CAST Commentary
QTA2019-1 March 2019

Enabling Open-source Data Networks in Public Agricultural Research

Authors:

Sylvie Brouder (Chair)
Purdue University
West Lafayette, Indiana

Alison Eagle
Environmental Defense Fund
Raleigh,
North Carolina

Naomi K. Fukagawa
USDA–ARS
Beltsville, Maryland

John McNamara (retired)
Washington State University
Pullman

Seth Murray
Texas A&M University
College Station

Cynthia Parr
USDA–ARS
Beltsville, Maryland

Nicolas Tremblay
Agriculture and Agri-Food Canada
St-Jean-sur-Richelieu, Canada

Reviewers:

Marianne Stowell Bracke
Whitworth University
Spokane, Washington

Paul Fixen (retired)
International Plant Nutrition Institute
Brookings, South Dakota

Jeffrey Volenec
Purdue University
West Lafayette, Indiana

CAST Liaison: **Drew Lyon**
Washington State University
Pullman

Abstract

The next generation of agricultural problem solving will require big science and linkages forged across data sets and disciplines. Currently, a lack of data sharing and data accessibility is a major barrier for making better decisions in agriculture. Business cases for data-sharing infrastructure include that pooling datasets and computational power efficiently extends sparse data resources, facilitates new discovery, derives better answers and decision making, lowers the barrier of entry, and ensures scientific reproducibility so that U.S. production agriculture can compete sustainably.

Immediate imperatives for facilitating data sharing to fully realize open access to public agricultural research are the following: (1) development and implementation of best practices for data—workflows and standards—in all future federally funded projects; (2) incentives and mechanisms for making available data not represented in the peer-review literature (grey and

Currently, a lack of data sharing and data accessibility is a major barrier for making better decisions in agriculture.

Photo source: Shutterstock—biDaala_studio (left); Semisatch (center); Power Up (right)

This publication was made possible through funding provided by the United Soybean Board ("USB"). As stipulated in the Soybean Promotion, Research, and Consumer Information Act, USDA's Agricultural Marketing Service ("AMS") has oversight responsibilities for USB. AMS prohibits the use of USB's funds to influence legislation and/or to influence governmental policy or action. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of USB, USDA, and/or AMS.

dark data); (3) coordination among existing and emerging data initiatives, networks, and repositories; and (4) dedicated and sustainable infrastructure—hardware, software, and human resources—to curate, preserve, and add value to data beyond the primary use for which they were collected.

Agriculture's pathway forward requires dedicated partnering among domain researchers, data scientists, science administrators and agencies, professional societies, and private publishing entities. To simultaneously achieve sustained and equitable data access, the authors suggest the most promise lies with a novel business model in which funding agencies pay directly for stewardship in proportion to grant volume. Further, they propose four major institutional strategies to advance data-driven research in agriculture: bridging gaps, reorienting institutions, leveraging assets, and connecting feedbacks. Teams must bridge expertise gaps through meaningful collaborations between agricultural researchers and data scientists. Institutions will need to reorient to prioritize team science and data sharing over smaller scale, individual efforts and to infuse an understanding of data sciences into curricula and learning outcomes. Initiatives to leverage assets should focus on surfacing grey/dark data not represented by peer-review publication, including high-value legacy datasets for which time and cost prohibit replication. Finally, for research data to achieve and maintain public value, it must connect feedbacks to ensure data are useful and useable for informing the end-user "apps" designed to enhance and secure our current food supply and address environmental and social challenges.

Agriculture's pathway forward requires dedicated partnering among domain researchers, data scientists, science administrators and agencies, professional societies, and private publishing entities.

Introduction

Research has created the most efficient food production system in history through accrual of massive amounts of data, information, and knowledge. The amount of research data collected to date, however, pales when compared to current ability to generate or compile data using an array of digital tools. Rapidly accruing datasets, each containing ever-larger quantities of data, have led to concepts such as "big data," "data sciences," and "data analytics." Yet, with much research data remaining unpublished, only partially available, or incompletely described, policy decisions and program design may lean disproportionately on expert opinion and partial information. The complex interconnections between agriculture, the natural environment, and social and physical well-being increase the need for researchers to use the full suite of data, but, for this to happen, access to data and analytical tools must be relatively unimpeded and open to all working within the scientific, nongovernmental organization, government, and business fields (Figure 1; Textbox 1). The idea that information, including all data, collected from publicly funded scientific activities belongs to the public and should be freely available and usable motivates this commentary.

The amount of research data collected to date . . . pales when compared to current ability to generate or compile data using an array of digital tools.

The idea that information, including all data, collected from publicly funded scientific activities belongs to the public and should be freely available and usable motivates this commentary.

For agriculture, the scope of opportunities and challenges linked to data is hard to overstate. Recent analyses suggest scientists have reached near universal agreement that data sharing has value and advances research toward solutions to complex problems (e.g., Kim and Stanton [2016] and references cited therein). Yet, current approaches to research design and data collection are rarely standardized across studies, even within narrowly focused disciplines. Data access and use by others remains dependent on individual agreements, facilitated by one-time trial-and-error solutions for data transfer (Cragin et al. 2010); a lack of funding for synthesis research using aggregated data remains a significant barrier. Issues of data privacy, security, and intellectual property further constrain nascent data-sharing efforts, especially those in which public-private partnerships are involved. The underlying problem is a continued absence of a coordinated infrastructure of equipment and people to support agricultural research data sharing and its routine synthesis into practice and policy.

In the United States, this infrastructure deficit threatens agriculture's ability to comply with "open access" mandates (Holdren 2013) and proposed legislation. The Federal Research Public Access Act (FRPAA 2012) succeeded by the Fair Access to Science Technology Research Act (FASTR 2017) require funding agencies to develop public access policies, reflecting increased public pressure for transparency in science use. While FRPAA guidelines do not define "free online public access," most federal funding agencies have interpreted open access to be inclusive of all nonproprietary, nonsensitive data collected in their sponsored research. Free and open access to information generated by federal funding is clearly in the spirit of the original legislation creating the U.S. Department of Agriculture (USDA) and the land-grant university system to develop and apply scientific knowledge in food production for the betterment of the U.S. population. Public expectations for accessing and using agricultural science to make informed choices span a myriad of health, land management, and lifestyle issues, from understanding food's nutritional value for diet modification to vastly improving application of precision technologies for profitable crop and animal production and environmental protection.

Public expectations for accessing and using agricultural science to make informed choices span a myriad of health, land management, and lifestyle issues.

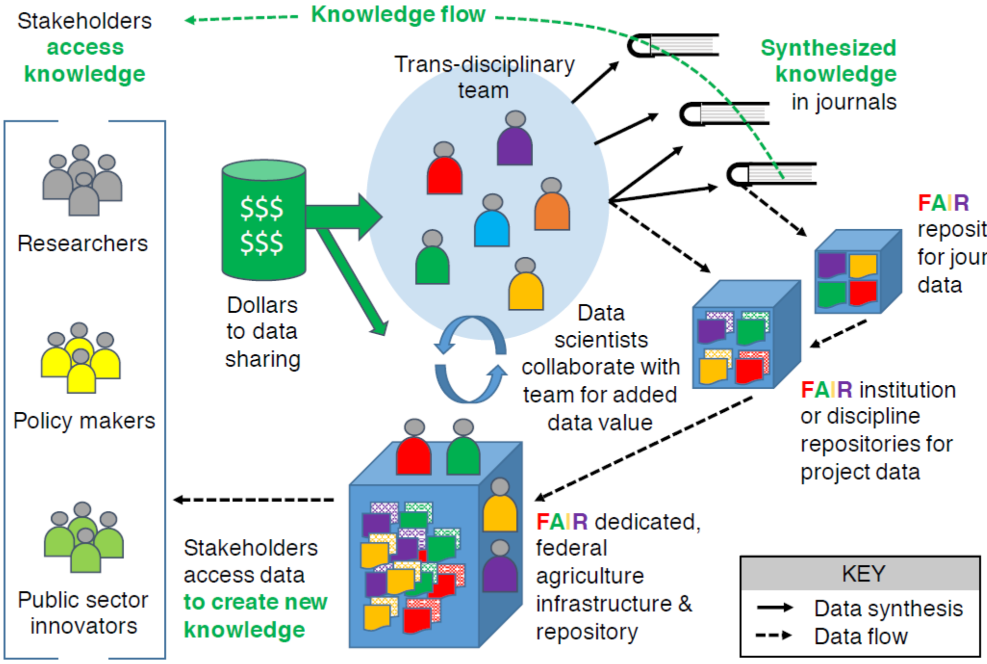

Stakeholders access knowledge
Knowledge flow
Trans-disciplinary team
Synthesized knowledge in journals
Researchers
$$$ $$$
Dollars to data sharing
FAIR repository for journal data
Policy makers
Data scientists collaborate with team for added data value
FAIR institution or discipline repositories for project data
Public sector innovators
Stakeholders access data to create new knowledge
FAIR dedicated, federal agriculture infrastructure & repository
KEY
Data synthesis
Data flow

Figure 1. The data ecosystem for team science (see Textbox 1).

Textbox 1: The Data Ecosystem for Team Science: Key attributes for data sharing

- All data are collected a priori anticipating reuse and in accordance with FAIR principles.
- Data synthesized in journal articles are published with and referenced in the article.
- All (unpublished and published) project data are FAIR (findable, accessible, interoperable, and reusable), residing in either disciplinary or institutional repositories and/or in a federal "knowledgebase."
- A newly created federal knowledgebase provides repository function with expert services to enhance data collections and their reuse.
- Dedicated knowledgebase experts collaborate with research teams and stakeholders to develop high-value data products including fusions from disparate repositories and knowledgebases (e.g., merge weather, soil, and agronomic data streams).
- Stakeholders in the data value chain (extension specialists, entrepreneurs, farmers, etc.) access knowledge from journals and data from the knowledgebase for innovation and new knowledge creation.

The purpose of this commentary is to document need for and anticipated benefits of developing data-sharing standards, incentivizing researchers to share data, and building a data-sharing infrastructure within agricultural research.

The purpose of this commentary is to document need for and anticipated benefits of developing data-sharing standards, incentivizing researchers to share data, and building a data-sharing infrastructure within agricultural research. The authors present the factors contributing to the current system of agricultural research that has fostered ambivalence toward data sharing; briefly review the success of data-sharing examples from other domains that offer promise for advancing agricultural research; and describe the advantages and shortcomings of emerging data-sharing platforms, networks, and repositories intended to facilitate data sharing in agriculture. Although they focus on accessing and using the full wealth of data generated by research, they realize impact from this effort also requires research in food production to de-emphasize smaller-scale, individual-effort studies and pursue larger efforts integrating social, economic, and environmental components. Thus, the ultimate goal is to advance the conversation among agricultural science partners to create a system conducive to data sharing and the team science that are needed to address the complex, “grand-challenge” questions in food systems (e.g., Bennett and Balvanera 2007; Daar et al. 2007; Mueller et al. 2012; Robertson and Swinton 2005). The authors highlight key strategies, roles, and responsibilities of partners in agriculture’s science and data enterprise, and they discuss the business case for data sharing as well as ingredients essential to data preservation and curation.

Justification: Data-driven Research Approaches for Agriculture

To date, agricultural research has generally been pursued as incremental aggregations of “small science,” hypothesis-driven research led by single researchers or small teams generating and analyzing their own results.

To date, agricultural research has generally been pursued as incremental aggregations of “small science,” hypothesis-driven research led by single researchers or small teams generating and analyzing their own results (Cragin et al. 2010). By 1935, Fisher had brought coherence to an ad hoc mixture of statistical ideas, exemplified by the work conducted at Rothamsted Experimental Station and rapidly emulated elsewhere (Speed 1992). Thereafter, Fisherian statistics dominated 20th-century thinking and methodology in agricultural research and many other disciplines (Efron 1998), embedding into curricula and scientific practice the principles of replication, randomization, analysis of variance, and elimination of heterogeneity with local control (Box 1980; Speed 1992). The scientific reward system in agriculture co-evolved with the small science model and has favored—via granting, promotion, and tenure systems—the researcher who contributes excellent, albeit modest and fragmented, knowledge. Although past research has clearly advanced agricultural productivity, some question future contributions of this small science approach (Figure 2; Textbox 2). For example, McNamara, Hanigan, and White (2016) suggested for livestock research that traditional approaches resulted in studies with narrow foci and undefinable global applicability.

Although past research has clearly advanced agricultural productivity, some question future contributions of this small science approach.

Unavoidable tensions and trade-offs in goals for system performance are hallmarks of agricultural “grand challenges.” Reconciling productivity and profitability with environmental integrity has long been considered tantamount to achieving sustainability (Davis et al. 2012; Robertson and Swinton 2005); yet among the limitations of agriculture’s small science approach is an inability to characterize such trade-offs (Caron, Biénabe, and Hainzelin 2014). Further, broader challenges concerning climate change, global hunger, food security, and societal sustainability and development goals all encompass agriculture (Campbell et al. 2017; Dobermann et al. 2013; Griggs et al. 2013). The complexity of human nutrition, natural resource sustainability, and socioeconomic problems linked to the global agricultural sector requires convergence across historically discrete disciplines and greater collaboration. The National Academies “Science Breakthroughs” report (2018a) highlights both systems research and integration of data sciences among major strategies for agriculture. A large part of future scientific inquiry may entail accessing a wealth of resources across different subject areas (Eisenhardt, Graebner, and Sonenshein 2016), requiring that scientists learn how to use data available through unstructured sources; problem solutions will entail complex optimizations with inherent uncertainties (van Mil et al. 2014). The envisioned approach is dramatically different from an agronomist simply asking an economist or other social scientist to play a cursory role on a research team.

A large part of future scientific inquiry may entail accessing a wealth of resources across different subject areas.

Concomitant with agriculture’s need for new, data-rich approaches to advance research on complex phenomena is a public mandate for credible solutions to be transparent to the underlying science (National Academies of Sciences, Engineering, and Medicine 2018b). Transparency is a core tenet of scientific endeavor and can be achieved through adhering to methodological standards that ensure repeatability and reproducibility, a “minimum necessary condition for a finding to be believable and informative” (Bollen et al. 2015). Recently, science’s commitment to these standards has been called into question by surveys citing dismal statistics for reproducibility of results (e.g., 40 and 70% failure in trying to reproduce one’s own or another researcher’s results, respectively) (Baker 2016). The biomedical and psychological literature suggest an apparent crisis in reproducibility (Harris 2017; Jarvis and

Williams 2016; Open Science Collaboration 2015; Pashler and Wagenmakers 2012; Stokstad 2018).

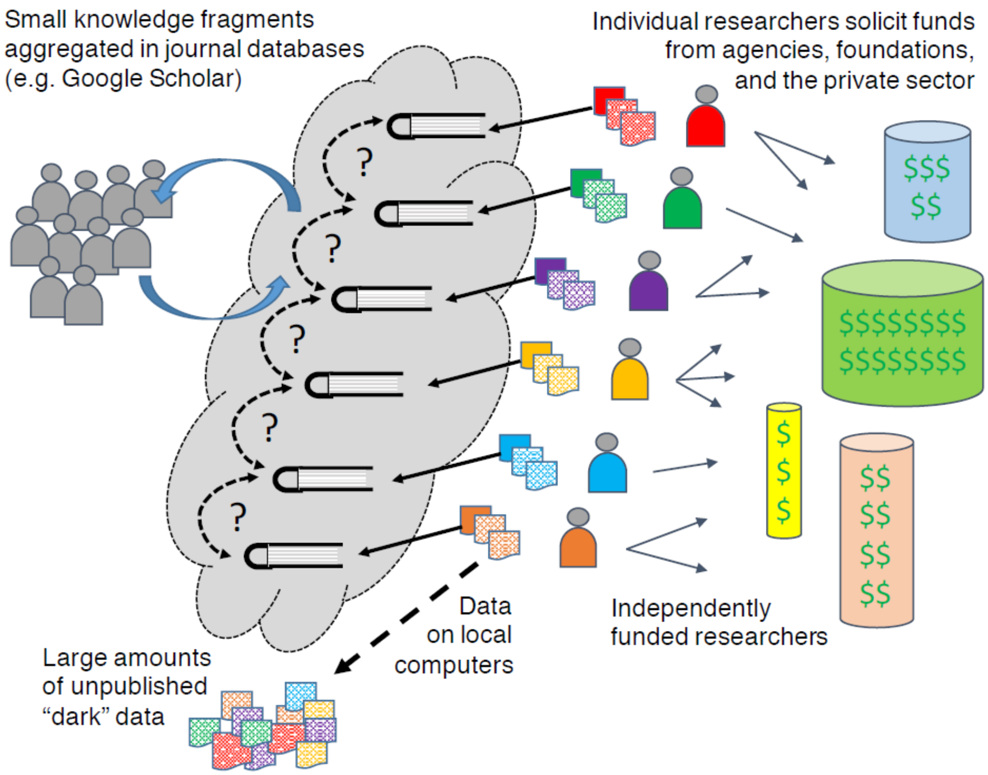

Small knowledge fragments aggregated in journal databases (e.g. Google Scholar)
Individual researchers solicit funds from agencies, foundations, and the private sector
?
?
?
?
?
$$$
$$
$$$$$$$$
$$$$$$$$
$
$
$
$$
$$
$$
$$
Data on local computers
Independently funded researchers
Large amounts of unpublished "dark" data

Figure 2. The small science environment (see Textbox 2).

Textbox 2: The Small Science Environment: Impediments to science without data sharing

- Data in an article is reproducible but may not be "open" (behind subscription paywalls).
- There are no meaningful linkages of articles and journals across disciplines and researchers.
- Research is not transdisciplinary and, thus, insufficient for "grand challenge" questions that interest donors and the public.
- Peer-reviewed journals emphasize novel results, creating bias in the scientific record.
- Unpublished data are inaccessible and often unrecoverable, representing lost research investment.
- Policy and recommendation developers pay to access fragmented, partial, and/or biased evidence to inform practice.

Professional pressures to publish have both decreased willingness to share data and increased the prevalence of smaller studies, which can lessen the likelihood of finding true, nonnull effects.

Crisis drivers include small study sizes, publication bias, and inaccessibility of original data. Professional pressures to publish have both decreased willingness to share data (Tenopir et al. 2015) and increased the prevalence of smaller studies, which can lessen the likelihood of finding true, nonnull effects (e.g., small studies with low statistical power) (Button et al. 2013). Simultaneously, journal bias toward publishing positive results (Fanelli 2012) has distorted the foundations for evidence-based practice by excluding reports of negative or null results. Although acknowledging the multitude of factors that can contribute to irreproducible results, solutions consistently stress complete reporting inclusive of data access (Begley and Ioannidis 2015; Button et al. 2013; Goodman, Fanelli, and Ioannidis 2016).

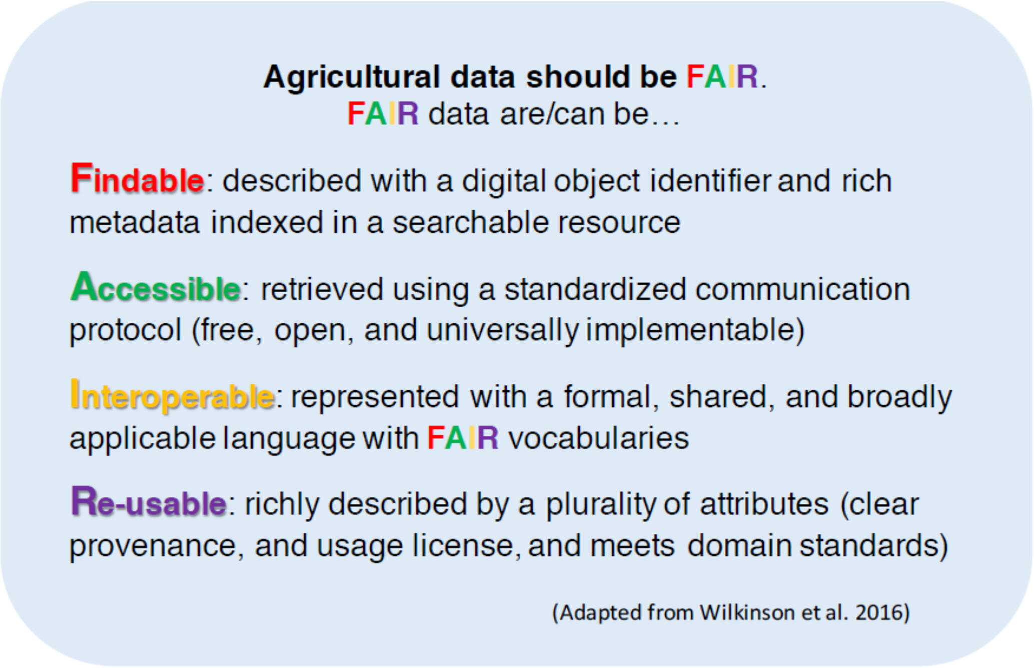
Agricultural research has not yet been highlighted for widespread procedural lapses in reproducibility, but the contributing factors exist and a culture for valuing open data has yet to be established.

Agricultural research has not yet been highlighted for widespread procedural lapses in reproducibility, but the contributing factors exist and a culture for valuing open data has yet to be established. Further, publication in the less rigorous grey literature (e.g., newsletters, reports, conference proceedings) is common when applied research is done to test or adapt the peer-review literature result for local considerations or constraints. Grey literature exacerbates transparency and reproducibility problems because it is generally harder to find and less rigorous, and persistence of both results and the data in the scientific record is notoriously poor. For reproducible science, Munafö and colleagues (2017) enumerate many advantages to data sharing in public repositories, stressing not only the transparency and openness essential to maintaining public confidence in science, but also the efficiencies that can accrue. Few expect funding allocations to agricultural research to increase dramatically in the coming years, making it even more imperative to improve efficiency of data collection and the use and reuse of existing data (see EPAR [2017] and USDA–ERS [2018] for trends in public and private support). The emergence of a host of new tools and technologies for e-sciences is potentially serendipitous for agriculture because they offer not only opportunity for higher-impact synthesis research, but also avenues for improving the overall efficiency, including creating big data and big science out of initially small efforts.

Transparent disclosure of methods and results, sharing of research materials and nonsensitive raw data, and collaboration to increase power and replicate findings will enhance reliability of many fields and increase public trust in and use of science.

In sum, transparent disclosure of methods and results, sharing of research materials and nonsensitive raw data, and collaboration to increase power and replicate findings will enhance reliability of many fields and increase public trust in and use of science. In agriculture, standardizing, organizing, and making publicly available the wide variety of specialty datasets produced in the numerous, small independent studies that typify publicly funded research is a critical first step to capitalizing on the opportunities and efficiencies afforded by e-sciences. An immediate benefit of fluent access to already-existing datasets is facilitation of meta-analyses, a powerful statistical approach for synthesizing multiple, independent studies to determine a more complete understanding of experimental results (Ehm 2016). It has been used routinely by medicine, education, and other disciplines to translate science into practice; results are generally considered robust, although outcomes can be biased by researcher decisions and judgement calls (de Vrieze 2018). Artificial intelligence also holds great promise for synthesizing agricultural data. Methods such as machine learning are tolerant to complex data characteristics (e.g., nonlinearity and outliers) and applicable to a wider range of tasks, including pattern detection and information extraction from raw data even if the underlying data model is unknown. Machine learning approaches can automatically incorporate new information, but any new data must be prepared for interoperability. As articulated by Wilkinson and colleagues (2016), the cornerstone principles for data in e-sciences are FAIR: data must be findable, accessible, interoperable, and reusable (Figure 3).

Artificial intelligence also holds great promise for synthesizing agricultural data.



Agricultural data should be FAIR.
FAIR data are/can be…

Findable: described with a digital object identifier and rich metadata indexed in a searchable resource

Accessible: retrieved using a standardized communication protocol (free, open, and universally implementable)

Interoperable: represented with a formal, shared, and broadly applicable language with FAIR vocabularies

Re-usable: richly described by a plurality of attributes (clear provenance, and usage license, and meets domain standards)

(Adapted from Wilkinson et al. 2016)

Figure 3. The FAIR principles (Wilkinson et al. 2016)

The Current Landscape for Agricultural Data and Data Sharing

Although agricultural research has been slow in developing e-infrastructure and mechanisms that promote efficiencies and transparency via open data, examples from other domains demonstrate that open data can catalyze new discoveries, decisions, and economic growth. The Cochrane Collaborative (Cochrane n.d.) is a global network of 10,000 members dedicated to improving human health outcomes through rigorous synthesis of basic medical and clinical research. The network's foundation is an open-access database of systematic reviews (SRs); the accompanying guidance on how to perform these reviews has been adopted by other domains (e.g., Collaboration for Environmental Evidence [2018]). With strong encouragement from the National Academy of Sciences, the Integrated Risk Information System of the Environmental Protection Agency recently adopted SR as a core methodology to create efficiencies and improve timeliness and responsiveness in environmental health assessment (USEPA 2017).

Although agricultural research has been slow in developing e-infrastructure and mechanisms that promote efficiencies and transparency via open data, examples from other domains demonstrate that open data can catalyze new discoveries, decisions, and economic growth.

Reports in the agricultural literature have repeatedly highlighted the potential for such infrastructure to improve the quality of the primary agricultural literature and its use in evidence-based decision making (Brouder and Gomez-Macpherson 2014; Eagle et al. 2017; Philibert, Loyce, and Makowski 2012). In neurosciences, richly diverse and important but heterogeneous and small datasets continue to be produced by individual researchers, leading to the development of infrastructure and best practices for data sharing and aggregating small data into bigger data. Neuroscience case studies provide proof-of-concept that data sharing across small, disparate research programs can successfully address larger questions and yield the advantages of big science approaches (Ferguson et al. 2014). Finally, numerous, large, data-sharing efforts initially developed for other, broader purposes are already bringing significant ancillary benefits to agricultural research. The NCBI (Textbox 3) has supported exponential growth and use of genomics inclusive of agricultural plant and animal objectives, whereas the National Oceanic and Atmospheric Administration has facilitated the integration of weather data into a myriad of human activities, including on-farm applications ("apps") for yield forecasting and management decision making.

Textbox 3: The National Center for Biotechnology Information (NCBI) has democratized biotechnology innovation with data access.

Well-supported public federal data repositories have catalyzed and democratized U.S. innovation and economic development. A prime example is the NCBI, a division of the National Library of Medicine established by congress in 1988. The NCBI provides a one-stop database for genomic sequence data, including innovative search and analysis tools, as well as an index database to relevant biotechnology and health science information. Often a mandatory repository for data collected in publicly funded genomics research, the NCBI has catalyzed diverse fields of research, making critical genomics information easy to find and freely available. Although the NCBI supports petabytes of agricultural species genomic data, it lacks capability or mandate to serve most other subdomains of agriculture. Thus, although the NCBI hosts a database of Genotypes and Phenotypes (dbGaP) for studies on humans, it offers no similar function for linking agricultural genomic data to the field studies that document phenotypes as a function of management and environment.

Moving agriculture from its present culture of short data life cycles and limited sharing to one valuing open data and data reuse requires development and implementation of best practices that ensure readability over time and between disciplines.

Moving agriculture from its present culture of short data life cycles and limited sharing to one valuing open data and data reuse requires development and implementation of best practices that ensure readability over time and between disciplines. The Wilkinson and colleagues (2016) global call for researchers and publishers to adopt FAIR (Figure 3) principles is specifically targeted at the disparate but important datasets that are not accommodated by existing well-curated, specialty repositories such as the NCBI. These principles emphasize consistent use of appropriate metadata and free, universally implementable protocols to permit authentication. The intent of FAIR and related initiatives is to set a high bar for scientific data stewardship and sharing, to facilitate transparency, and to spur innovation and impact. Machines must be able to assist in finding, obtaining, and subsequently using relevant data. This, in turn, requires the re-imagination of data workflows and storage solutions that are software agnostic. Certainly it will be a challenge for agriculture to move toward common metadata and data standards and generic workflows, but new organizations, infrastructures, and services are emerging with potential to streamline and professionalize the data management life cycle (Textbox 4). A primary limitation common to these entities, however, is the ability to maintain effective communication and engagement in a widely dispersed, fast-moving field.

Textbox 4: Examples of emerging alliances, coalitions, and networks that may help agriculture with implementation of FAIR principles:

The Research Data Alliance coordinates efforts of a broad array of data managers, information scientists, and data policymakers (Berman, Wilkinson, and Wood 2014). Its Interest Group on Agricultural Data oversees work on best practices, interoperable data standards, on-farm data sharing, and agrisemantics (Research Data Alliance n.d.).

The Ag Data Coalition (Agricultural Data Coalition n.d.) and the International Agroinformatics Alliance (Gustafson et al. 2017) are partnerships working to allow farmers, university researchers, and industry to exchange data safely and securely for analysis and innovation.

DataONE (DataONE n.d.) is a distributed network focused on environmental science, with significant coverage of agricultural research areas; it promotes best practices in data management via educational tools.

AgBioData (AgBioData 2017–2019) is a coalition of major agricultural genomics platform managers seeking to promote best practices in genomics, genetics, and breeding data.

CyVerse (formally iPlant Collaborative) (CyVerse n.d.) provides resources for data management and analysis emphasizing computational infrastructure for huge, complex datasets.

GODAN (GODAN n.d.) supports global efforts in open data by building high-level policy and public/private institutional support.

Decreasing the variety of and increasing the depth of data descriptions remain critical needs.

Along with this array of alliances, coalitions, and networks, an ecosystem of repositories and aggregators is emerging with the intent of fostering FAIR compliance; these also are not without limitations. Agricultural research data can be accessed in a large number of domain databases (e.g., MaizeDB, Soybase), as well as in general purpose publishing repositories (e.g., Dryad [Dryad 2018]) and institutional research repositories (e.g., PURR n.d.; Texas A&M University Libraries 2016). These platforms make data available for free and are tailored to the needs of their immediate stakeholders. They have the potential to contribute to a data-driven future where data can be found, integrated, and used easily regardless of source. The sheer number and variety of these platforms poses a challenge for a coordinated landscape (Parr, Antognoli, and Sears 2017), however, and each faces sustainability issues that constrain them in achieving their full potential, including universal interoperability. Domain-relevant aggregators such as DataONE (Textbox 4) serve a niche but do not yet explicitly serve core agricultural fields such as agronomy, crop genetics, or agricultural economics. Ag Data Commons (USDA–NAL n.d.) is emerging as a central catalog to a wide variety of USDA-supported research data. It harvests relevant metadata from many other repositories, including both scattered domain-specific sources and more general sources filtered to agricultural results. It also provides a repository for selected data with no domain-specific home. Regardless, decreasing the variety of and increasing the depth of data descriptions remain critical needs.

A key to FAIR data is the use of metadata and data standards. Agricultural standards are under active development. Several International Organization for Standardization (ISO) standards are relevant, particularly ISO 19115 (ISO n.d.), now endorsed for federal geospatial data (FGDC n.d.). Industry standards for agricultural equipment data are emerging at AgGateway (AgGateway 2018). For the Agricultural Model Intercomparison and Improvement Project (AgMIP 2014), crop model datasets were harmonized using the vocabularies and standards developed by the International Consortium for Agricultural

Systems Application. Their data dictionary describes terms and units for data related to field crop experiments (White et al. 2013); this standard allowed AgMIP tools to be rapidly developed to translate data so that it could be used in ensemble modeling (Rosenzweig et al. 2013). The proliferation of agricultural ontologies and thesauri is described on the Agrisemantics website (Agrisemantics n.d.). Pruning and mapping across these ontologies is underway. For example, the global multilingual agricultural concept scheme combines the most useful terms from three broadly used sources (Baker, Caracciolo, and Arnaud 2016). These standards have potential for vastly improving discoverability and integration across agricultural data, leading to larger and more interdisciplinary discoveries and innovations (Devare et al. 2016). Broad adoption by existing infrastructures for bench and field researchers, however, remains a challenge. More engagement in standards development by academic, federal, and nonprofit researchers is needed.

Broad adoption by existing infrastructures for bench and field researchers . . . remains a challenge. More engagement in standards development by academic, federal, and nonprofit researchers is needed.

Pathways Forward: Strategies, Partnerships, and Business Models

Designing a singular system or mechanism for sharing agricultural data seems inherently untenable given the array of initiatives currently underway; distributed but interoperable infrastructure likely holds more promise for linking databases. Additionally, the pathway forward must include assurances for security and information quality as well as incentives to publish both confirmatory studies and data in its entirety (i.e., individual replicates) in order to identify false positives and strengthen the characterization of true effects. Agricultural subject experts must solicit partnerships with data scientists. Numerous recent surveys and analyses of data-sharing behaviors in science provide insights into creation of the environment conducive of data-driven, actionable solutions to complex problems (Kim and Stanton 2016; Tenopir et al. 2015). Critical attributes are functional, low-barrier solutions, including desktop tools, for individuals to manage their data in keeping with FAIR principles; the infrastructure, resources, and access to ongoing support to permit real-time FAIR compliance and data curation and preservation; and a system of rewards to incentivize team science and data sharing (Figure 4).

The pathway forward must include assurances for security and information quality as well as incentives to publish both confirmatory studies and data in its entirety (i.e., individual replicates) in order to identify false positives and strengthen the characterization of true effects. Agricultural subject experts must solicit partnerships with data scientists.

Strategies for Transportation to Team Science and Open Data

Simultaneous pursuit of four strategies will facilitate agriculture's pathway forward into data-driven research; these entail bridging gaps, reorienting institutions, leveraging assets, and connecting feedbacks (Figure 4).

Bridging Gaps with Novel Teams and Data Sciences

Dunn and Bourne (2017) identify fostering collaborations with data scientists as a key strategy for building the biomedical data science workforce. In agriculture, data scientists will be critical to successful translation of new knowledge or innovative products from basic research into commercial applications and policies. Researchers and decision makers typically have domain expertise but lack the data-science skills essential for maintaining and making sense of data. Partnering with data scientists to build workflows, analysis tools, and education materials promises to improve both the rate and efficiency of discovery. Once available in accessible, machine-readable forms, scholarly articles and associated data or even unanalyzed data can be mined for trends, filtered for promising ideas, and translated or visualized for end users (Porcel et al. 2012). Data scientists can provide guidance on the correct use of meta-analytical tools to make sense of conflicting studies and harness artificial intelligence and machine learning to explore complex relationships within large and heterogeneous data. Initiatives such as Ag Data Commons and CyVerse provide both tools and infrastructure for specific analyses. Yet the expertise barrier remains high, and the overall process of making decisions from big and open data at the farm, consumer, and policy levels requires substantial further investments and democratization. In regard to artificial intelligence, approaches and tools currently being developed by the Defense Advanced Research Project Agency (DARPA) (Shen n.d.) are being piloted on some agricultural questions. Ownership and development after DARPA by the USDA will require additional resources to create "low-barrier" solutions. Data scientists can also contribute to improving

In agriculture, data scientists will be critical to successful translation of new knowledge or innovative products from basic research into commercial applications and policies.

interoperability (data and coding) and establishment of standards for models and applications; they are well positioned to improve educational technology to reduce training lags.

Data scientists can . . . contribute to improving interoperability (data and coding) and establishment of standards for models and applications; they are well positioned to improve educational technology to reduce training lags.

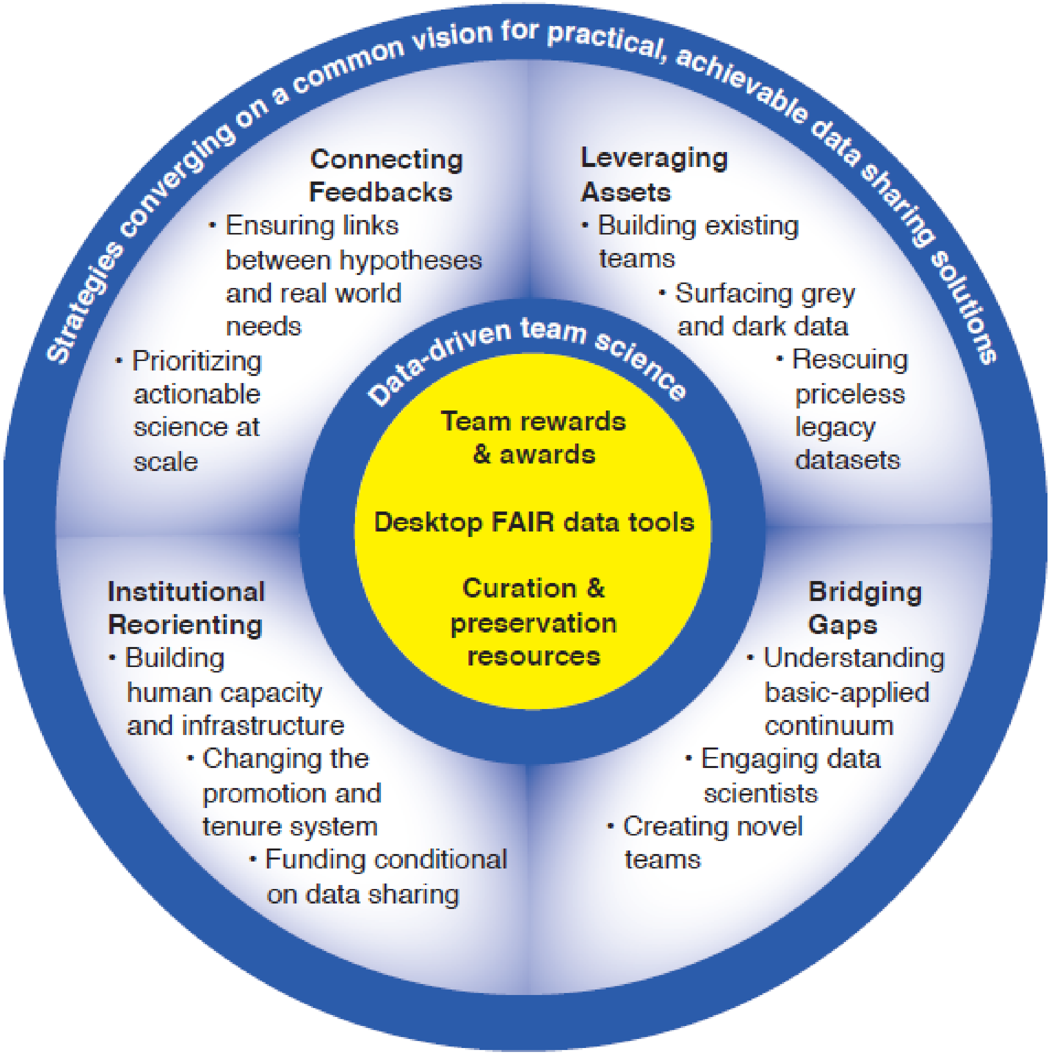

Strategies converging on a common vision for practical, achievable data sharing solutions
Connecting Feedbacks
• Ensuring links between hypotheses and real world needs
• Prioritizing actionable science at scale
Leveraging Assets
• Building existing teams
• Surfacing grey and dark data
• Rescuing priceless legacy datasets
Data-driven team science
Team rewards & awards
Desktop FAIR data tools
Curation & preservation resources
Institutional Reorienting
• Building human capacity and infrastructure
• Changing the promotion and tenure system
• Funding conditional on data sharing
Bridging Gaps
• Understanding basic-applied continuum
• Engaging data scientists
• Creating novel teams

Figure 4. Creating the data ecosystem for public agricultural research.

Institutional Facilitation of Team Science and Data Sharing

Institutions will need to reorient to support team science and data sharing.

Institutions will need to reorient to support team science and data sharing. Undergraduate and graduate curricula must include content that ensures some understanding of data sciences and their importance and use in food systems research. Each individual will not need to be a "data scientist," but all will need a data sciences foundation such that, irrespective of their component of focus, they can recognize how it fits with the system as a whole (Dunn and Bourne 2017); they will need a thorough grounding in FAIR principles, tools for their application, and the ethics of implementing open data. As forecast for the workforce in general (BHEF 2017), agricultural domain specialists with computational, mathematical, and/or statistical skills to help manage and use the large body of data generated by research teams will be in high demand, with many positions anticipated at the bachelor's and master's degree levels.

Any new curriculum must be seamlessly integrated with FAIR infrastructure (workflows, relevant ontologies, standards, repositories, etc.), supporting resource investments must be adequate, and the professional reward system must be altered to incentivize FAIR implementation and team participation. With few exceptions, university researchers are measured and rewarded on a small array of individual metrics: amount of funding garnered, number of papers published and the perceived quality of the journal, and the number and subsequent success of graduate students. In their analysis of institutional and individual factors affecting data sharing, Kim and Stanton (2016) found normative pressure within a discipline and perceived career benefit to have a positive association with data-sharing behaviors. A culture of scholarship for data as a product of scientific endeavor and mechanisms to acknowledge the broader array of activities associated with team science must be developed and infused into the assessment system.

A culture of scholarship for data as a product of scientific endeavor and mechanisms to acknowledge the broader array of activities associated with team science must be developed and infused into the assessment system.

Leveraging Assets and Surfacing Grey/Dark Data

A data-sharing infrastructure would accommodate data not currently represented by peer-review publications. This can increase the reach of research results beyond a given region or even the initial research question and, perhaps more importantly, decrease publication bias in meta-analysis. Existing, successful efforts in data sharing should be examined for insights and mechanisms to emulate and replicate. Several of the Coordinated Agricultural Projects (CAPs) funded by the USDA's Agriculture and Food Research Initiative (AFRI) invested significantly in data-sharing infrastructure to facilitate results integration across teams. For example, the Iowa State University-led Corn CAP compiled standardized crop, soil, and environmental outcome data from more than 30 different research sites over five years, publishing the dataset in the National Agricultural Libraries repository (Abendroth et al. 2017).

A data-sharing infrastructure would accommodate data not currently represented by peer-review publications.

Current grey literature and data not associated with a peer-review publication will require mechanisms to ensure visibility and validity. Simply identifying the existence of documents in grey literature is difficult (Debachere 1995). Permitting formal data publication—inclusive of assigning a digital object identifier—out of repositories (PURR n.d.; USDA–NAL n.d.) is a strategy that can meet open data requirements of funding organizations but may be insufficient for broader discovery objectives. Indeed, availability of repositories alone has not been identified as encouraging data sharing (Kim and Stanton 2016). Sansone and colleagues (2017) describe a data tag suite to enable findability and accessibility over 60 biomedical data sources as well as a general architecture that could be applied to the agricultural data pipeline. Van Tuyl and Whitmire (2016) consider stand-alone data publication that follows citation conventions essential for incentivizing data preparation and ensuring wide accessibility.

Commonly, researchers are unable or unwilling to invest the substantial effort needed to publish studies with negative or nonnovel (replicative) results.

Commonly, researchers are unable or unwilling to invest the substantial effort needed to publish studies with negative or nonnovel (replicative) results. Yet such studies are critical to creating an unbiased foundation to evidence-based practice. To promote inclusion of negative and replication studies in the literature, Nosek and Lakens (2014) propose the use of "Registered Reports" in which peer-reviewed and accepted research proposals are registered prior to data collection, assuring authors that results will be published irrespective of outcome. Kupferschmidt (2018) highlights preregistration as critical to "a recipe for rigor." Finally, resources should be invested to rescue valuable legacy datasets whose initiation predates the digital era and to capture information that cannot be captured by replication (e.g., older animal feed data compilations, genetic records, observations on environmental change and its impacts).

Connecting Feedbacks to Ensure Data Are Useful and Usable

For research data to achieve and maintain public value, they must inform end-user apps designed to enhance and secure our current food supply and address environmental and social challenges.

For research data to achieve and maintain public value, they must inform end-user apps designed to enhance and secure our current food supply and address environmental and social challenges. For data, the Holdren (2013) memo specifically highlights the importance of coordination and collaboration across all relevant entities in public and private sectors. A highly diverse array of end-users—extension, nongovernment institutions, foundations, private entities—implement policies and strategies for achieving a range of outcomes from healthy diets (e.g., MyPlate [USDA n.d.]) to effective environmental markets and sustainability branding (e.g., Fieldprint^{R} calculator [Field to Market n.d.]). Effective apps and tools depend on accurate information from both the demand and supply sides of the equation, and the data needs for end-user apps and tools may not be fully understood by researchers. Within research, disconnects between data needs for model development and data generated in empirical studies has been frequently acknowledged if largely unaddressed (Craufurd et al. 2013). When such poorly calibrated models are converted to apps and deployed at scale, the error is propagated and related management decisions may have widespread and long-lasting negative consequences. Close coupling between researcher and end-user interests is essential to efficient use of resources in data collection, preservation, and curation.

Close coupling between researcher and end-user interests is essential to efficient use of resources in data collection, preservation, and curation.

Partnerships for the Agricultural Data Value Chain

An array of new or strengthened partnerships will be needed to underpin the strategies outlined earlier. Beyond engaging the data scientist, building infrastructure and human capacity within institutions require that agricultural scientists solicit the informatics expertise housed within libraries. In recent years, library scholarship has focused on digital assets and developed tools, workflows, and education materials that are readily adapted to any domain. For example, library sciences were instrumental in the development of the DMPTool (DMPTool 2010–2019) and the Data Curation Profiles Toolkit (Brandt and Kim 2014). The former is a funding agency-compliant, online resource for data management planning, and the latter is used to enhance digital products from any domain; both are examples of the practical solutions libraries can bring to agriculture, and library expertise can be harnessed for development of the architecture of data pipelines and related curricula.

Beyond engaging the data scientist, building infrastructure and human capacity within institutions require that agricultural scientists solicit the informatics expertise housed within libraries.

The articulation of needs and sponsorship of development and implementation of low-barrier solutions to infrastructure also necessitates partnering with science administrators, professional societies, and private publishing entities. In most academic institutions, faculty set the domain-relevant metrics to meet general policy and standards for promotion and tenure; thus, agricultural faculty must not only create new achievement metrics but also assume responsibility for communicating changes to their administrators. Likewise, where administrators have resourced initial development of infrastructure such as repositories to facilitate enhanced competitiveness via funding agency compliance (PURR n.d.), they need to remain engaged to understand ongoing resourcing needs. Sustainability requires supporting repository managers in providing high-level oversight for minimum data standards, including sufficiency of metadata and formats that are actionable (Van Tuyl and Whitmire 2016). As more scientists use repositories, support costs may escalate given current low levels of familiarity with data-sharing best practices. Van Tuyl and Whitmire (2016) also argue that funding agencies that have required data sharing but have thus far been relatively unengaged in the development of best practices have contributed to an “environment of confusion and low-quality shared data.”

When scientists perceive that a funding agency does not enforce their data-sharing policies, the stated policies have not changed data-sharing behaviors.

The Kriesberg and colleagues (2017) analysis of public access plans found the USDA to lag behind most federal agencies in presenting a thorough discussion on addressing all elements of the Holdren (2013) memo. When scientists perceive that a funding agency does not enforce their data-sharing policies, the stated policies have not changed data-sharing behaviors (Kim and Stanton 2016). In their capacity as a major funder of U.S. agricultural research, USDA National Institute of Food and Agriculture (NIFA) AFRI administrators should consider a more active role in incentivizing data sharing, including collaborating with researchers to rapidly pilot specific requirements rather than wait for grassroots emergence of metadata and data standards.

Supplying access to data via published papers appears to confer some citation advantage, and journal policies requiring authors to share the raw data underpinning a publication have been identified as one of the main reasons why researchers share data.

Scientific societies have long served as key arbitrators of professional standards, including what constitutes a scholarly contribution and minimum publishable unit; policies and practices of affiliated journals are extremely influential in driving professional behavior. Supplying access to data via published papers appears to confer some citation advantage (Piwowar and Vision 2013), and journal policies requiring authors to share the raw data underpinning a publication have been identified as one of the main reasons why researchers share data (Mongeon et al. 2017).

In addition to encouraging and supporting systems-oriented research, McNamara, Hanigan, and White (2016) identified the updating of publication guidelines to encourage and support sharing of complete datasets as a key opportunity to advance knowledge. Van Tuyl and Whitmire (2016) recommend publishers not only assume responsibility for setting sharing policies, but also assume some responsibility for assuring the quality of data shared with their journals. In their *A Data Citation Roadmap for Scientific Publishers*, Cousijn and colleagues (2017) suggest publishers provide guidance to authors on suitable repositories. A team of major publishers, repositories, and researchers in the Enabling FAIR Data initiative are beginning to address these concerns for earth and space sciences (Stall et al. 2017). Ultimately, designing functional architecture for open data access in agriculture requires partners to commit to collaborative, iterative analysis of successes and failures in design, implementation, and utility.

Ultimately, designing functional architecture for open data access in agriculture requires partners to commit to collaborative, iterative analysis of successes and failures in design, implementation, and utility.

The Business Case and the Business Model for Data Sharing

Physical and cyber infrastructure require a business case for making open access data and data tools viable to start and sustain over the long term. A major challenge that interoperable infrastructure would overcome is the financial inefficiency of multiple organizations implementing their own stop-gap solutions to similar data problems without seeking economies of scale (Sewadeh and Sisson 2018). For example, individual USDA agencies (e.g., Agricultural Research Service, Farm Service Agency, Forest Service, National Agricultural Statistics Service [NASS]) separately maintain and pay for their geospatial data platform. Pooling datasets and computational power and creating one-stop shopping would extend sparse data resources. Importantly, beyond a business case of efficiency is the case of facilitating new discovery and derivation of better answers. The NCBI (Textbox 3) is a tremendous example. By storing genetic sequence data from thousands of species in one data center and developing tools to search these data, researchers are rapidly identifying the functions of genes, causes of disease, and human evolutionary signatures. This simply could not be done if each data collection project resided on a single lab's computer or in a domain- or species-specific database. The critical business case for data-sharing infrastructure in agriculture is that it will enable better understanding and decision making, with a low barrier of entry, so that U.S. production agriculture can compete sustainably.

Physical and cyber infrastructure require a business case for making open access data and data tools viable to start and sustain over the long term.

Toward this end, competitive grants programs (e.g., USDA–NIFA, National Science Foundation) could be extremely useful to build tools and apps but would not be efficient mechanisms for long-term data storage and curation, much as a library or the NCBI could not subsist on competitive grants. Short-term competitive funding cycles target innovation in research projects, not maintenance of supporting infrastructure or databases; other mechanisms are needed to support data infrastructure post innovation (Gabella, Durinx, and Appel 2017). In their analysis of biological database longevity, Attwood, Agit, and Ellis (2015) found that persistence of web-based data assets was relatively rare (23% of 326 databases were still "alive" after 18 years); databases with longevity were almost always a function of core, institutional support. Recognizing the wasted investment such failures of persistence represent, these authors cautioned against creating more databases without developing long-term financial strategies inclusive of persistent, institutional leadership. The model of the USDA–NASS, which surveys and stores long-term agricultural production data, would be more relevant than traditional research grant programs as a path toward sustained financial support.

As agriculture considers pathways forward for data, careful examination of the various financial models currently under active consideration by other domains should be undertaken.

As agriculture considers pathways forward for data, careful examination of the various financial models currently under active consideration by other domains should be undertaken. Infrastructure to support open data is, by definition, a public good and, as such, sustainability strategies must consider governance and community and not just cost (Neylon 2017). For domains in which data sharing is common, funding uncertainties have opened discussion on functional, hybrid business models that could supplement and stabilize prevailing models of public financing via short-duration research grants and/or strictly national funding (e.g., USDA–NASS).

Reiser and colleagues (2016) explore the question of who pays with a case study focused on The Arabidopsis Information Resource (TAIR n.d.) and suggest subscription fee-based models could be an option, but the challenge is balancing maximized access with incentives for subscriptions. The Arabidopsis Information Resource uses tiering based on usage to set fees in an approach analogous to the current business models used by private publishers of peer-review journals. Dryad (2018) offers both volume-based and usage-based submission fee payment plans for researchers, funders, universities, and publishers. Even as private publishers are moving into the data-sharing space via data publications (e.g., Elsevier 2019), however, their current business model (typically selling content to library subscribers) has been flagged as incompatible with open access (Björk 2017; Schimmer, Geschuhn, and Vogler 2015). Indeed, Allahar (2017) argues that it is only a matter of time before Internet technology and open-access philosophies combine to disrupt traditional private publisher-subscription models.

Ultimately, some combination of institutional support and infrastructure models may offer the most promise for public agricultural research data.

Other models that are fully compatible with open access include the institutional support and infrastructure models as well as an array of models considered less stable because they rely on commercial or other partnering and/or willingness to contribute. In a case study of the Universal Protein Resource (UniProt n.d.), Gabella, Durinx, and Appel (2017) review these models, as well as four other models not considered fully compliant with open access, and conclude that the infrastructure model is the most sustainable and equitable option for core life sciences data. In this yet untested model, public and private funding agencies pay directly for stewardship in contributions proportionate to grant volume. In contrast, PURR (n.d.) exemplifies an institutional support approach because Purdue administration not only funded the development of the repository, but currently subsidizes most of the recurring costs. The growth in costs associated with growth in use, however, are unclear as are any potential limits to internal funding for long-term curation of data as a public good versus as an asset primarily available to institutional employees. Ultimately, some combination of institutional support and infrastructure models may offer the most promise for public agricultural research data.

Even with stronger requirements from funders for data preparation, some activities such as anonymization remain beyond the scope of the funded research.

The cost of the data infrastructure envisioned here is largely unknown but, without doubt, substantial investment will be needed. Cost concerns and uncertainties reflect the additional workflows and human resources needed for FAIR data stewardship as compared to simple storage (Bourne, Lorsch, and Green 2015). For journal articles, Karp (2016) estimated the additional costs of curation for open access articles to be minimal (an added $219 for five years), but the existing publication process is designed to produce curatable digital objects. In contrast, Gabella, Durinx, and Appel (2017) argue that data generated in a research project is not necessarily a finished product and will likely require additional, dedicated funding to convert data and databases to FAIR-compliant "knowledge bases."

Even with stronger requirements from funders for data preparation, some activities such as anonymization remain beyond the scope of the funded research. Service and technology investment required to deliver data online and without price or permission barriers—subscriptions, pay-per-view fees, licensing restrictions, etc.—is far from cost free. Gabella, Durinx, and Appel (2017) estimate 1% of the entire life-science budget would be sufficient to support an infrastructure model, but it is unknown how costs would scale with the smaller overall budgets of agriculture. On the bright side, new return-on-investment studies in domains with large research data infrastructure are demonstrating substantial payoffs from data reuse (Beagrie and Houghton 2014; Sullivan, Brennan-Tonetta, and Marxen 2017). For example, the public life-science data managed by the European Bioinformatics Institute brings estimated research and development value of 6 to 7 times its annual operating expenses (Beagrie and Houghton 2016).

It behooves the agricultural research community to not only invest in learning from the experiences of other scientific domains but also in piloting agricultural case studies to simultaneously strengthen the business case for sharing data and assess the value proposition of candidate business models.

In sum, it behooves the agricultural research community to not only invest in learning from the experiences of other scientific domains but also in piloting agricultural case studies to simultaneously strengthen the business case for sharing data and assess the value proposition of candidate business models. The most cost-effective and robust solutions for infrastructure may involve established tools and methods with the best service for the lowest cost coming from a mix of the innovative with the proven (Reiser et al. 2016). Leadership and oversight

Convening stakeholders in public listening sessions and USDA agency leadership in high-level meetings . . . are logical first steps . . . for developing and implementing infrastructure for open data.

by the USDA Research Education and Economics office, and specifically the Office of the Chief Scientist (OCS) in partnership with the Office of the Chief Information Officer (OCIO), would be logical. Stewardship of public research data is a natural extension of their historic role in support of the crop and animal research facilities that directly improved public well-being and provided the foundation for additional innovation by the private sector. The 2018 Farm Bill created the Agriculture Advanced Research and Development Authority (AgARDA) under the direction of the chief scientist that was envisioned to have the scope, authority, and investment needed to initiate improvement in integrating USDA's data. Full appropriation of authorized funds for AgARDA would position a federal authority specifically to address the issues raised in this commentary; convening stakeholders in public listening sessions and USDA agency leadership in high-level meetings led by the OCS and OCIO are logical first steps for AgARDA in leading a partnership for developing and implementing infrastructure for open data that will deliver the necessary benefits to all stakeholders in the agricultural data value chain.

Literature Cited

Abendroth, L. J., D. E. Herzmann, G. Chighladze, E. J. Kladivko, M. J. Helmers, L. Bowling, M. Castellano, R. M. Cruse, W. A. Dick, N. R. Fausey, and J. Frankenberger. 2017. Sustainable corn CAP research data (USDA–NIFA Award No. 2011-68002-30190). National Agricultural Library, USDA–ARS.

AgBioData. 2017–2019. Welcome to AgBioData, https://www.agbiodata.org (11 December 2018)

AgGateway. 2018. Standards and guidelines, http://www.aggateway.org/GetConnected/StandardsGuidelines.aspx (11 December 2018)

Agricultural Data Coalition. n.d. Agricultural Data Coalition: Putting farmers in the driver's seat, http://agdatacoalition.org (11 December 2018)

Agricultural Model Intercomparison and Improvement Project (AgMIP). 2014. Homepage, http://www.agmip.org (11 December 2018)

Agrisemantics. n.d. Semantics for the interoperability of agricultural data, https://agrisemantics.org (11 December 2018)

Allahar, H. 2017. Academic publishing, Internet technology, and disruptive innovation. *Tech Innov Manag Rev* 7 (11), https://timreview.ca/article/1120 (7 January 2019)

Attwood, T. K., B. Agit, and L. B. Ellis. 2015. Longevity of biological databases. EMBnet.journal 21:e803, https://journal.embnet.org/index.php/embnetjournal/article/view/803/1209 (7 January 2019)

Baker, M. 2016. Is there a reproducibility crisis? *Nature* 553:452–454.

Baker, T., C. Caracciolo, and E. Arnaud. 2016. Global agricultural concept scheme (GACS): A hub for agricultural vocabularies. In *7th International Conference on Biomedical Ontologies, ICBO* 16:2.

Beagrie, N. and J. W. Houghton. 2014. *The Value and Impact of Data Sharing and Curation: A Synthesis of Three Recent Studies of UK Research Data Centres*. Jisc. 26 pp.

Beagrie, N. and J. Houghton. 2016. *The Value and Impact of the European Bioinformatics Institute*. EMBL-EBI. 88 pp., http://vuir.vu.edu.au/33707/1/EBI-impact-report.pdf (7 January 2019)

Begley, C. G. and J. P. Ioannidis. 2015. Reproducibility in science: Improving the standard for basic and preclinical research. *Circ Res* 116 (1): 116–126.

Bennett, E. M. and P. Balvanera. 2007. The future of production systems in a globalized world. *Front Ecol Environ* 5 (4): 191–198.

Berman, F., R. Wilkinson, and J. Wood. 2014. Guest editorial: Building global infrastructure for data sharing and exchange through the research data alliance. *D-Lib Mag* 20 (1/2), http://www.dlib.org/dlib/january14/01guest_editorial.html (25 January 2019)

Björk, B. C. 2017. Scholarly journal publishing in transition—From restricted to open access. *Electron Mark* 27 (2): 101–109.

Bollen, K., J. T. Cacioppo, R. Kaplan, J. Krosnicj, and J. L. Olds. 2015. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. National Science Foundation, Arlington, Virginia. 29 pp., https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf (7 January 2019)

Bourne, P. E., J. R. Lorsch, and E. D. Green. 2015. Perspective: Sustaining the big-data ecosystem. *Nature* 527 (7576): S16–S17.

Box, J. F. 1980. R. A. Fisher and the design of experiments, 1922–1926. *Am Stat* 34 (1): 1–7.

Brandt, D. S. and E. Kim. 2014. Data curation profiles as a means to explore managing, sharing, disseminating or preserving digital outcomes. *Int J Perf Arts Digit Media* 10 (1): 21–34.

Brouder, S. M. and H. Gomez-Macpherson. 2014. The impact of conservation agriculture on smallholder agricultural yields: A scoping review of the evidence. *Agr Ecosyst Environ* 187:11–32.

Business Higher Education Forum (BHEF). 2017. *Investing in America's Data Science and Analytics Talent: The Case for Action*. 24 pp., http://www.bhef.com/sites/default/files/bhef_2017_investing_in_dsa.pdf (7 January 2019)

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14 (5): 365–376, https://www.nature.com/articles/nrn3475 (8 January 2019)

Campbell, B. M., D. J. Beare, E. M. Bennett, J. M. Hall-Spencer, J. S. I. Ingram, F. Jaramillo, R. Ortiz, N. Ramankutty, J. A. Sayer, and D. Shindell. 2017. Agriculture production as a major driver of the Earth system exceeding planetary boundaries. *Ecol Soc* 22 (4): 8, https://doi.org/10.5751/ES-09595-220408 (8 January 2019)

Caron, P., E. Biénabe, and E. Hainzelin. 2014. Making transition towards ecological intensification of agriculture a reality: The gaps in and the role of scientific knowledge. *Curr Opin Env Sust* 8:44–52.

Cochrane. n.d. Homepage, https://www.cochrane.org (10 December 2018)

Collaboration for Environmental Evidence. 2018. Homepage, http://www.environmentalevidence.org (10 December 2018)

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, P. Polischuk, S. Taylor, M. Martone, and T. Clark. 2017. *A Data Citation Roadmap for Scientific Publishers*. *Sci Data* 180259, https://www.nature.com/articles/sdata2018259 (4 January 2019).

Cragin, M. H., C. L. Palmer, J. R. Carlson, and M. Witt. 2010. Data sharing, small science and institutional repositories. *Philos T Roy Soc A* 368 (1926): 4023–4038.

Craufurd, P. Q., V. Vadez, S. V. Jagadish, P. V. Prasad, and M. Zaman-Allah. 2013. Crop science experiments designed to inform crop modeling. *Agr Forest Meteorol* 170:8–18.

CyVerse. n.d. Homepage, https://www.cyverse.org (11 December 2018)

Daar, A. S., P. A. Singer, D. L. Persad, S. K. Pramming, D. R. Matthews, R. Beaglehole, A. Bernstein, L. K. Borysiewicz, S. Colagiuri, N. Ganguly, and R. I. Glass. 2007. Grand challenges in chronic non-communicable diseases. *Nature* 450 (7169): 494.

Data Observation Network for Earth (DataONE). n.d. What is DataONE?, https://www.dataone.org (11 December 2018)

Davis, A. S., J. D. Hill, C. A. Chase, A. M. Johanns, and M. Liebman. 2012. Increasing cropping system diversity balances productivity, profitability and environmental health. *PLOS One* 7 (10): e47149.

de Vrieze, J. 2018. The metawars. *Science* 361 (6408): 1184–1188.

Debachere, M. C. 1995. Problems in obtaining grey literature. *IFLA J* 21 (2): 94–98.

Devare, M., C. Aubert, M. A. Laporte, L. Valette, E. Arnaud, and P. L. Buttigieg. 2016. Data-driven agricultural research for development: A need for data harmonization via semantics. In *CEUR Workshop Proceedings*, http://ceur-ws.org/Vol-1747/IT205_ICBO2016.pdf (8 January 2019)

DMPTool. 2010–2019. Homepage, https://dmptool.org (4 January 2019)

Dobermann, A., R. Nelson, D. Beever, D. Bergvinson, E. Crowley, G. Denning, K. Giller, J. d'Arros Hughes, M. Jahn, J. Lynam, W. Masters, R. Naylor, G. Neath, I. Onyido, T. Remington, I. Wright, and F. Zhang. 2013. *Solutions for Sustainable Agriculture and Food Systems*. Sustainable Development Solutions Network, http://unsdsn.org/wp-content/uploads/2014/02/130919-TG07-Agriculture-Report-WEB.pdf (8 January 2019)

Dryad. 2018. Homepage, https://datadryad.org (11 December 2018)

Dunn, M. C. and P. E. Bourne. 2017. Building the biomedical data science workforce. *PLOS Biol* 15 (7): e2003082, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2003082 (8 January 2019)

Eagle, A. J., L. E. Christianson, R. L. Cook, R. D. Harmel, F. E. Miguez, S. S. Qian, and D. A. Ruiz Diaz. 2017. Meta-analysis constrained by data: Recommendations to improve relevance of nutrient management research. *Agron J* 109 (6): 2441–2449.

Efron, B. 1998. R. A. Fisher in the 21st century. *Stat Sci* 13 (2): 95–122.

Ehm, W. 2016. Reproducibility from the perspective of meta-analysis. Ch. 7. In H. Atmanspacher and S. Maasen (eds.). *Reproducibility: Principles, Problems, Practices, and Prospects*. John Wiley & Sons, Inc., Hoboken, New Jersey, https://doi.org/10.1002/9781118865064.ch7

Eisenhardt, K. M., M. E. Graebner, and S. Sonenshein. 2016. Grand challenges and inductive methods: Rigor without rigor mortis. *Acad Manage J* 59 (4): 1113–1123.

Evans School Policy Analysis and Research Group (EPAR). 2017. Blog, https://evans.uw.edu/policy-impact/epar/blog/private-public-and-philanthropic-funding-global-agricultural-and-health (8 January 2019)

Fair Access to Science and Technology Research Act (FASTR). 2017. HR 3427, S 1701. 115th Cong., https://sparcopen.org/our-work/fastr/ (8 January 2019)

Fanelli, D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90 (3): 891–904.

Federal Geographic Data Committee (FGDC). n.d. ISO geospatial metadata standards, https://www.fgdc.gov/metadata/iso-standards (11 December 2018)

Federal Research Public Access Act (FRPAA). 2012. HR 4004. 112th Cong., 2d sess., https://www.congress.gov/bill/112th-congress/house-bill/4004/text (8 January 2019)

Ferguson, A. R., J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone. 2014. Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nat Neurosci* 17 (11): 1442–1447.

Field to Market. n.d. Homepage, https://calculator.fieldtomarket.org/#/ (4 January 2019)

Gabella, C., C. Durinx, and R. Appel. 2017. Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. *F1000Research* 6 (ELIXIR): 2051, https://www.elixir-europe.org/platforms/data/funding-models (8 January 2019)

Global Open Data for Agriculture and Nutrition (GODAN). n.d. Homepage, https://www.godan.info (11 December 2018)

Goodman, S. N., D. Fanelli, and J. P. Ioannidis. 2016. What does research reproducibility mean? *Sci Transla Med* 8 (341): 341ps12–341ps12, http://stm.sciencemag.org/content/8/341/341ps12 (8 January 2019)

Griggs, D., M. Stafford-Smith, O. Gaffney, J. Rockström, M. C. Öhman, P. Shyamsundar, W. Steffen, G. Glaser, N. Kanie, and I. Noble. 2013. Policy: Sustainable development goals for people and planet. *Nature* 495 (7441): 305-307.

Gustafson, A., J. Erdmann, M. Milligan, G. Onsongo, P. Pardey, T. Prather, K. Silverstein, J. Wilgenbusch, and Y. Zhang. 2017. A platform for computationally advanced collaborative agroinformatics data discovery and analysis. P. 2. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, New Orleans, Louisiana, 9–13 July 2017, ACM DL.

Harris, R. 2017. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. Basic Books, New York, New York. 288 pp.

Holdren, J. P. 2013. 2013 Memorandum for the Heads of Executive Departments and Agencies: 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 (30 May 2018)

International Organization for Standardization (ISO). n.d. ISO 19115-1:2014—Geographic information—Metadata—Part 1: Fundamentals, https://www.iso.org/standard/53798.html (11 December 2018)

Jarvis, M. F. and M. Williams. 2016. Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol Sci* 37 (4): 290–302.

Karp, P. D. 2016. How much does curation cost? *Database* 2016:baw110, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4976296/ (9 January 2019)

Kim, Y. and J. M. Stanton. 2016. Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *J Assoc Inf Sci Tech* 67 (4): 776–799.

Kriesberg, A., K. Huller, R. Punzalan, and C. Parr. 2017. An analysis of federal policy on public access to scientific research data. *Data Sci J* 16, https://datascience.codata.org/articles/10.5334/dsj-2017-027/ (9 January 2019)

Kupferschmidt, K. 2018. A recipe for rigor. *Science* 361 (6408): 1192–1193.

McNamara, J. P., M. D. Hanigan, and R. R. White. 2016. Invited review: Experimental design, data reporting, and sharing in support of animal systems modeling research. *J Dairy Sci* 99 (12): 9355–9371.

Mongeon, P., N. Robinson-Garcia, W. Jeng, and R. Costas. 2017. Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science. *Aslib J Inform Manag* 69 (5): 545–556.

Mueller, N. D., J. S. Gerber, M. Johnston, D. K. Ray, N. Ramankutty, and J. A. Foley. 2012. Closing yield gaps through nutrient and water management. *Nature* 490 (7419): 254.

Munafò, M. R., B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simonsohn, E. J. Wagenmakers, J. J. Ware, and J. P. Ioannidis. 2017. A manifesto for reproducible science. *Nat Hum Behav* 1:0021, https://www.nature.com/articles/s41562-016-0021 (9 January 2019)

National Academies of Sciences, Engineering, and Medicine. 2018a. *Science Breakthroughs to Advance Food and Agricultural Research by 2030*. The National Academies Press, Washington, D.C., https://doi.org/10.17226/25059 (9 January 2019)

National Academies of Sciences, Engineering, and Medicine. 2018b. *Open Science by Design: Realizing a Vision for 21st Century Research*. The National Academies Press, Washington, D.C., https://doi.org/10.17226/25116 (9 January 2019)

Neylon, C. 2017. Sustaining scholarly infrastructures through collective action: The lessons that Olson can teach us. *KULA* 1 (1): 3, https://kula.uvic.ca/articles/10.5334/kula.7/ (9 January 2019).

Nosek, B. A. and D. Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Hogrefe* 45:137–141, https://doi.org/10.1027/1864-9335/a000192 (9 January 2019)

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251): aac4716.

Parr, C., E. Antognoli, and J. Sears. 2017. How agricultural researchers share their data: A landscape inventory. *Proc TDWG* 1:e20434, https://doi.org/10.3897/tdwgproceedings.1.20434 (9 January 2019)

Pashler, H. and E. J. Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspect Psychol Sci* 7 (6): 528–530.

Philibert, A., C. Loyce, and D. Makowski. 2012. Assessment of the quality of meta-analysis in agronomy. *Agr Ecosyst Environ* 148:72–82.

Piwowar, H. A. and T. J. Vision. 2013. Data reuse and the open data citation advantage. *PeerJ* 1:e175.

Citation:
Council for Agricultural Science and Technology (CAST). 2019. *Enabling Open-source Data Networks in Public Agricultural Research*. CAST Commentary QTA2019-1. CAST, Ames, Iowa.

Porcel, C., A. Tejeda-Lorente, M. A. Martínez, and E. Herrera-Viedma. 2012. A hybrid recommender system for the selective dissemination of research resources in a technology transfer office. *Inform Sciences* 184 (1): 1–19.

Purdue University Research Repository (PURR). n.d. Research data management for Purdue, https://purr.purdue.edu (11 December 2018)

Reiser, L., T. Z. Berardini, D. Li, R. Muller, E. M. Strait, Q. Li, Y. Mezheritsky, A. Vetushko, and E. Huala. 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database* 2016:baw018, https://academic.oup.com/database/article/doi/10.1093/database/baw018/2630208 (9 January 2019)

Research Data Alliance. n.d. Agricultural data interest group, https://rd-alliance.org/groups/agriculture-data-interest-group-igad.html (11 December 2018)

Robertson, G. P. and S. M. Swinton. 2005. Reconciling agricultural productivity and environmental integrity: A grand challenge for agriculture. *Front Ecol Environ* 3 (1): 38–46.

Rosenzweig, C., J. W. Jones, J. L. Hatfield, A. C. Ruane, K. J. Boote, P. Thorburn, and S. Asseng. 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agr Forest Meteorol* 170:166–182.

Sansone, S. A., A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J. S. Grethe, H. Xu, I. M. Fore, J. Lyle, A. E. Gururaj, X. Chen, and H. E. Kim. 2017. DATS, the data tag suite to enable discoverability of datasets. *Sci Data* 4:170059.

Schimmer, R., K. K. Geschuhn, and A. Vogler. 2015. Disrupting the subscription journals' business model for the necessary large-scale transformation to open access, doi:10.17617/1.3.

Sewadeh, M. and J. Sisson. 2018. Disseminating government data effectively in the age of open data. Pp. 13–28. In F. A. Batarseh and R. Yang (eds.). *Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence*. Academic Press, Cambridge, Massachusetts.

Shen, W. n.d. Data-driven discovery of models (D3M). Defense Advanced Research Project Agency (DARPA), https://www.darpa.mil/program/data-driven-discovery-of-models (13 December 2018)

Speed, T. P. 1992. Introduction to Fisher (1926)—The arrangement of field experiments. Pp. 71–81. In S. Kotz and N. L. Johnson (eds.). *Breakthroughs in Statistics*. Springer, New York, New York.

Stall, S., K. A. Lehnert, E. Robinson, M. Parsons, J. Cutcher-Gershenfeld, B. A. Nosek, B. Hanson, and L. Yarmey. 2017. Enabling FAIR data. OSF Home, https://osf.io/jy4d9/ (9 January 2019)

Stokstad, E. 2018. The truth squad. *Science* 361 (6408): 1189–1191, doi:10.1126/science.361.6408.1189.

Sullivan, K. P., P. Brennan-Tonetta, and L. J. Marxen. 2017. *Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank*. Office of Research Analytics, Rutgers New Jersey Agricultural Experiment Station, New Brunswick, New Jersey, https://cdn.rcsb.org/rcsb-pdb/general_information/about_pdb/Economic%20Impacts%20of%20the%20PDB.pdf (9 January 2019)

Tenopir, C., E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS One* 10 (8): e0134826, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134826 (9 January 2019)

Texas A&M University Libraries. 2016. OAKTrust, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134826 (11 December 2018)

The Arabidopsis Information Resource (TAIR). n.d. Homepage, https://www.arabidopsis.org (7 January 2019)

U.S. Department of Agriculture (USDA). n.d. Homepage, https://www.ChooseMyPlate.gov (4 January 2019)

U.S. Department of Agriculture–Economic Research Service (USDA–ERS). 2018. Agricultural research funding in the public and private sectors, https://www.ers.usda.gov/data-products/agricultural-research-funding-in-the-public-and-private-sectors/ (9 January 2019)

U.S. Department of Agriculture–National Agricultural Library (USDA–NAL). n.d. Ag data commons beta, https://data.nal.usda.gov (11 December 2018)

U.S. Environmental Protection Agency (USEPA). 2017. Procedures for chemical risk evaluation under the amended toxic substances control act. Rule 82FR33726. *Fed Regist* 82 FR 33726: 33726–33753, July 20, 2017, https://www.federalregister.gov/documents/2017/07/20/2017-14337/procedures-for-chemical-risk-evaluation-under-the-amended-toxic-substances-control-act (25 January 2019)

UniProt. n.d. Homepage, https://www.uniprot.org (7 January 2019)

van Mil, H. G. J., E. A. Foegeding, E. J. Windhab, N. Perrot, and E. Van Der Linden. 2014. A complex system approach to address world challenges in food and agriculture. *Trends Food Sci Tech* 40 (1): 20–32.

Van Tuyl, S. and A. L. Whitmire. 2016. Water, water, everywhere: Defining and assessing data sharing in academia. *PLOS One* 11 (2): e0147942.

White, J. W., L. A. Hunt, K. J. Boote, J. W. Jone, J. Koo, K. Soonho, C. Porter, P. W. Wilkins, and G. Hoogenboom. 2013. Integrated description of agricultural field experiments and production: The ICASA version 2.0 data standards. *Comput Electron Agr* 96:1–12, doi:10.1016/j.compag.2013.04.003.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. ‘t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018, doi:10.1038/sdata.2016.18.