incorporates stratified sampling that is mostly representative at the state level and in some counties. Key research topics in dairy economics over the next 10 yr that would benefit from minimal dataset development by the research community are given in Supplemental Table S2.

2.3 | Using and sustaining data repositories

2.3.1 | Using data repositories

Managing data locally and responding personally to individual requests for data is common practice among the research community. A number of attributes make a data repository trusted by the research community of a particular discipline, and these repositories may pursue certification (e.g., https://www.coretrustseal.org/). The use of trusted data repositories is not yet pervasive throughout the agricultural community, but this use is an important best practice that should be followed to ensure long-term, broad availability of valuable data. With or without formal certification, a repository must be secure, have stable funding support, and provide sufficient infrastructure and metadata to ensure understandability and usability of datasets. These repositories must also be affordable. When multiple repositories serve similar disciplines, shared policies and standards will allow users of these repositories to combine datasets. Data repositories and big international databases are only as good as their data submissions and their ease of use. The Ag Data Commons (https://data.nal.usda.gov) is an example of a national, cross-disciplinary repository suitable for U.S.-funded agricultural research data. The majority of agricultural and natural resources researchers (80.2%) are willing to share data across a broad group of researchers (Tenopir et al., 2020), which bodes well for the future utilization of data repositories. Examples of large international repositories are given in Table 1.

2.3.2 | Sustaining data repositories

In order to sustain data repositories, DIDAg participants emphasized that funders could require the use of specific data repositories. Subject matter repositories, such as the

National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and European Bioinformatics Institute (EBI; https://www.ebi.ac.uk/), are well established for specific disciplines. Some institutional and commercial repositories, including several associated with journals, may be trusted by communities as well, although they are often not able to archive large amounts of data. Certain types of repositories are more suitable for active use and analysis, whereas others are ideal for long-term storage and less frequent data use. The cost of data deposit, as quoted by the appropriate repository for the type of data generated by the study, should be included in project proposal budgets to ensure that data will be archived properly in suitable repositories. Data repositories need to have long-term support for storage infrastructure with the ability to adapt to different data needs and emerging technology. Depending on funding, some repositories could themselves facilitate data integration and meta-analysis rather than relying on individual researchers to do this (and if they do, the integrated or harmonized data should also be shared).

2.3.3 | Managing inconsistent data repository standards

Different data repositories can follow different practices (e.g. disparate requirements for ontologies or data dictionaries or acceptance criteria) that result in inconsistent and evolving data standards. For example, NCBI no longer accepts non-human genetic variation data as this type of data has become increasingly voluminous and challenging to manage. Because of potential repository data requirement changes, researchers often want assurances that data from their long-term studies will continue to be accepted without major requirements to modify data dictionaries or ontologies. Transparency from data repositories on their data requirements and criteria will help encourage data submission from researchers performing long-term studies. Many data users and researchers indicate that the user community does not have much leverage over current practices, but the ability for users to choose between different repositories will help them identify the most appropriate requirements for their data. International databases do not all have the same definitions for different terms, but consistent data dictionaries are important to ensure that data is not misused. For example, the World Bank (https://www.worldbank.org/) and the Organization for Economic Co-operation and Development (OECD; http://www.oecd.org/) do not have the same definition of foreign direct investment.

It is important for users to know the database or repository and its standards in order to correctly analyze data and interpret results. Given that many databases are multinational, clearly defined metadata schemas and emphasis on common terms, consistent data dictionaries and units of measure should be a priority. Currently, the Consultative Group for International Agricultural Research (CGIAR) is working to lead the development of metadata standards and ontologies (Arnaud et al., 2020). A useful best practice for integrating data from different sources and repositories is the use of smart templates that can check for accuracy and validity of data inputs, and identify common variables that can be used to link multiple datasets.

In general, it is a best practice that research data should not be archived in proprietary formats. However, for data that can only be interpreted through a proprietary platform, all attempts should be made to provide at least some of the data in non-proprietary formats or the data submitter should provide software that allows others to use the closest approximation to the proprietary-formatted data. In any case, data format information must be included in metadata to ensure interpretability, keeping in mind that if the data is stored in a form that is difficult to use, then potential users will not use the data.

2.3.4 | Dataset appraisal

As the volume of research and observational data increases, repositories may increasingly need methods for conducting appraisals for scientific importance in order to ensure that they are accepting and keeping high value data. This is not peer review if the appraisers are not peer researchers but repository managers. The number of citations and reads for a journal article associated with a potential dataset, or for a published dataset itself could be a usefulness score. Data appraisal performed by repositories could be improved if researchers provide examples in the metadata description of how their submitted data is or could be useful to others.

Replaceability should also be an important metric for assigning value to datasets to determine suitability for long-term preservation. For example, sequence data is cheaper to produce than preserve, therefore storing the physical sample could be better than storing the sequence data. The USDA National Agricultural Library (https://www.nal.usda.gov/) and the University of Maryland (UMD; https://umd.edu/) are collaborating on data rescue protocols to create rubrics to help determine when the value of data is great enough to justify the costs of rescue and preservation (Shiue et al., 2021).

2.4 | Best practices for citizen science in agricultural research

Citizen science (also known as community or participatory science), including crowdsourcing, provides substantial opportunity to increase observational data collection, inform model development, and increase engagement between agricultural researchers and people who are not

trained scientists, including farmers. For example, on-farm replicated strip trial research can be performed to evaluate the impact of different practices and products on productivity (Kyveryga et al., 2018), and precision agriculture technology allows for enhanced data collection. Precision agriculture technology now allows farmers and scientists to collect GIS coordinates in concert with agricultural data that can be used for precisely selecting varieties, fertilizer and water needs, and pesticide application strategies (Fulton & Port, 2018). New technologies such as these also come with greater data management challenges.

In the case of farmer-scientist research partnerships, scientists may be reluctant to share data that is owned by the farmer. Farmers may not be able to access, much less share, their own data due to issues with proprietary software and hardware. Additionally, for research conducted in partnership, "dual ownership" of data can cause confusion. We recommend creating data management plans and data sharing agreements before projects start to avoid such challenges.

Because of the differences in data collection methods, citizen science repositories, or more typically databases, must be evaluated differently than traditional scientific repositories. Citizen science repositories can have additional value that traditional repositories cannot necessarily provide: very large sets of observations from many citizen data contributors present the ability to identify likely outliers and unbiased trends; citizens often have little reason to lie unless they have a vested interest (e.g., siting a new industrial plant). Many citizen scientists engage with the work out of a deep passion for a particular subject. For example, the public produces butterfly monitoring data (i.e., eButterfly, http://www.e-butterfly.org/) whose contributors create accounts that show who they are and where they live. eButterfly communicates to their users to inform them what the data has been used for, what researchers have learned from the data, and also asks for citizen feedback when the database tools are updated. Giving citizen participants the option to use the data themselves enhances participation in data collection. It is important that the data collection and submission process is not too complicated, or participants will not take the time to submit data. However, educating participants to use good data collection practices, and building platforms that follow existing standards will raise the quality of the data for subsequent analysis, such as modeling.

In agricultural research, it would be most beneficial to increase the use of citizen science with farmers as the participants, as farmers have a vested interest in agricultural research directions and results. Because of these vested interests, this type of citizen science tends toward knowledge coproduction (Cash et al., 2003; Djenontin & Meadow, 2018) because farmers can provide valuable knowledge, including but not limited to data collection, throughout all stages of the research process. However, agricultural knowledge coproduction often operates within (or is limited by) long-standing institutional systems for translating scientific research into usable information for farmers. In the United States, agricultural extension agents at land-grant colleges and universities act as an educational resource for agricultural producers and rural communities; translating scientific research into usable information. Additionally, the Natural Resources Conservation Service (NRCS; https://www.nrcs.usda.gov/) brings scientific research to agricultural producers through technical assistance in a range of areas such as soil health, water quality, conservation, and livestock management. While these institutional systems are very helpful to agricultural producers in many ways, there is still a disconnect between the agricultural scientist producing the research and the farmer or rancher who is often the beneficiary or end-user of the research results. This disconnect limits the agricultural scientist's ability to understand producer perspectives ("where the farmers are coming from") and can yield research results that are not seen as legitimate by producers due to lack of trust.

Lack of trust in research results is especially prevalent in agricultural model development in which the producers were not involved in the model creation (e.g., empirical, process, predictive, risk assessment, etc.). Models can be highly useful for crop analysis, but unfortunately, model results are often not trusted by agricultural producers. Older models and their results are often lost as revisions are made, and model improvements can change the predictions and subsequent recommendations to farmers. For farmers to accept new model findings there must be greater transparency to show how the model changes have impacted the predictions. Based on their own experiences, DIDAg participants acknowledged that farmers often don't trust management decisions made from sensor output, or by the people who are collecting the data (government or scientists), unless there is a strong local agricultural extension presence. The farmer–scientist data collection process should include a trusted advisor, such as an extension specialist, NRCS staff member, or local expert such as a certified crop advisor (CCA) to improve transparency during the collaboration.

Participatory research can give farmers a bigger role in shaping the research development process beyond just contributing data. Initial participants can be found through enhanced outreach by extension agents at field days and show-and-tell events. Continued outreach and education by Extension will help build valuable participatory research communities. In many areas participatory research is conducted by industry, such as the testing of precision agricultural methods (Fulton & Port, 2018). Precision ag is a smaller part of a larger emerging structure named smart farming, or smart farming technologies (SFT) that integrates data into farming practices (Balafoutis et al., 2017). As more farmers adopt "smart" techniques, further trust and strategies to share data are needed in the agricultural community. Increased engagement of farmers in the research process through

participatory methods will improve the perceived legitimacy of research results, and improve trust between farmers, Extension, and researchers. Farmer engagement with researchers will also yield research that is more relevant to farmers and their operations. A valuable best practice for facilitating farmer and citizen participation is to make participation simple and not be overly time consuming. Using and digitizing farmer and citizen analog data sets will help expand participation of individuals who are on the other side of the data divide. Understanding the participant community is important for these projects; using short survey instruments to gather demographic data and gauge participant knowledge can benefit the project and also situate citizens in the science.

A very positive example of successful citizen science that resulted in community best practices is the Wheat Stripe Rust disease effort (Kolmer, 2005). In this case, the citizen scientists were farmers, whose boots were on the ground quickly to meet this time-sensitive challenge of the rapidly spreading Stripe Rust disease. Farmer participation had clear value in fighting the spread of the Stripe Rust disease that was harming crop production, and many data points were collected from a broad range of participants. Agricultural extension offices taught the farmers who wanted to participate how to contribute data, and those farmers taught other farmers. The RustTracker website (CIMMYT, 2020) was made available for the data to be input, and early identification of the disease was made in order for it to be contained. In particular, successful efforts were made to contain Stripe Rust disease in the Walla Walla Valley of Washington and Oregon. The Strip Rust disease effort revealed a number of points that can be used to improve future agricultural citizen science efforts:

- Many farmers and participants are close to retirement and their knowledge will be lost when they leave the profession.
- Larger corporate farms don't participate in citizen science efforts.
- Collaboration between Extension and Researchers doesn't happen as often as it should, often due to the reduction in the number of Extension agents.
- Better communication is needed between Extension and Natural Resource Conservation Service.

Citizen science collaboration, participation, satisfaction, and data literacy could be enhanced by building a platform (like eButterfly) that can link Extension with farmers and researchers to make Extension fully available. Successful use cases often involve tracking the spread of crop diseases because they impact a farmer's bottom line. Expanding effective infrastructure like the Stripe Rust example could be used for future agricultural disease citizen science (e.g., citrus greening). However, there are many more areas that are ripe for participatory research methods. For example, water management could also be tracked to improve the bottom line for farmers, as irrigation is an important common management decision. Additionally, farmers may be likely to participate in data collection for agricultural economics models if an output of those models could give them predictions to improve yields and profit. Also, citizens could be involved in agricultural product data collection such as milk production data, but data collection and management protocols for these data would need to be fully developed.

2.5 | Supporting agricultural community best practices

There are many common practices that result in poor data management and minimal public access to publicly funded research data, such as storing analog data in notebooks, storing data on hard drives that are not publicly accessible, or publishing in journals that do not require data public access. Many scientists believe that they own the data that they have collected and that sharing the data reduces opportunities to use the data in future analyses and scientific papers. When data is jointly owned by farmers and scientists, many farmers are concerned that sharing the data may open them up to future lawsuits or that they are sharing information with their competitors. Agricultural research is often jointly funded by the public and industry, which complicates data sharing and public access. However, significant progress has been made to expand best practices in data management and public access. The first level of support for agricultural community best practices of data management and public access is with funding agencies when they require explicit public access policies and digital scientific data standards in order to secure and maintain funding. Positive incentives such as citations, awards, and credit are also very desirable best practices for encouraging data public access. Publishers can require community standards as well as requiring or encouraging datasets to be publicly accessible after publication. Additionally, journals should require researchers to cite all data sources that are used. CrossRef and Scholix make it easier to track the use of data and give it more value. Third parties could be contracted for independent assessment and enforcement of data public access including data repositories or regulatory bodies, if adequate funding is available. It is important for experienced scientists to inform their younger colleagues that some results accepted by a research community may not be evidence-based. Strong evidence is crucial for decision-making and developing policy.

Data dictionaries and data standards are essential for other researchers to interpret datasets and correctly analyze them for future use, and these are often promoted by data repositories. Librarians could assist in the choice of these dictionaries and metadata standards and repositories, where standards exist. Standard languages and data formats should be

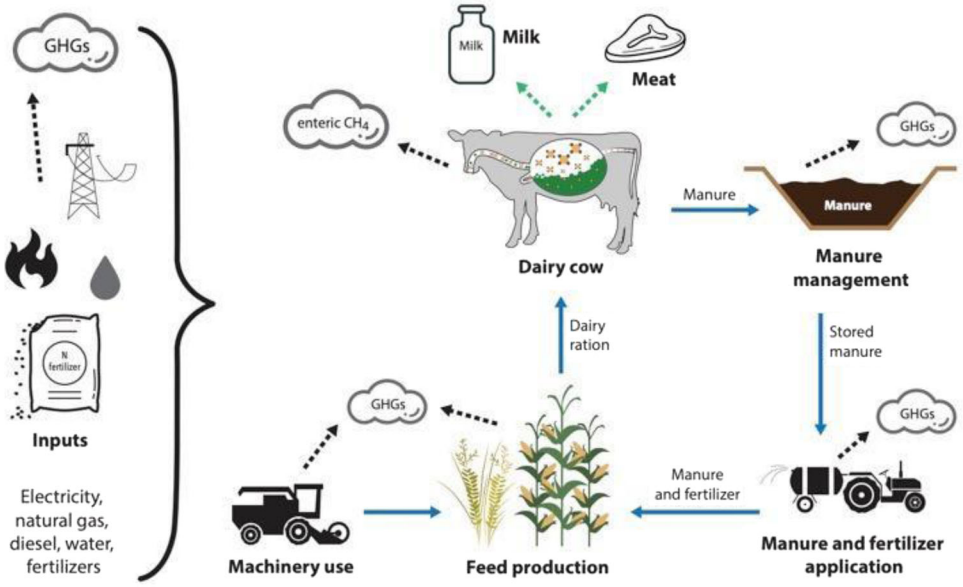

FIGURE 2 Understanding sustainability and impacts of complex food systems for dairy products requires data for inputs and outputs at all stages, including large numbers of processing steps when going from dairy cows to the products made from their milk. Modified from Aguirre-Villegas et al., 2018)

defined for different data types or disciplines, and researchers need to incorporate these in their metadata. Acceptable data formats and data dictionaries can be defined by the funding agency, a standard-making body, or the repository where the data will be stored. Clear descriptions of data dictionaries and standards should be done at the beginning of the workflow description in the metadata. Policies should be put in place that describes what process will be used to manage inconsistencies in data dictionaries for long- term curation.

Long-term data standardization of community best practices should be achieved through collaboration and communication within a given discipline. CGIAR, an organization that includes 15 international research centers, is developing agricultural data standards to create their own associated ontology with input from many sources. The Research Data Alliance Interest Group on Agricultural Data promotes dialog across international agricultural research communities. Community efforts help build momentum for data standardization, resulting in community expectations and requirements that data standards must be met to warrant publication.

University leadership support for agricultural community best practices for data management and public access is crucial. This is why both Association of Public & Land-grant Universities (APLU; https://www.aplu.org/) and Association of American Universities (AAU – i.e., private universities; https://www.aau.edu/) are engaged on this topic. Deans, Provosts, and Department Heads can influence their departments and communities to adopt data management and public access best practices, in partnership with their research libraries (Chodacki et al., 2020). Currently, a lot of money is going into data creation via precision agriculture but much of it is private. Competitions such as the Gates Foundation Grand Challenges (https://gcgh.grandchallenges.org/about) competition using Microsoft FarmBeats (https://www.microsoft.com/en-us/research/project/farmbeats-iot-agriculture/) data can encourage faculty to take advantage of data, public or private, in innovative ways. Success of these data-intensive researchers can engage the leadership of their colleges and universities, who in turn can use this success to illustrate the benefits of community best practices among the rest of their faculty. New types of data are now available that agricultural researchers never dreamed of in the past, and the research community and university leadership are beginning to recognize the value of that data for promoting the public good.

3 | CASE STUDY: DATA BEST PRACTICES AND DIET AND GENETICS IMPACTS ON DAIRY CATTLE GREENHOUSE GAS EMISSIONS

Supporting data management best practices will help the dairy sector translate research and industry results into policy and productive options for farmers. The USDA-ARS "Dairy Agriculture for People and the Planet" Grand Challenge Synergy Project aims to improve the availability of safe and nutritious dairy products, and decrease the environmental impact of dairy production (Tricarico et al., 2019). Additionally, new innovations in diet and genetics are being developed to reduce greenhouse gas emissions in the dairy industry and the greater livestock industry (Beauchemin et al., 2009; Boadi et al.,

2004; Buddle et al., 2011) (Figure 2). A food systems analysis of the diary sector recognizes that data is a significant constraint towards integrating the different disciplines, spatial and temporal scales, and multiple vocabularies that is necessary for reducing the environmental impact of dairy production (Finley & Fukagawa, 2019). In order to take advantage of emerging greenhouse gas remote sensing technologies, Geenhouse gas Reduction through Agricultural Carbon Enhancement network (GRACEnet) (Jawson et al., 2005)] can provide examples of minimal greenhouse gas dataset development as a foundation for remote sensing data. Industry data repositories are a valuable historical data resource to use in concert with research data repositories, and developing new crowdsourcing techniques will help enhance consumer data among younger generations. Further discussion is available in Additional Online Materials.

4 | CONCLUSIONS AND RECOMMENDATIONS

Agricultural data is crucial to many aspects of production, commerce, and research involved in feeding the global community. However, standard best practices for agricultural research data management and publication do not exist given the wide range of disciplines associated with agriculture. Support for agricultural community best practices should come from funders, institutions, and organizations; the support from these entities will facilitate faster adoption of best practices data management by researchers.

A wide range of best practices identified by DIDAg participants could replace data management common practices and improve data-intensive research in agriculture. The following key recommendations emerged from the DIDAg workshops to improve data management without overburdening agricultural researchers and data repositories. (a) Peer review is important for ensuring quality data publication. Broad-purpose repositories can enable peer review, whether it happens before or after publication, by ensuring that adequate metadata is present, particularly regarding collection and analysis methodology. (b) Minimal dataset development that includes detailed metadata and data dictionaries is a crucial best practice that should be adopted by agricultural sub-disciplines. Agricultural research communities should develop minimal dataset requirements that will make archived data more useful and interoperable for researchers within and across different disciplines or locations. (c) Agricultural researchers should use data repositories that provide long-term data preservation and consistent collection criteria and other standards. Appraisal processes should be used to ensure high standards of data quality and value for data going into these repositories. (d) Funding agencies, scientific journals and other publications, and university leadership are crucial partners that should be centrally involved in promoting agricultural community standards and best practices. Funding agencies should encourage and provide funding for data publication and archiving in trusted data repositories. (e) Citizen science has a strong potential to drive innovation in agricultural research by generating new or improved observational datasets, improving the salience and perceived legitimacy of research results, and building trust between researchers, Extension, and agricultural producers. More citizen science and participatory research efforts, involving farmers in particular, should be pursued by agricultural researchers.

ACKNOWLEDGMENTS

We thank the National Agricultural Library for hosting the first DIDAg Workshop in June, 2018 and we thank Virginia Tech for hosting the second DIDAg Workshop in August 2019. This work was supported in part by the U.S. Department of Agriculture National Institute for Food and Agriculture AFRI FACT no. 2018-67023-27843 and by USDA Agricultural Research Service no. 8070-13000-014-00D and 8260-88888-001-00D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Eli K. Moore https://orcid.org/0000-0002-9750-7769
Adam Kriesberg https://orcid.org/0000-0002-9240-4998
Steven Schroeder https://orcid.org/0000-0001-9103-5150
Kerrie Geil https://orcid.org/0000-0001-7604-3267
Inga Haugen https://orcid.org/0000-0002-9414-9097
Carol Barford https://orcid.org/0000-0002-4941-8710
Erica M. Johns https://orcid.org/0000-0002-3068-3143
Dan Arthur https://orcid.org/0000-0001-6471-0149
Megan Sheffield https://orcid.org/0000-0002-4030-2674
Stephanie M. Ritchie https://orcid.org/0000-0002-0054-4409
Carolyn Jackson https://orcid.org/0000-0001-5450-9716
Cynthia Parr https://orcid.org/0000-0002-8870-7099

REFERENCES

Aguirre-Villegas, H. A., Kim D., Thoma G., Larson R. A., & Ruark M. D. (2018). Life cycle assessment of greenhouse gas emissions from dairy farms in the great lakes region *(Sustainable dairy fact sheet series* A4131-12/GWQ084). Madison, WI: University of Wisconsin-Extension.

Arnaud, E., Laporte M. A., Kim S., Aubert C., Leonelli S., Miro B., Cooper L., Jaiswal P., Kruseman G., Shrestha R., Buttigieg P. L., Mungall C. J., Pietragalla J., Agbona A., Muliro J., Detras J., Hualla V., Rathore A., Rani Das R., … King, B. (2020). The ontologies community of practice: A CGIAR initiative for big data in

agrifood systems. *Patterns*, *1*(7), 100105. https://doi.org/10.1016/j.patter.2020.100105

Assante, M., Candela L., Castelli D., & Tani A. (2016). Are scientific data repositories coping with research data publishing? *Data Science Journal*, *15*(0), 6. https://doi.org/10.5334/dsj-2016-006

Balafoutis, A. T., Beck B., Fountas S., Tsiropoulos Z., Vangeyte J., van der Wal, T., Soto, I., Gómez-Barbero, M., & Pedersen, S. M. (2017). Smart farming technologies – Description, taxonomy and economic impact. In S. M. Pedersen & K. M. Lind (Eds.), *Precision agriculture: Technology and economic perspectives* (pp. 21–77). Berlin: Springer International Publishing.

Beauchemin, K. A., McAllister T. A., & McGinn S. M. (2009). Dietary mitigation of enteric methane from cattle. *CAB Reviews: Perspectives in Agriculture, Veterinary Science*. *Nutrition and Natural Resources*, *4*(035), 1–18. https://doi.org/10.1079/PAVSNNR20094035

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from The American Economic Review. *The American Economic Review*, *81*(5), 1041–1067

Boadi, D., Benchaar C., Chiquette J., & Massé D. (2004). Mitigation strategies to reduce enteric methane emissions from dairy cows: Update review. *Canadian Journal of Animal Science*, *84*(3), 319–335. https://doi.org/10.4141/A03-109

Buddle, B. M., Denis M., Attwood G. T., Altermann E., Janssen P. H., Ronimus R. S., Pinares-Patiño C. S., Muetzel S., & Wedlock D. N. (2011). Strategies to reduce methane emissions from farmed ruminants grazing on pasture. *Veterinary Journal*, *188*(1), 11–17. https://doi.org/10.1016/j.tvjl.2010.02.019

Cash, D. W., Clark W. C., Alcock F., Dickson N. M., Eckley N., Guston D. H., Jäger J., & Mitchell R. B. (2003). Knowledge Systems for Sustainable Development. *Proceedings of the National Academy of Sciences*, *100*(14), 8086–91. https://doi.org/10.1073/pnas.1231332100

Chodacki, J., Hudson-Vitale, C., Meyers, N., Muilenburg, J., Praetzellis, M., Redd, K., Ruttenberg, J., Steen, K., Cutcher-Gershenfeld, J., & Gould, M. (2020). *Implementing effective data practices: Stakeholder recommendations for collaborative research Support*. Washington, DC: Association of Research Libraries. https://doi.org/10.29242/report.

CIMMYT. (2020). RustTracker.org –A global wheat rust monitoring system. Retrieved 23 Sept. 2020 from https://rusttracker.cimmyt.org/

Djenontin, I. N. S., & Meadow A. M. (2018). The Art of Co-Production of Knowledge in Environmental Sciences and Management: Lessons from International Practice. *Environmental Management*, *61*(6), 885–903. https://doi.org/10.1007/s00267-018-1028-3

Finley, J. W., & Fukagawa N. K. (2019). Integrated data across multiple and diverse disciplines are essential for developing a sustainable food system. *Journal of Soil and Water Conservation*, *74*(6), 632–638. https://doi.org/10.2489/jswc.74.6.632

Fulton, J. P., & Port K. (2018). Precision agriculture data management. In D. K. Shannon, D. E. Clay, & N. R. Kitchen (Eds.), *Precision agriculture basics* (pp. 169–187). New York: John Wiley & Sons, Ltd.

Jawson, M. D., Shafer S. R., Franzluebbers A. J., Parkin T. B., & Follett R. F. (2005). GRACEnet: Greenhouse gas reduction through agricultural carbon enhancement network. *Soil Tillage Research*, *83*(1), 167–172. https://doi.org/10.1016/j.still.2005.02.015

Kitchen, N. R., Shanahan J. F., Ransom C. J., Bandura C. J., Bean G. M., Camberato J. J., Carter P. R., Clark J. D., Ferguson R. B., Fernández F. G., Franzen D. W., Laboski C. A. M., Nafziger E. D., Qing Z., Sawyer J. E., & Shafer M. (2017). A public–industry partnership for enhancing corn nitrogen research and datasets: Project description, methodology, and outcomes. *Agronomy Journal*, *109*(5), 2371–2389. https://doi.org/10.2134/agronj2017.04.0207

Kolmer, J. A. (2005). Tracking wheat rust on a continental scale. *Current Opinion in Plant Biology*, *8*(4), 441–449. https://doi.org/10.1016/j.pbi.2005.05.001

Kriesberg, A., Huller K., Punzalan R., & Parr C. (2017). An analysis of Federal policy on public access to scientific research data. *Data Science Journal*, *16*, 27. https://doi.org/10.5334/dsj-2017-027

Kundzewicz, Z. W., & Koutsoyiannis D. (2005). Editorial–The peer-review system: Prospects and challenges. *Hydrological Sciences Journal*, *50*(4), 1. https://doi.org/10.1623/hysj.2005.50.4.577

Kuru, T. H., Wadhwa K., Chang R. T. M., Echeverria L. M. C., Roethke M., Polson A., Rottenberg G., et al. (2013). Definitions of Terms, Processes and a Minimum Dataset for Transperineal Prostate Biopsies: A Standardization Approach of the Ginsburg Study Group for Enhanced Prostate Diagnostics. *BJU International*, *112*(5), 568–77. https://doi.org/10.1111/bju.12132

Kyveryga, P. M., Mueller T. A., & Mueller D. S. (2018). On-farm replicated strip trials. In D. K. Shannon, D. E. Clay, & N. R. Kitchen (Eds.), *Precision agriculture basics* (pp. 189–207). New York: John Wiley & Sons, Ltd.

Lane, D., Murdock E., Genskow K., Betz C. R., & Chatrchyan A. (2019). Climate change and dairy in New York and Wisconsin: Risk perceptions, vulnerability, and adaptation among farmers and advisors. *Sustainability*, *11*(13), 3599. https://doi.org/10.3390/su11133599

Langfeldt, L. (2006). The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation*, *15*(1), 31–41. https://doi.org/10.3152/147154406781776039

Massey, C. G. (2014). *Creating linked historical data: An assessment of the Census Bureau's ability to assign protected identification keys to the 1960 Census*. Working paper number CARRA-WP-2014-12. Washington, DC: U.S. Census Bureau: Center for Economic Studies.

Northey, S. A., Mudd G. M., Saarivuori E., Wessman-Jääskeläinen H., & Haque N. (2016). Water footprinting and mining: Where are the limitations and opportunities? *Journal of Cleaner Production*, *135*, 1098–1116. https://doi.org/10.1016/j.jclepro.2016.07.024

Office of Science and Technology Policy (OSTP). (2013). *Increasing access to the results of Federally-funded scientific research*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Shiue, H. S. Y., Clarke, C. T., Shaw, M., Hoffman, K. M., & Fenlon, K. (2021). Assessing Legacy Collections for Scientific Data Rescue. In *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16* (pp. 308–318). Springer International Publishing. https://link.springer.com/book/10.1007/978-3-030-71305-8

Schwartz, P. M., & Solove D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. *New York University Law Review*, *86*(6), 1814–1894.

Tenopir, C., Dalton E. D., Allard S., Frame M., Pjesivac I., Brich, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8). https://doi.org/10.1371/journal.pone.0134826

Tenopir, C., Rice N. M., Allard S., Baird L., Borycz J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists

worldwide. *PLOS ONE*, *15*(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Tricarico, J. M., Slimko M. L., Graves W. B., Eve M. D., & Thurston J. A. (2019). Elevating dairy research and extension through partnership: Outcomes from the United States Department of Agriculture and National Dairy Council collaborative meeting to develop a coordination roadmap. *Journal of Dairy Science*, *102*(10), 9518–9524. https://doi.org/10.3168/jds.2019-16579

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, *92*(1), 16–24. https://doi.org/10.3168/jds.2008-1514

Van Tassell, C. P., Smith T. P. L., Matukumalli L. K., Taylor J. F., Schnabel R. D., Lawley C. T., Haudenschild C. D., Moore S. S., Warren W. C., & Sonstegard T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, *5*(3), 247–252. https://doi.org/10.1038/nmeth.1185

Veltman, K., Rotz C. A., Chase L., Cooper J., Ingraham P., Izaurralde R. C., Jones C. D., Gaillard R., Larson R. A., Ruark M., Salas W., Thoma G., & Jolliet O. (2018). A quantitative assessment of Beneficial Management Practices to reduce carbon and reactive nitrogen footprints and phosphorus losses on dairy farms in the US Great Lakes region. *Agricultural Systems*, *166*, 10–25. https://doi.org/10.1016/j.agsy.2018.07.005

Whitmire, A., Carlson J., Westra B., Hswe P., & Parham S. (2017). The DART Project: using data management plans as a research tool. https://doi.org/10.17605/OSF.IO/KH2Y6

Wolfert, S., Ge L., Verdouw C., & Bogaardt M.-J. (2017). Big data in smart farming – A review. *Agricultural Systems*, *153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Moore EK, Kriesberg A, Schroeder S, et al. Agricultural data management and sharing: best practices and case study. *Agronomy Journal*. 2021;1-11. https://doi.org/10.1002/agj2.20639