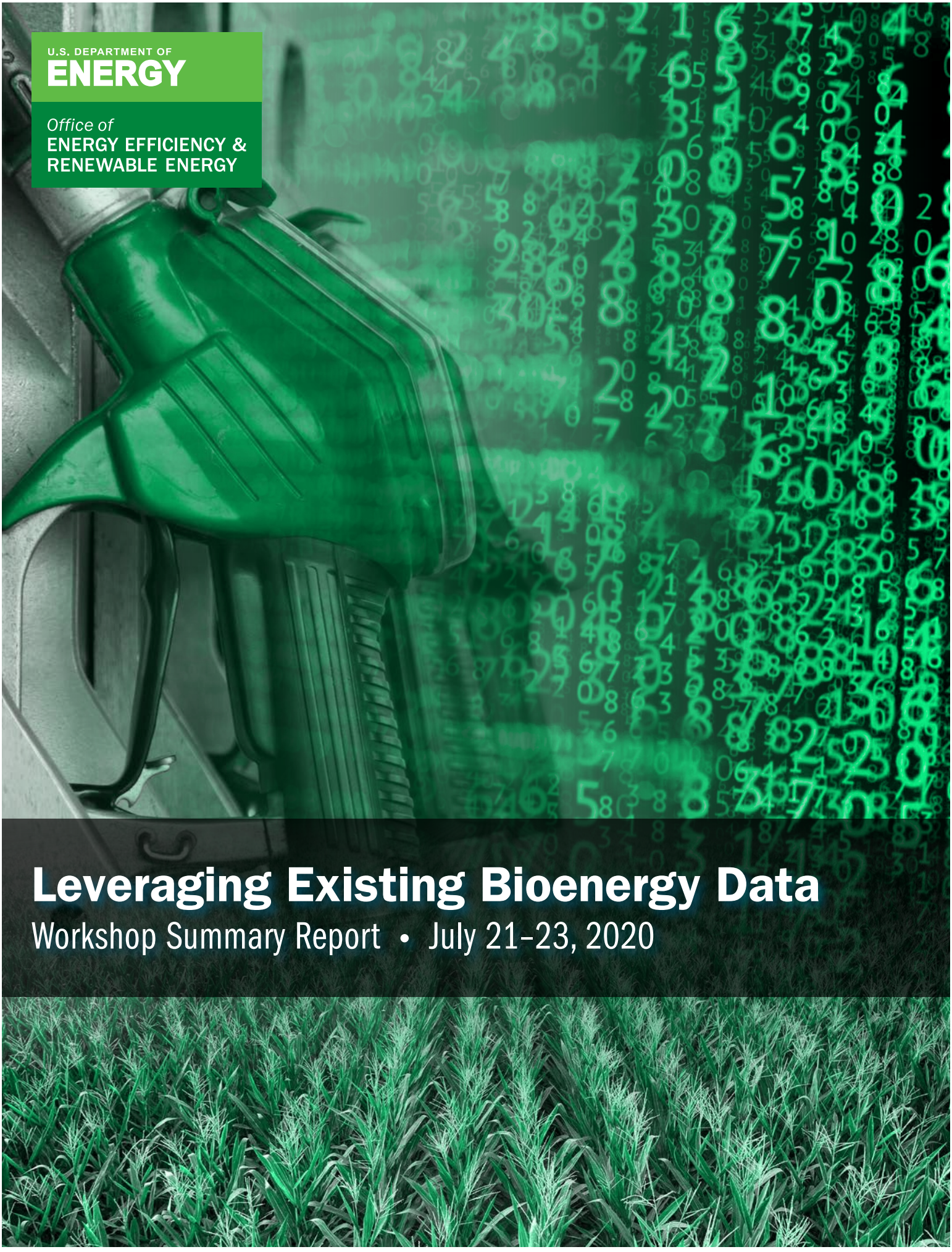


U.S. DEPARTMENT OF
ENERGY

Office of
ENERGY EFFICIENCY &
RENEWABLE ENERGY

Leveraging Existing Bioenergy Data

Workshop Summary Report • July 21–23, 2020

Disclaimer

Foreword

The U.S. Department of Energy's (DOE's) Office of Energy Efficiency and Renewable Energy (EERE) invests in a diverse portfolio of technologies to ensure domestic energy security, continued economic competitiveness, environmental sustainability, and the availability of cleaner fuels and power. The mission of EERE's Bioenergy Technologies Office (BETO) is to develop transformative and revolutionary sustainable bioenergy technologies for a prosperous nation. BETO develops technologies that convert domestic biomass and waste resources into fuels, products, and power to enable affordable energy, economic growth, and innovation in renewable energy and chemicals production.

This report summarizes the input received from attendees of the public workshop sponsored by BETO on July 21–23, 2020, as well as input received from a Request for Information (RFI) run in August and September 2019.

Acknowledgments

This summary report was prepared by Lauren Illing, Liz Burrows, and Becca Szymkowicz, with contributions from Jamie Meadows, Andrea Bailey, Nichole Fitzgerald, and Beau Hoffman. The authors would like to sincerely thank the workshop participants for their contributions, which provided input for this publication. The full list of individuals who registered for the workshop is provided in Appendix B.

Workshop conceptualization and planning was led by Liz Burrows with support from Beau Hoffman, Becca Szymkowicz, Jamie Meadows, Julia Moody, Andrea Bailey, Evan Mueller, Colleen Tomaino, and Mark Shmorhun. Julia Moody and Kim Graber also provided invaluable legal perspective to frame the workshop. The workshop plenary speakers also contributed their time, sharing their expert insights with workshop organizers to help inform and shape the workshop discussions (Table 1).

Table 1. Workshop Plenary Speakers

Plenary Speaker	Job Title and Affiliation
Julia Moody	Deputy Chief Counsel for Intellectual Property at DOE's Golden Field Office
Kim Graber	Legal Counsel at DOE
John Ellersick	President of Next Rung Technology
Kjiersten Fagnan	Chief Informatics Officer and Data Science and Informatics Leader at the DOE Joint Genome Institute
Charles Tait Graves	Partner at Wilson Sonsini Goodrich & Rosati
Didier Navez	Vice President of Strategy & Alliances at Dawex
Doug Laney	Principal Data Strategist at Caserta
Debbie Brodt-Giles	Group Manager – Data, Analytics, Tools, and Applications at National Renewable Energy Laboratory

Lauren Illing led the breakout session facilitation planning, including group discussion process, questions, virtual format, and training the team of breakout session leaders. Stacey Young conducted logistics planning for the initial in-person workshop date, as well as the postponed virtual workshop. We gratefully acknowledge the team of moderators, facilitators, and scribes who guided participants in each breakout session discussion, with special appreciation to the rapporteurs who agreed to create and report out breakout session summaries with quick turnaround (Table 2). Breakout session leader affiliations are provided in Appendix B.

Table 2. Breakout Session Leaders

Session Name	Moderator	Facilitator	Scribe	Rapporteur
Data Quality	Liz Burrows	Lauren Illing	Jamie Meadows	Justin Billing
Data Acquisition	Beau Hoffman	Lauren Illing	Jamie Meadows	Bill Smith
Data Valuation	Liz Burrows	Lauren Illing	Becca Szymkowicz	Bruce Wilson/Joe Sagues
Feedstock Handling and Biorefineries	Mark Shmorhun	Liz Burrows	Clayton Rohman	Rachel M. Emerson
Thermochemical Conversion	Andrea Bailey	Seth Menter	Ben Simon	Asanga Padmaperuma
Microorganisms in Biotechnology	Beau Hoffman	Rob Naranjo	Robert Natelson	Deepti Tanjore
Algae	Daniel Fishman	Colleen Tomaino	Jessica Phillips	Alina Corcoran

List of Acronyms

BETO	Bioenergy Technologies Office
DOE	U.S. Department of Energy
DOI	digital object identifier
EERE	Office of Energy Efficiency and Renewable Energy
FAIR	Findable, Accessible, Interoperable, and Reusable
IP	intellectual property
NMDC	National Microbiome Data Collective
RFI	Request for Information

Executive Summary

The "Leveraging Existing Bioenergy Data" online workshop brought together a wide range of experts in data acquisition and data valuation, as well as bioenergy stakeholders, to discuss how to collect and valorize underused data sets and associated knowledge with the goal of making this information public on existing databases.

The main rationale behind this workshop was that there are existing high-impact, industrially relevant bioenergy data sets that are currently underused and could potentially be acquired at prices much lower than they cost to generate. These data sets could come from companies that have pivoted or failed, or from existing companies that have subsets of data that may be non-sensitive. Providing a small monetary incentive to acquire this information and make it public could *strengthen* companies and accelerate the bioeconomy by increasing the relevance of basic research, avoiding duplication of efforts, building on others' findings, and streamlining methods and operating conditions. The data, and often more importantly the outcomes or findings from the research, could be made public on existing established databases so that minimal resources would be spent on data curation.

This report summarizes the workshop and the associated Request for Information (RFI) published by the Bioenergy Technologies Office (BETO) in 2019. The aggregated input received from stakeholders has informed a project developed to leverage existing bioenergy data. The project, led by Oak Ridge National Laboratory, is titled "Accelerating Bioenergy Technology Advancement Through Findable, Accessible, Interoperable, and Reusable (FAIR) Data Delivery." The project team is widely soliciting data requests and data offers via its website, https://fair-bioenergy-data.pages.ornl.gov/.

The workshop aimed to accomplish the following goals:

Data Quantity and Quality

- Generate a list of potential data sources (e.g., current or former companies, labs)
- Discuss the most useful and available data types (e.g., methods, operating parameters, innovations, market analyses, resource assessments)
- Establish the best ways to measure and/or ensure quality of existing data sets (e.g., sign-off from scientists/engineers who collected it, ability to obtain missing metadata, repeatability of design).

Data Monetization and Valuation

- Propose strategies to adequately determine the value of data given the extremely large number of variables (e.g., potential impact, return on investment, level of interest, number of guaranteed users, age of the data, completeness of metadata, type of data).

Data Acquisition

- Clarify the legal processes required to obtain certain types of data
- Develop a process by which users can submit requests for data, suppliers can provide data and be compensated, and data can be uploaded to existing public databases.

Key takeaways from the workshop included:

Data Quantity

The workshop participants and respondents of the associated RFI generated a total of 49 existing data sets. These data sets were ranked and prioritized using a combination of (1) the impact they would have on the industry and (2) the likelihood of acquiring the data. For the highest-priority data sets, workshop participants suggested next steps for acquiring the data. The most promising data sets across the bioenergy technology space were industry data, with many specific requests for commercial-scale data on feedstock quality changes under different operating conditions between harvest and conversion to an energy product.

Data Quality

Availability, completeness, and context of data and metadata were identified as the most important factors in determining data quality and usefulness. And it was emphasized that the usefulness of any given data set is highly dependent on who will use the data and for what purpose. Workshop participants listed 31 data quality metrics with 94 suggested processes for determining quality.

Data Monetization and Valuation

Workshop participants suggested that data valuation could start with determining fit for use or fit for purpose. Resources discussed for validating data value included independent assessments (e.g., review boards), offers from potential buyers, market analyses, or comparable recent transactions. Establishing a general procedure for valuation of bioenergy data could be useful in the future.

Data Acquisition

There are several opportunities for potential data buyers to participate in or improve current data acquisition procedures, including attending data room sales with the objective of reducing the rate that data goes fallow and giving new opportunities to industry to sell information. An exchange platform for different parties to connect and interact could be helpful, serving as a neutral broker and defining rules or requirements to meet requisite quality.

Table of Contents

Executive Summary vii
Introduction 1
Plenary Session Summary 4
Open Forum Presentation Overview 10
Breakout Session Overview 10
Data Quality 11
Overview 11
Sample Public Databases 12
Determining Data Usefulness 13
Data Quality Metrics and Processes 14
Implementation Roles and Procedures 20
Data Quality Conclusions 21
Data Acquisition 21
Overview 21
Data Acquisition Approaches Based on Owner Classifications 22
Adapted Data Acquisition Approaches 23
Data Acquisition Conclusions 24
Data Monetization and Valuation 25
Overview 25
Fit for Purpose and Use 25
Data Set Valuation Examples 26
Data Valuation Factors 26
Data Valuation Strategies 27
Data Monetization and Valuation Conclusions 27
Existing Bioenergy Data 28
Feedstock Handling and Biorefineries 29
Thermochemical Conversion 31
Microorganisms in Biotechnology 33
Algae 34
Breakout Session Report Outs 37
Summary and Next Steps 38
Appendix A: Agenda 40
Appendix B: Registrant List 42

List of Figures

Figure 1. Workshop focused on currently used non-sensitive data and previously used data..... 1
Figure 2. Workshop registrants by affiliation 2
Figure 3. Workshop registrants' indication of previous BETO workshop attendance 3
Figure 4. Workshop participant data transaction experience 4
Figure 5. Average degree of difficulty for determining data usefulness 14
Figure 6. Required data quality rigor 15
Figure 7. Workshop participant group photo 37

List of Tables

Table 1. Workshop Plenary Speakers iv
Table 2. Breakout Session Leaders v
Table 3. Participant Estimates of Sources for Potentially Acquirable Bioenergy Data Sets 28
Table 4. Feedstock Handling and Biorefinery Data Sets 29
Table 5. Thermochemical Conversion Data Sets 32
Table 6. Types and Sources of Potentially Acquirable Algae Data Sets 35

Introduction

On July 21–23, 2020, the U.S. Department of Energy (DOE) Bioenergy Technologies Office (BETO) hosted the "Leveraging Existing Bioenergy Data" online workshop to discuss how to collect and valorize underused data sets and associated knowledge, with the goal of making this information public on existing databases. This workshop sought to connect industry scientists, data owners, and lawyers with representatives from the federal government, academia, and national laboratories to shepherd valuable data sets and other knowledge to be used to maximum benefit. The intended outcome would be to bolster the growing bioeconomy with industrially relevant data across the supply chain, resulting in accelerated development and utilization of biotechnologies.

The specific focus of this workshop was the finite subset of data that lies in the middle of the spectrum between highly sensitive trade secrets and fully open-access data (Figure 1). The thought behind this workshop was that there exists information that, if acquired and made public, would *not* decrease the competitive edge of any current organizations, but would be extremely useful to stakeholders in many sectors of the bioeconomy. Conversely, although there is much work to be done to help technically public data be made more accessible, this workshop was solely focused on the processes required to disclose truly inaccessible data. The focus of this workshop was partially informed by responses to a Request for Information (RFI) released by BETO in August 2019, in which respondents largely believed that there are existing bioenergy data that are not public and/or not widely distributed with potential utility to help current researchers and advance the field, but that data quality is extremely difficult to assess.

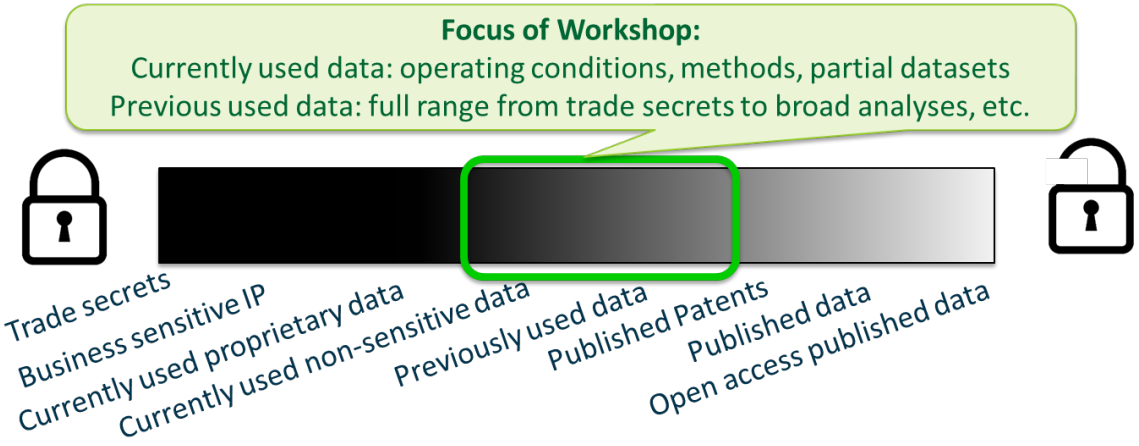


Figure 1. Workshop focused on currently used non-sensitive data and previously used data

Participants of this virtual workshop listened to speaker presentations on topics related to data management, access, legal considerations, and economics and monetization. Additionally, several attendees presented their own perspectives via open-forum 3 × 5 talks (3 slides, 5 minutes). Finally, participants contributed input through a series of breakout sessions.

Workshop attendees represented various bioenergy stakeholders including government organizations (federal and state), national laboratories and research institutes, universities, law firms, data management firms, and bioenergy production companies. The breakdown of the 189 registrants shows that participants were evenly distributed among these sectors (Figure 2). Of the government attendees, at least six different DOE offices and five additional agencies were represented, likely because the overall workshop goal of acquiring existing data will be applicable to any agency funding research and development. Only about half of the participants had attended a previous BETO workshop (Figure 3) because in addition to experienced bioenergy stakeholders, BETO solicited input from data valorization experts, venture capitalists, and lawyers familiar with startups, intellectual property, and bankruptcy.

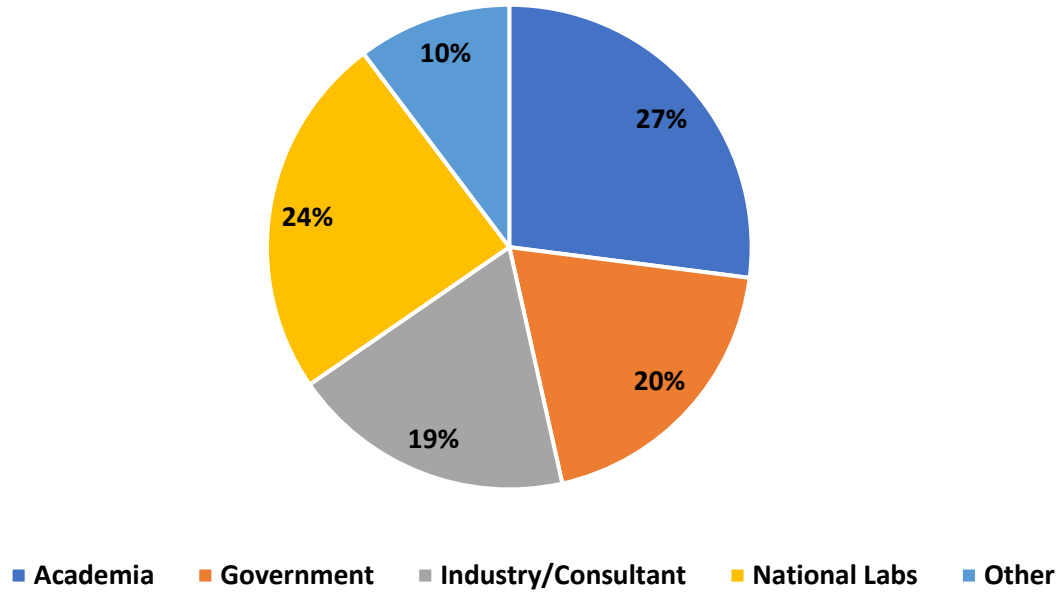
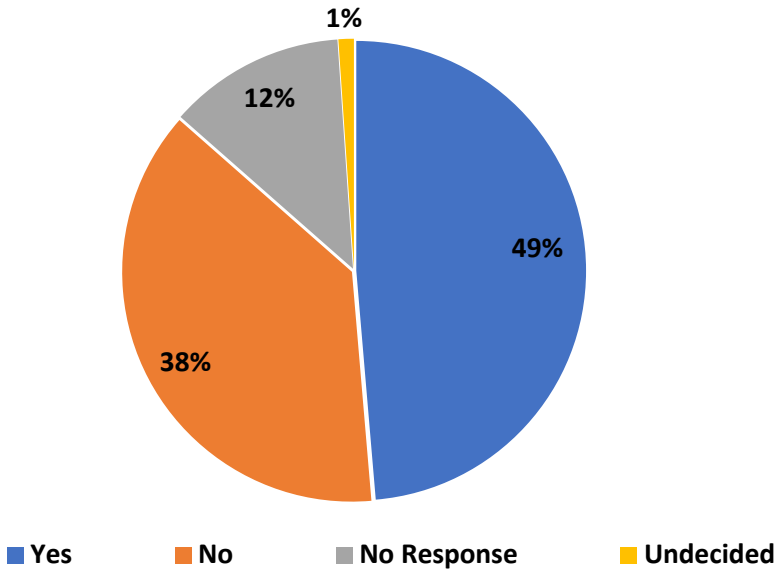


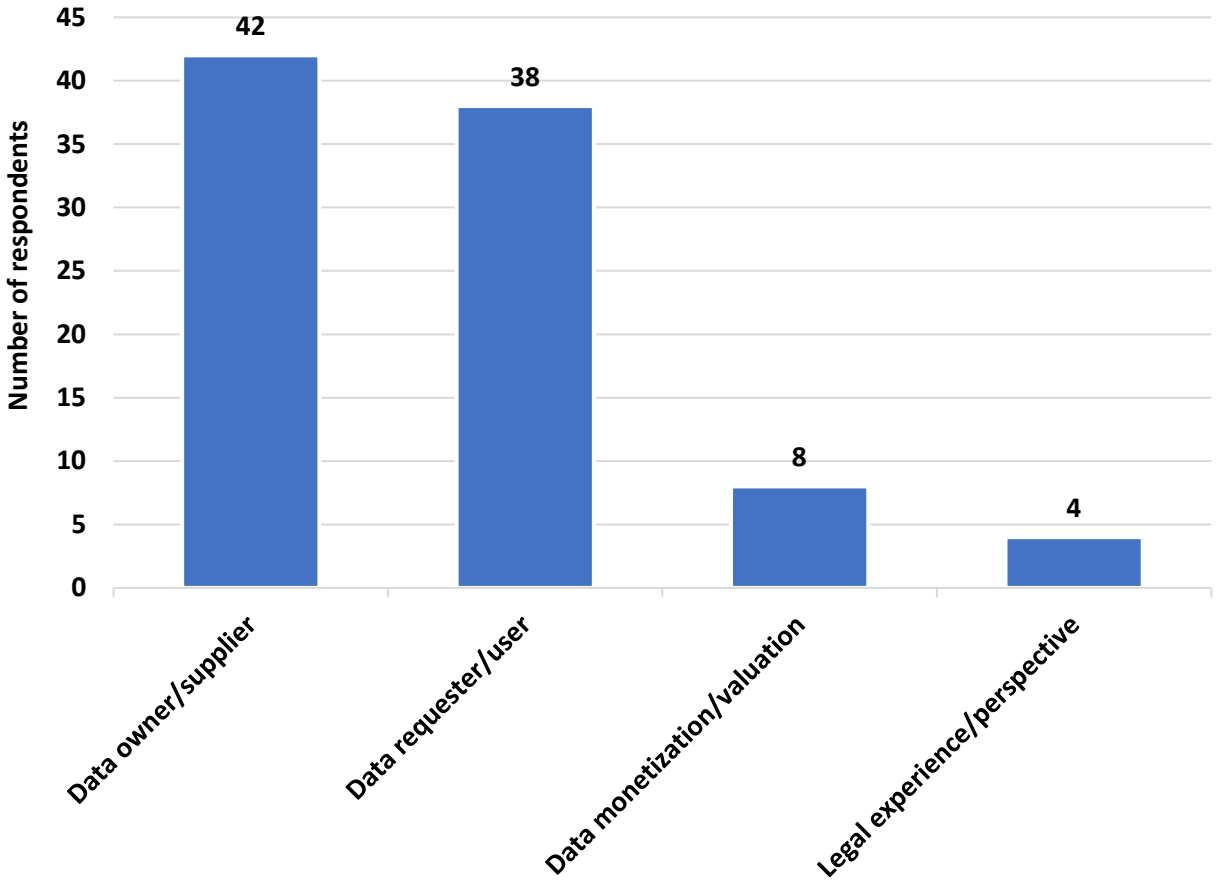
Figure 2. Workshop registrants by affiliation



Percentage of respondents out of 189 total participants.

Figure 3. Workshop registrants' indication of previous BETO workshop attendance

During a short networking session at the beginning of the event, participants responded to a poll indicating their experience related to data transactions (Figure 4). There was a nearly even spread between data users and data suppliers, which is promising for the success of future data-sharing efforts.



70 participants responded to this question; they were each able to choose more than one answer.

Figure 4. Workshop participant data transaction experience

Participants also responded to questions during the networking session about what they hoped to learn or offer at the workshop, and whether they had an "aha" moment regarding data sharing. Twelve respondents indicated that they had, and shared remarkable stories of the value of reusing data. One participant was eager to offer cellulosic ethanol scale-up data, and several participants had deep expertise on preparing data for sale in data rooms.

Plenary Session Summary

The first day of the workshop began with several plenary presentations meant to provide expert insight on each of the key components needed to realize the vision of acquiring existing data and making it publicly available. These included presentations on what would be possible within DOE's legal framework, perspectives on trade secret law, overviews of the types of existing bioenergy data and the subsequent reuse value, efforts to evaluate and monetize data, and current public databases and data exchange platforms.

U.S. Department of Energy Legal Perspective

Julia Moody, Deputy Chief Counsel for Intellectual Property at DOE's Golden Field Office, and Kim Graber, Legal Counsel at DOE, gave a presentation about government data, including the mechanisms through which DOE develops data, government data rights, and general guidelines for when DOE is willing to pay for data. The projects that DOE fund produce two types of data: data produced at least partially with government funds and proprietary data produced with all private funds, outside of any government funding. For this first type of data, DOE is generally required to make all data and findings public and has unlimited rights to these data. For the second type of data, DOE still has limited or restricted rights. A few statutes allow for limited protection of data that would have been a trade secret if privately funded for a period of up to 5 years.[1] This option to delay publication of results for 5 years is a way to aid commercialization of taxpayer-funded technology. Ms. Moody and Ms. Graber emphasized that because DOE-funded data are already required to be made public, it would be extremely important to determine that data purchased or collected as part of the effort were not created as a result of government-funded research (i.e., the government will not pay for data twice).

Perspectives on the Life Cycle of Critical Data Assets in Technology Development and Commercialization

John Ellersick, Founder and President of Next Rung Technology, gave a presentation on the life cycle of bioenergy data, including its creation, valorization, and disposition. The primary question for the audience's consideration was whether any data are valuable and to what degree they are useful. Mr. Ellersick discussed the background of how a few example technologies developed and how the changing landscape of sustainability has affected sustainability targets, which included examples of sustainable technologies from previous decades. Mr. Ellersick then reviewed the life cycle of technology development, including the various stages and who is typically involved, and the types of data available at each stage. He noted that various things are produced at the stages of technology development that may not be traditionally considered data; however, it is challenging to sift what is high-value from what is low-value. Mr. Ellersick suggested that data (broadly defined) value is not binary, but rather exists on a spectrum of value and presented a color-coded schema of data value. He presented more than 30 types of data, color-coded according to his suggested schema, and suggested potential sources for these data. To conclude his remarks, Mr. Ellersick shared the outcomes of some of the earlier technologies from the beginning of the presentation and what ultimately happened to the data, important sources of data from years past, and some questions to consider moving forward.

One workshop attendee asked a follow-up question about one of the technology development facilities discussed during Mr. Ellersick's presentation: When that facility was transferred to new ownership, how was the operational know-how provided to the new owners, if it was transferred? Mr. Ellersick said that plant operating information was transferred, along with

[1] Energy Policy Act of 1992, Pub. L. No. 102-486, 106 Stat. 2776 (1992); Energy Policy Act of 2005, Pub. L. No. 109-58, 119 Stat. 594 (2005).

limited process information, but information like material balances were probably not transferred, and the intellectual property (IP) stayed with the original owners.

Another participant asked what happened to operational data and IP for companies that had filed for Chapter 11 bankruptcy. Mr. Ellersick stated that he did not have this personal experience, but that in his acquisition experience, sellers valued IP above all else in the sale of data.

The National Microbiome Data Collective: Building a FAIR Data Resource

Kjiersten Fagnan, the National Microbiome Data Collective (NMDC) Infrastructure Lead at DOE's Joint Genome Institute, gave a presentation on the work NMDC is doing to advance microbiome science by building an open-source, agile, integrated data system. Dr. Fagnan discussed the cost of data, including the need for contextual information to use data. She noted the need to bring large amounts of data together to run statistical sign analyses and that metadata and other information about data frequently does not find its way to the scientists. According to Dr. Fagnan, the scale of omics data is immense; if an easy way for scientists to submit their data and metadata exists, we would be able to answer really interesting questions. Access to a large amount of omics data would allow scientists to explore a broad range of hypotheses. The goal of NMDC, therefore, is to remove barriers to accessing high-quality microbiome data.

NMDC has two strategic priorities: establishing infrastructure and engaging with scientists. NMDC plans to create an inclusive engagement strategy, which will include partnering with research teams, both by collaborating with teams to support individual projects and by working across research programs and initiatives; leveraging societies to network with broad stakeholder groups; engaging funders to support data management plans across agencies; and allying with journals to improve links across data and publications. NMDC also plans to create an integrative data infrastructure, including standards and expert curation (minimal information about any sequence and mapped ontologies and harmonized sample metadata); Findable, Accessible, Interoperable, and Reusable (FAIR) data; standardized open-source workflows for omics data; and streamlined search and accessibility. Dr. Fagnan also discussed some of the challenges associated with metadata (or lack thereof), including determining from the beginning what another researcher would need to be able to use these data, and how to make data discoverable by both machines and humans. Dr. Fagnan suggested that research sponsors should allocate a portion of award funding to collecting metadata, and that the ability to reuse data be aligned with publishing. Ultimately, NMDC envisions the infrastructure being built as a distributed network of nodes, linked by a central metadata store, contributing data and metadata that would be accessed through some sort of portal or platform, with a solid system for tracking metadata associated with a data set.

A workshop attendee asked the following: One of the first steps is establishing standards for data. How have you gone about ensuring that those standards are broadly disseminated and updated? How have you been working with a wide variety of universities to incentivize, or police them, to adhere to those standards? Dr. Fagnan replied that NMDC ran a workshop last fall as an engagement effort to stress the importance of the intricacies that they were trying to establish;

attendees for that event included board members from standards creators. Dr. Fagnan said that leveraging the relationship between those groups to standardize the metadata is key, which is a large focus of their community engagement. NMDC is also looking at data already generated by the Joint Genome Institute, which still contains large gaps in the metadata. NMDC has 600 metadata terms for which they need values; when they do outreach, they hope to obtain at least subsets of those metadata requirements. NMDC is trying to work with the boards for those groups that develop standards and hold workshops to engage. They are also examining multiple schemas to determine what works.

Another participant asked how NMDC has considered this from a user interface perspective. Dr. Fagnan noted that this is challenging, but is part of the reason why there was a huge financial investment at the start of the project. NMDC is looking into what the industry is already using and working with a team skilled in this area. Thinking about different ways to query, NMDC is also querying content and analysis, by gene or by species, and linking those back to the samples. Presenting these findings back to the user will be a challenge; the process will be iterative, take time, and hinge on good metadata. They are also planning demonstrations. Further, NMDC is considering metadata quality scores and doing an analysis on how frequently lower-quality data are excluded in search results. Ultimately, they will also need to use stories about how users interact with the searches.

Another workshop participant asked what differences NMDC has experienced with regard to mining preexisting data versus active and ongoing input, and how they capture older data sets. Dr. Fagnan said that reprocessing existing data is not problematic—they are able to take sequence data, run a quality control, and reprocess it to create higher-quality outputs—but the greater challenge comes from gaps in the metadata. Reprocessing existing data sets provided insight into what is really important. NMDC has a few active collaborations to determine the balance of how much metadata can be collected and improve past data; this is something that NMDC is still working to determine. It is a huge effort to go back and curate old data sets by hand, and another option would be to look forward and disregard old data; however, that results in a huge loss of investment.

Trade Secret Law, Licensing, and Acquiring Bioenergy Data

Charles Tait Graves, partner at Wilson Sonsini Goodrich & Rosati, gave a presentation on trade secrets and the mechanisms for transferring trade secrets. Mr. Graves discussed the basics of trade secrets, highlighting that trade secrets are the broadest category of IP but may be the weakest. A trade secret is defined as (1) secret (not published or known in the industry, not readily ascertainable/not easily duplicated, protected by its owner, and disclosed only under non-disclosure agreements) and (2) current (not stale or obsolete). Trade secrets differ from patents in a number of ways: trade secrets do not need to be registered with a government, are not protected, are not typically written down (i.e., companies do not typically keep lists of trade secrets), are a broader category than patents, and can be defeated (i.e., another party can publish or make public the same or similar information). Trade secrets will last until published, which

could be potentially in perpetuity, but trade secrets can be defeated at any time. Mr. Graves also discussed trade secret licensing. Trade secrets are property and as such, they may be acquired like any other property. However, there is typically a ceiling on the licensing or acquisition price a buyer might pay, due to the fact that it is theoretically possible to do the research and development to develop the secret yourself. Potential trade secret buyers will do their homework prior to negotiations to make sure that the sale price is appropriate. Mr. Graves discussed when bioenergy information may be a trade secret. There is a distinction between internally developed or created information versus information that a company has gathered from other sources; in the latter, it is possible to go back and compile this information again, but in the case of the former, it is difficult or impossible for others to replicate this data. Trade secrets tend to be lab experiments that have not been published. Bioenergy data assets are typically acquired from existing companies, usually under contract rather than by trade secret rights; it is often easier to take everything than negotiate individual items. In such negotiations, Mr. Graves reminded the audience to be sure to obtain adequate representations and warranties regarding conflicting IP claims from other parties.

A workshop attendee asked Mr. Graves about the event of a company going through a bankruptcy or merger process where they are establishing a data room and liquidating assets, and what he would recommend to data generators as a way to advertise a trade secret, since advertising the knowledge will reduce its value. Mr. Graves replied that only in a formal bankruptcy would there be a public notice of sale; the company would not lose the rights by announcing the sale and would be able to negotiate confidentiality through normal means. In all other cases, the sale would be through contract and would not be advertised publicly.

Leveraging the Value of Data Assets in the Bioeconomy through Better Data Circulation and Monetization

Didier Navez, Senior Vice President of Strategy and Alliances at Dawex, gave a presentation on Dawex's work in the data exchange area. Dawex's mission is to create the conditions for the smooth development of the data economy by facilitating the exchange of data among companies and organizations. Dawex works as both a data provider, operating a global data marketplace, and a data exchange platform for corporations, consortia, or public organizations to use to share and monetize data internally or externally. Dawex built a platform technology for organizations to use to operate data exchanges, which they themselves also use to sell data. Mr. Navez discussed data exchange, its importance, and its value. The volume of data being produced globally is growing exponentially, and it is reshaping our world. Data is essential for a variety of applications and uses, and its value lies in its use and reuse, but there are several barriers to data availability. These include fragmentation of data sources, lack of trust between economic operators, imbalances in negotiating power, fear of data misappropriation by third parties, lack of legal clarity on who can do what with the data, and lack of private sector data availability for use by the public sector. Dawex believes that data exchanges can provide a number of functions on both the supply and demand side to alleviate these issues. Data exchanges can circulate data at scale across sectors and borders, creating direct and indirect economic value. Companies that are

data savvy are more valuable on the market; the most valuable business use cases combine internal and external data sources to create actionable insights. The data exchange can automate and industrialize these processes through specialized procurement, marketing, sales, legal, logistics, and IT processes, providing a building block in the global data value chain. Mr. Navez then discussed several data use cases from around the world.

Infonomics: The New Economics of Information

Doug Laney, Principal Data Strategist at Caserta, gave a presentation on the economics of information, or "infonomics." Mr. Laney began his presentation with a rebuttal to the popular idea that "information is the new oil," pointing out that data is nondepleting and regenerative, easy to store, move, and share, but also easy to steal and impossible to clean up if spilled. Then, Mr. Laney introduced the concept of infonomics, or treating information as an asset, including monetizing, managing, and measuring information. Monetizing information is generating economic benefits, both direct and indirect, from data. Examples of economic benefits from data include selling, bartering, or trading information; using information to enhance products or services; improving process performance or effectiveness; and using data to develop new solutions. Mr. Laney shared specific examples of data monetization information valuation models. These included models for ascertaining foundational measures of data value:

- The intrinsic value of information, or how correct, complete, and exclusive the data are
- The business value of the information, or how good and relevant the data are for specific purposes
- The performance value of information, or how the data affect key business drivers.

Mr. Laney also discussed models for the financial value of information, including how to determine:

- The cost value of information, or what it would cost if the data were lost
- The market value of information, or what might be obtained from selling/trading the data
- The economic value of information, or how the data contribute to the bottom line.

Mr. Laney then shared ways to apply the information valuation models, including ways to guide investment, determine where additional value can be captured, and manage the life cycle expenses of data appropriately. He also made broad recommendations for how to manage, think about, and value business information going forward.

Ensuring Bioenergy Data Can Be Accessible, Usable, and Useful

Debbie Brodt-Giles, Group Manager for Data, Analytics, Tools, and Applications at the Strategic Energy Analysis Center at the National Renewable Energy Laboratory, gave a presentation on the FAIR principles, why these are important, and their applications. The FAIR principles refer to the idea that federally funded data should be Findable, Accessible, Interoperable, and Reusable so that they may spur innovation, reduce duplicative work, and

facilitate rapid advancements of industries and technologies. Findable data refers to data and metadata that are easy to find for both humans and machines; these should include a globally unique and persistent identifier or a digital object identifier (DOI) and be registered or indexed in a searchable resource. Accessible data are data that people are able to understand how to access once they have found the data; should be retrievable by the identifier using an open, free, and accessible standard communications protocol; and that are available for long-term use and citation. Interoperable data are easily integrated with other data and/or applications or workflows for analysis, storage, and processing. Reusable data are optimized for reuse and reutilization by being well defined, providing provenance about the data sources, and allowing for others to build on the data. Ms. Brodt-Giles offered a few examples of data operating under FAIR principles, including OpenEI and its subsidiaries, such as the Marine and Hydrokinetic Data Repository and the Geothermal Data Repository. Ms. Brodt-Giles concluded her presentation by offering next steps toward applying FAIR principles to bioenergy data.

Open Forum Presentation Overview

In addition to the plenary session presentations, nine snapshot presentations, also known as 3 × 5 presentations (3 slides, 5 minutes), were given by experts in their field during the first day of the workshop. These presentations were chosen to complement the plenary presentations by adding perspectives on all major aspects of the workshop, including IP law, data sharing, and case studies in existing underused data:

- Intellectual Property: Types, Eligibility, and Protection, *Charles Naggar, Alston & Bird LLP*
- Stranded Data from KiOR, *Bruce Adkins, Oak Ridge National Laboratory*
- Scale-Up Data: A Hidden Asset, *Joe Sagues, North Carolina State University*
- Knowledge Representation to Capture Lessons Learned in Bioprocessing, *Deepti Tanjore, Lawrence Berkeley National Laboratory*
- Data Qualification Framework, *Rachel Emerson, Idaho National Laboratory*
- Building Fungal and Algal Multi-omics, *Igor Grigoriev, DOE Joint Genome Institute*
- Computational Catalyst Property Database and Catalyst Deactivation, *Carrie Farberow, National Renewable Energy Laboratory*
- Time and the Value of Data, *Bruce Wilson, Oak Ridge National Laboratory*
- Generating and Transferring Technology to Fill Knowledge Gaps, *Vijaya Gopal Kakani, Oklahoma State University*.

Breakout Session Overview

The online workshop featured a total of seven breakout sessions. The highlights of each of these sessions are summarized in their own separate sections of this report.

Each session focused on high-level questions designed to let the group effectively contribute their specific input. A facilitator guided the group through session activities, including short-answer questions and quantitative assessments. All input was visible in real time for participant review and response via the ThinkTank collaboration software tool.

On the second day of the event, participants provided their input on data quality, acquisition, and valuation. The objective of these sessions was to identify data quality metrics, acquisition strategies, and valuation approaches that can be applicable to a variety of bioenergy technology markets.

On the third day of the event, participants joined small groups focused on the specific technology areas: (1) feedstock handling and biorefineries, (2) thermochemical conversion, (3) microorganisms in biotechnology, and (4) algae. The objective of these sessions was to identify potential data sets for acquisition and next steps to support public access.

Data Quality

Overview

The objective of this session was to establish metrics, processes, procedures, and roles to determine the quality and/or usefulness of existing data sets. This breakout session expanded upon points regarding data quality made previously by respondents to the 2019 RFI, including that data quality is difficult to ascertain, level of quality needed is ultimately defined by the user, maintaining quality data is related to organizational culture, and involving those knowledgeable about the data is critical. The RFI respondents also suggested the ideas of including third-party assessment of data quality (potentially including multiple perspectives) and building upon existing rubrics for determining quality, such as FAIR. Discussions focused on assessing quality, specifically related to data that already exist (not how to make future data better). Workshop organizers set parameters, including an assumption that the data are legally accessible (as developed in the Data Acquisition breakout session) and a preference for data that are non-sensitive but not published.

Importantly, the distinction was made between the term “data quality” as it relates to the ability for data to be used by a new research group compared to a standard definition of the term that refers more directly to data cleanliness, accuracy, or completeness. For clarity throughout the discussions, groups used the term “useful” to describe data (or associated knowledge) that may be utilized to inform or advance new research.

Nearly 50 bioenergy stakeholders representing DOE national laboratories, universities, industry, and consultants joined the session. Of these, 27 identified themselves in the collaboration software and 38 attendees responded to a poll identifying their roles in data transactions. Respondents were able to choose more than one role, and 55% identified as data owners/suppliers, 72% as users/requesters, 3% as brokers, 31% as reviewers, and 21% as other.

In some cases, the data requester will already know what they are looking for, and generally trust the quality of the data they are requesting. In other cases, the data supplier will need to convince the community of the quality and usefulness of their data. As one participant suggested, this may involve generating case studies to show how to use the data, including demonstration of models or algorithms, examples of process scale-up, or business intelligence storytelling.

Sample Public Databases

Participants were initially asked to identify existing avenues for data sharing and vetting. They suggested:

- Bioenergy Knowledge Discovery Framework
- National Center for Biotechnology Information
- Bioenergy Feedstock Library
- Joint Genome Institute databases (PhycoCosm, MycoCosm)
- Phyllis 2 (Energy Research Centre of the Netherlands) Biomass Database
    - ECN.TNO does a quick check before uploading and serves as gatekeeper. One must download and use ECN.TNO's Microsoft Excel template to enter data.
    - Excel data are then sent to ECN.TNO for review. It is not fully automated or self-service.
- SILVA rRNA Database Project
- United States Patent and Trademark Office Published Patent/Patent Applications
    - Currently, vetting this data set is manual.
- Mendeley Data
- Data.gov (United States)
- Open Energy Data at DOE
- DOE Energy Data eXchange
    - Tool is designed for users to upload data and then work through a quality assurance and approval process to make their data public.
- U.S. Department of Agriculture Natural Resources Conservation Services Soils SSURGO Database
    - Contains information about soil as collected by the National Cooperative Soil Survey.
- National Oceanic and Atmospheric Administration Climate Data Online
    - Archive of global historical weather and climate data.

- EMODnet Human Activities
- Databases within BETO Consortia (i.e., Agile BioFoundry, Feedstock-Conversion Interface Consortium, Chemical Catalysis for Bioenergy)
- Federal LCA Commons
    - Access to a collection of data repositories for use in life cycle assessment.
- National Renewable Energy Laboratory Biomass Compositional Analysis Laboratory Procedures
    - Requires data suppliers to vet the procedures through interlaboratory studies.
- AgMIP (Agricultural Model Intercomparison and Improvement Project)
    - This data interchange was designed to be the central repository for AgMIP data used in and resulting from modeling activities.

Determining Data Usefulness

The first part of this discussion considered the level of difficulty for determining data usefulness for different data types. Participants indicated that the quality of data from large-scale tests was the most difficult to determine, whereas the quality of standard operating procedures would be the least difficult to determine (Figure 5). Data derived from experimental data but with independent quality attributes, such as techno-economic analysis and life cycle analysis, may also be easier to assess and use in a new research setting.

Figure 5 shows the average of 32 participants' rankings. The scale was from 0–5, with the most difficult data sets to assess usefulness (and/or quality) assigned as 5 and the least difficult assigned as 0. Results had a 25%–30% relative standard deviation. Even the data types ranked most difficult to determine quality averaged 2.5 out of 5, suggesting that the participants were overall optimistic about being able to determine data quality.

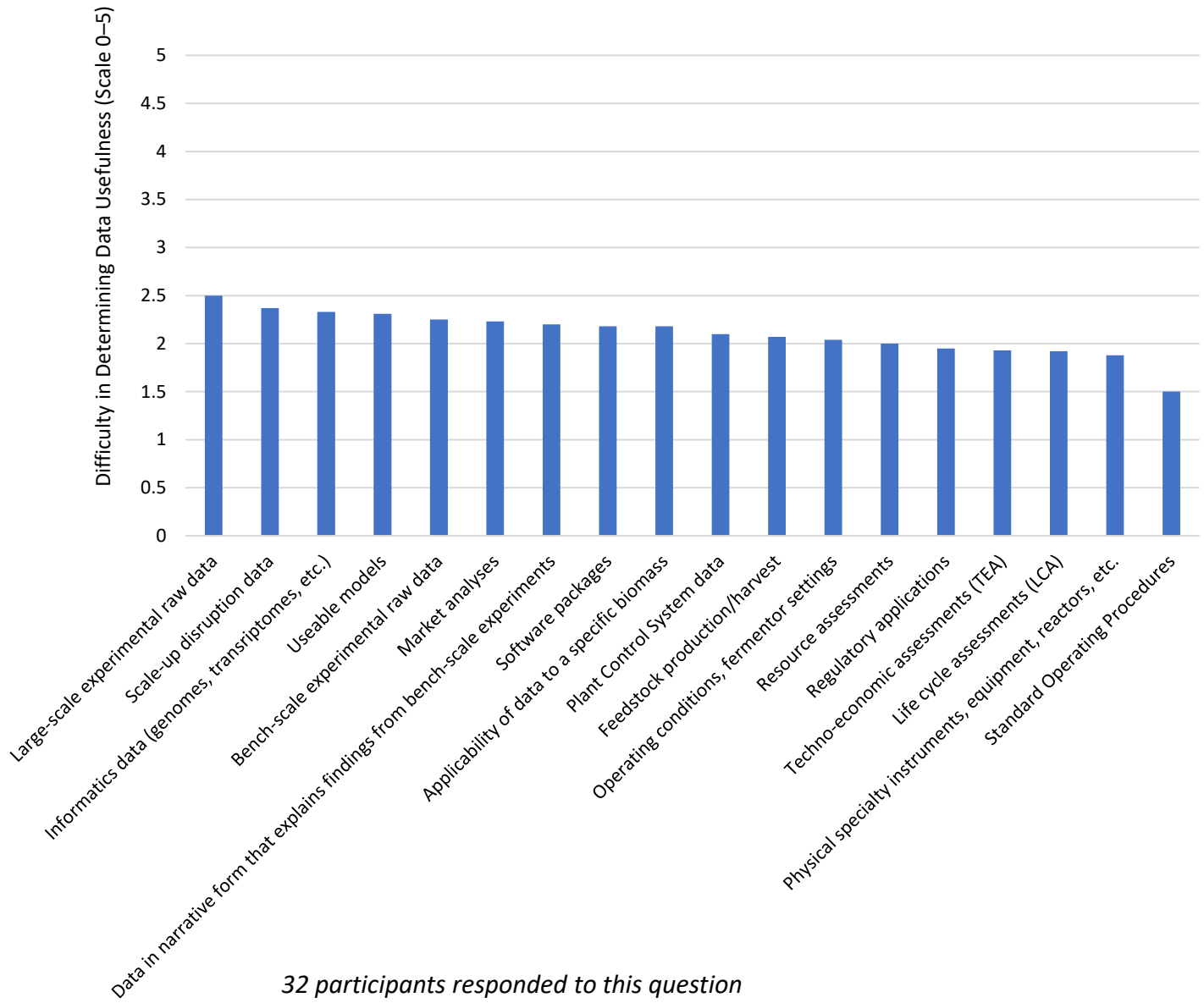


Figure 5. Average degree of difficulty for determining data usefulness

Data Quality Metrics and Processes

The breakout session group identified over 30 different data quality metrics and ultimately focused on relationships between data quality rigor and potential impact of a given data set. Participants suggested that metrics and processes depend on end use, scale, and technical field. Discussions also pointed out that during initial technology development, higher data-quality rigor is required for higher-impact projects, such as due diligence during operational scale-up (see Figure 6 for an illustration of this relationship). As one participant described this, the closer to commercialization and demonstration (i.e., the higher the technology readiness level), the higher

the bar is for data quality and the greater the liability is for low-quality data. This suggests that there may be better documentation for the more valuable, large-scale data sets.

Other findings from the contributors were that context and metadata are critically important in addition to traditional quality assurance and quality control; participants emphasized using the SI system (International System of Units) to share metadata when possible. Additionally, they suggested using keyword searching for large narrative data sets, such as operator logs and weekly status updates, which can contain valuable information. Finally, participants noted that one way to address incomplete metadata would be if funding agencies and/or journal publishers mandated that complete metadata be part of any final report or publication.

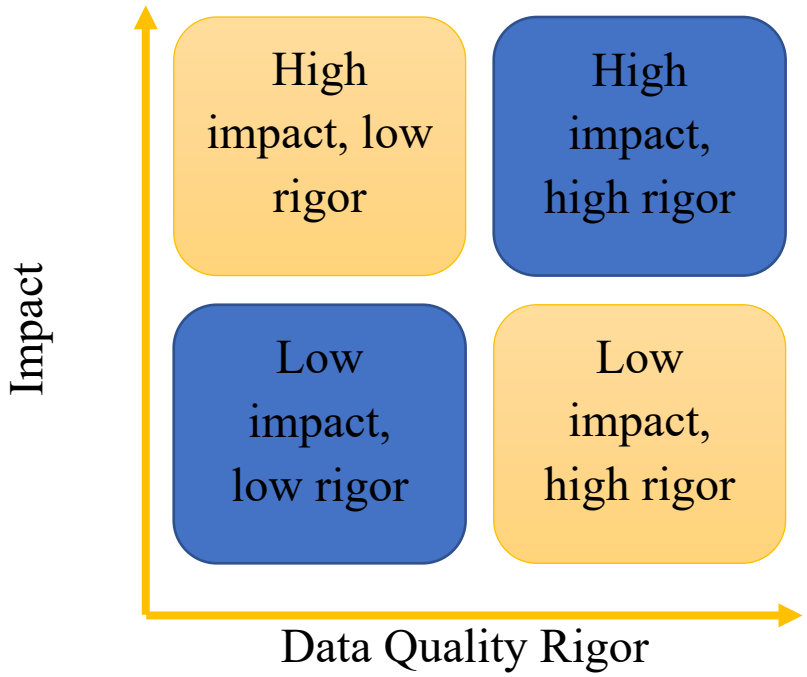


Figure 6. Required data quality rigor

For each proposed metric or quality component, the group discussed which processes could be utilized to collect the information. The metrics and processes that the group listed during a brainstorming activity include:

- Sufficiency of data
  - Elements of a complete data set may include the list of variables for each data type, meaningful data dictionary, batch numbers or lot numbers, time stamps to tie offline data with online data (especially for plants), and confirmation of relevant parameters for each type of data set (e.g., batch vs. continuous experiments, as they do not require the same parameters).
  - Potential data user can use a tool that scans through all of the files to assess what variables have been collected and then display those as part of a data dictionary. This

is a standard part of some publication processes. This is a key factor for users to assess whether or not the data are relevant to their particular need.

- Supporting documentation quality is often used to assess the underlying data quality
- Demonstration or diagram of complete data collection and processing scheme is needed
    - Data provider should describe the laboratory equipment utilized and the context in which data was collected
    - Describe accuracy and detection limits of instruments and any relevant instrument certifications
    - Include documentation of any custom (noncommercial or precommercial) lab equipment used to generate data
    - Include equipment operation standard operating procedures, scope/build-out reports, and manuals
- A data quality program could be similar to the Nuclear Quality Assurance (NQA-1) Certification
- The activities necessary for confirming a given data set's level of completeness may range from a very cursory review to an extended due diligence period that involves outside companies and tremendous amount of detail
- Source data accuracy for analyses or modeled data sets
    - Overview of statistical analysis that was applied and any relevant statistical results (e.g., standard deviation, standard error, sample size)
    - Description of design of experiments
    - Description of data uncertainty
- Description of statistical process control
- Explicitly stated sampling plan to ensure that analyses are performed on representative samples of whatever material or process is being sampled
    - Description of data collection objectives, including what material or process was being sampled/measured, how the samples were selected (random, time-based, "grab"), whether there was a chain of custody or other formal sample handling plan in place, and if so, was it based on any specific standards (e.g., U.S. Environmental Protection Agency standards)
    - As applicable, differentiate between bench-scale data and process data
- Types and impacts of bias: describe assumptions, hypotheses, and rationale for data selection

    - Report the standards applied
        - Describe the standard methods and the use and analytical results from standard reference materials
        - Report the method. For example, TAPPI, ASTM International (formerly the American Society for Testing and Materials), published laboratory instructions, standard reference materials from the National Institute of Standards and Technology, shared biomass materials, and in-house standards.
        - Standard Reference Materials analyses, ideally presented as a control chart.
        - Calibration and underlying certification (e.g., from the American National Standards Institute, International Organization for Standardization)
        - Independent testing, round-robin tests for reproducibility: same materials, same methods, similar instruments.
- Sufficiency of metadata
    - Assess using the Dublin Core™ Metadata Initiative specifications
    - Assess using FAIR Guiding Principles
    - Develop a standardized form
        - Many of these metrics (especially collection forms and scores) will need to be domain-specific because different applications will require various information. The definition of "complete" will vary by type of data and context. Different sets of forms and templates will be required for bench-scale data and process data.
    - Use a metadata review team
    - Establish "metadata scores" for particularly high-importance types of information, as was done by NMDC
- List of information that would contribute to complete metadata
    - Research scale (bench, pilot, demo, commercial)
    - Description of whether the research was conducted with a batch or continuous process. Additionally, future researchers need information regarding the "time on stream" and steady-state conditions vs. number of batch replicates.
    - For large scales (demo and commercial), the list of process disruptions. This would require metrics for process disruptions, such as mean time between failure, mean time to repair, or some metric of "uptime." These metrics should indicate both scheduled maintenance as well as an allowance for unplanned maintenance/repair.

- List of intellectual property, especially patents, tied to research data, ideally linking to an IP library because it can be difficult to tie specific patents to data sets.
- Consistency of mass and energy balances
    - These are very hard to get in manufacturing plants, so the data user/requestor might need to do the assimilation. Even in demonstration-scale/small production plants, mass and energy balances are crucial to ensure the process is fully understood from a scale-up point of view. Extra effort must be taken that would not be necessary in a fully established production process.
    - Data providers should provide replicated quantification of all substances consumed and produced during the experiment.
- List of publications resulting from data, with DOIs
    - If the list of publications is not provided by the data supplier, leverage a process of searching multiple citation databases (e.g., Google Scholar, Scopus) and free text internet searches to get at citations for a data set. This process requires that the data set has a DOI and is most effective for finding citations for scholarly literature.
- In situations where data have abnormal or unexpected results and don't fit into a specific section of a data entry form, data suppliers can place additional detail into an unused section to add data
- Documented methodology for basic data cleansing
- Vetting with another reference (e.g., canonicalized simplified molecular-input line-entry system [SMILES] representation in new database vs. published canonicalized SMILES representation such as PubChem), including a list of published benchmarks
    - Identification of acceptable/definitive reference (i.e., benchmarking)
- Read quality for sequencing, which can be done using quality-control tools like FastQC
- Developer's interest/ability to answer questions from the data users (e.g., descriptions of the data review process, quality assurance procedures, data quality objectives of the sampling and analysis plan, and outlier rejection criteria in data documentation)
- Involved parties
    - Names of scientists involved with developing data set (links to ORCID, if possible)
    - Names of analysts, technicians, and operators who collected the data
    - Funding agencies that supported data set development and level of funding

- Laboratory notebooks (electronic or hard copy)
  - Notebook utility depends greatly on the quality of the scientists' input, whether they adhere to standard notebook/record-keeping practices.
  - There is a benefit to electronic notebooks (e.g., no messy handwriting to interpret, access to embedded files and spreadsheets, ability to be searched). However, not all e-notebooks are accessible without access to proprietary (or even arcane, unsupported, or extinct) software.
- Spreadsheets
  - Clean files, compile, and assess using R, Python, or other programming software
- Narrative documents with notes
  - Templates supplied by the data exchange host to be completed by the data supplier
  - Templates could include requests for process flow diagrams, mass balances, energy balances, description of available data, explanation of missing data, definition of error bars, etc.
  - Standardized test plans
- Standard operating procedures
- Human-machine interfaces
  - Data supplier should provide continuous data, including amperage, temperature, and pressure, and depending on specifications, recorded analog and digital outputs
- Supervisory control and data acquisition data
  - iFIX and LabVIEW are examples of supervisory control and data acquisition software currently being used. Data from these can then be downloaded to databases.
  - Creating standard formats and ontologies is a key data-management principle
  - Minimizing data cleanup is the ideal. However, we need to balance the value that can be derived from unformatted data (pre-data-management enlightenment) and the need for data cleanup. This might prevent the duplication of previously derived information.
- Open-source/available databases
  - Collaborate with international organizations (e.g., International Energy Agency) and foreign entities (equivalent to BETO) to exchange data.

- Data from PDFs/manuscripts
  - Research data management software
  - LabKey can help integrate various data from Structured Query Language (SQL) databases
  - MongoDB can be used for non-SQL databases
- Software or models (user manuals or instructions critical)
- Case studies to show how to use the data
  - Models or algorithms
  - Examples of process scale-up
  - Specific location information variables (e.g., environment conditions) that affect the experiment
  - Business intelligence storytelling
- Methodology documentation (overview of standard methods used, experiment design, statistical treatment, replication, whether the experiment is reproducible, statements on appropriate data use, and bounds/limitations of data).
  - Gathered from the data provider on a separate form for each metric: methods metadata, technical reports on instrumentation, and/or other key parameters.

Implementation Roles and Procedures

The following is a list of proposed roles and responsibilities related to data quality:

- Data owner/data supplier: Needs to ensure accuracy, works to a user-defined quality checklist, establishes expectations for data use, and manages licensing, copyright, and attribution
- Data producer: Distinguished from owner/supplier as the difference between the scientists/engineers and the organization (subject to change), primarily responsible for downstream data quality
- Data requester/data user: Establishes expectations and judges suitability of the data for the intended use, provides reference uses of data if and when publications are generated, asks about limitations and licensing
- Data quality reviewer: Needs complete data sets or sufficient descriptions of data sets, with methods, standards, protocols, cross-checks, and benchmarks/references
- Data broker: Establishes trust between data users and data suppliers, ensures that legal regulations are followed, and negotiates a fair price for the data given the quality.

Data Quality Conclusions

In summary, the Data Quality session provided DOE with extensive lists of more than 20 relevant established databases housing bioenergy data and more than 30 data quality metrics that will be valuable in supporting greater levels of data sharing. Establishing data usefulness is an important first step because the process of exchanging existing data requires a resource investment that must be justified by the value of the data. Breakout group participants suggested that assessing data quality and/or potential usefulness is feasible for all data types but will require effort, especially for data from large-scale tests.

The group agreed upon a definitional clarification that the term "data" includes narrative data sets such as operator logs and status updates as well as tables of numbers. Additionally, "data quality" should be defined beyond traditional quality assurance/quality control data processes to include context and metadata as critical metrics for determining data usefulness.

Stakeholders expect that the metrics and processes used to assess data quality/usefulness will vary depending upon the end use, scale, and technical field. They also anticipate a higher data-quality rigor will be required for due diligence during scale-up. Generally, as scale, impact, or technology readiness level of a data set increases, the data quality rigor should also increase.

Data Acquisition

Overview

Respondents to the 2019 RFI identified several challenges to obtaining existing underused data sets, including issues with data quality or reusability (particularly over time) and protecting the needs of the data owner (for valorization and IP). RFI respondents suggested focusing on data from public/private partnerships and crosscutting or noncompetitive data as the easiest to obtain. The Data Acquisition breakout session group set out to identify the best processes, procedures, and roles to acquire existing data sets. These ideas were framed around data owner categories and legal classifications. Once a baseline was established, the participants suggested new processes, procedures, and roles to acquire existing data sets. For example:

- Fill gaps in existing data room processes, with the objective of reducing the rate that data goes fallow
- Expand access of data to academia or non-private-sector researchers that are not necessarily well capitalized
- Give new opportunities to industry to sell information, such as useful methods or data with retracted metadata.

Workshop organizers asked attendees to assume that the applicable data has a checklist/cover sheet that conveys the quality and potential usefulness of the resource (as developed in the Data Quality breakout session). The group was told to assume that any data type could be assigned

any given legal classification (e.g., patent, trade secret), depending upon its value to the creator. These data legal classifications were the foundation of the discussion.

There were more than 20 participants in the group, and they represented scientists from national laboratories and government agencies, data scientists, private sector data exchange platform hosts, and academic researchers.

Data Acquisition Approaches Based on Owner Classifications

The following section reviews currently understood acquisition approaches that may apply to various data owners. This participant input provided value by establishing a baseline for understanding how data are acquired and a starting point for considering how gaps can be addressed.

- Venture capital firm: If a venture capital firm has the rights to data and any associated intellectual property, these assets could be spun out into a new startup company. The venture capital firm could also be approached by purchasers interested in specific aspects of the existing data. This could be applicable to data classified as informal proprietary information (assuming it has an acceptable risk factor), engineering packages, and design manuals. Informal proprietary information is data that do not have an official classification and could include a wide range of materials (e.g., operators' logs, capital and operating costs, equipment operation and performance data, regulatory applications).
- Active company (pivoting away from a business unit): The company can grant the acquirer exclusive or nonexclusive licenses to the data. This could be applicable to informal proprietary information as well as formal information.
- Active company (interested in monetizing their data/info): The company can offer buyers either exclusive or nonexclusive license or outright purchase of data. This could be applicable to formal and informal proprietary information. Breakout session participants noted that in this situation, the company should set up comprehensive templates to illustrate that all sorts of process parameters and information were at least considered, if not collected.
- Company that filed for Chapter 7 bankruptcy: Banks are likely to be represented by a broker, who will facilitate the transaction. Banks often use brokers that specialize in the company's specific field because they have specialized knowledge and skills tailored to these types of transactions.
- Company that filed for Chapter 11 bankruptcy: An appointed trustee brokers the transaction of intellectual property (commonly a data room).
- Company that emerged from Chapter 11 bankruptcy: The potential data buyer can coordinate directly with the investor who is effectively the data owner, which is typically a venture capital firm. Session attendees suggested waiting until all lawsuits are resolved

because ongoing litigation regarding disbursement of assets/IP can be a barrier to moving the patents to additional users. Participants also noted that the venture capital firm may have immediate interest in monetizing the patent portfolio, so that may be acquirable while litigation is ongoing; however, if the owner thinks the greatest value lies in the combination of patents and proprietary databases, there may be no choice but to wait out the resolution of lawsuits.

- "Failed" company that has not (or not yet) filed for bankruptcy: Typically approached directly by a potential buyer performing due diligence and looking for ways to optimize the business or to turn it around to profitability. Session participants suggested that these types of buyers are predominately interested in specific technology or talent and not necessarily traditional "data."
- University, government, or nonprofit laboratory: This may be handled by the organization's technology transfer office, if they have one.
    - For university-owned data, there are generally licensing processes established by the technology transfer office. Session participants noted that these offices often actively try to license patents, even if they are pending. There is a general recognition that pending patents have value, but is unclear how this affects valuation. One attendee explained that the value is more about the quality of the patent and whether there is a continuation application that can add further claims to cover the space more broadly.
    - For federally funded research and development organizations, data sharing or acquisition can occur through developing cooperative research and development projects. This often involves licensing agreements between national laboratories and other institutions.

Adapted Data Acquisition Approaches

After establishing a baseline for currently used data acquisition approaches, the breakout group was asked to identify adaptations for each existing process gap. For each type of data owner, they were asked to consider how the current processes might be adapted to give the government the ability to acquire data and make it available on existing public databases.

The group suggested that DOE could provide a simple way for different parties to connect and interact, serving as a neutral broker and defining rules or requirements to meet requisite quality. Breakout session attendees noted that this type of central database of available data sets would be useful for many of the acquisition approaches discussed. Potential acquirers could access such a database to view available data, and data offerors could access resources to assist with facilitating transactions. This workshop was designed with the idea that the potential acquirer would be the federal government, with the intent to make the data public, but this type of centralized database could be useful for data that the government considered but did not purchase.

Participants noted the advantages of a centralized database, especially if the database assumed a two-sided platform/marketplace approach and could facilitate the search and discovery process for potential data users. Additionally, the database platform could offer data licensing advising, which requires specialized skills that not all organizations have available. Another benefit of a data exchange platform could be simplifying the process by proposing mechanisms to easily produce fair, balanced license agreements (based on templates and configurable options) that parties can rely on, while at the same time allow the parties to use their own license (often proposed by the data providers). Further, the operator of the platform could assume the role of inviting data providers, vetting them (based on defined terms of service), and onboarding them on the platform. The platform operator would also be responsible for defining and implementing the acquisition strategy. As discussed in the Data Quality breakout session, there are clear challenges in assembling enough background information on each available data set to make a centralized database worthwhile.

Discussions also covered suggestions for potential data users on how to conduct their data acquisition process. Attendees suggested beginning with conducting market research to find companies with potentially useful data sets and directly contacting them. Then, the potential data user should conduct due diligence and vet the organization that developed the data. Next, depending on the acquirer and intended use case, certain characteristics of the data would need to be validated, as discussed in the Data Quality section. Responses to the 2019 RFI suggested that broad data availability in neutral platforms with easy access may incentivize data providers to contribute their data.

As part of the data validation process, the acquirer should keep in mind that there may be concerns with data quality associated with pilot-scale data (e.g., calibrations, lack of mass balance closures, missing elements). If the data are deemed fit for use, the acquirer could proceed with evaluating licensing terms such as duration, territory, business sector, permitted usages, sublicensing rights, and exclusivity clauses. It may be worth considering developing a trial license first before a full data license agreement.

Data Acquisition Conclusions

The focus of this discussion was to establish processes, procedures, and roles to acquire existing data sets. Further, the breakout group provided input related to clarifying the types of data owners and the current processes for acquisition. Finally, the participants made suggestions for new processes, procedures, and roles to acquire existing data sets.

The group identified at least nine types of data owners that may have different legal processes in place for conducting data transactions. Participants discussed the current data acquisition approaches for each type of data owner, with specific insights related to different data types and varying legal statuses (e.g., classified data, intellectual property, trade secrets). After establishing a baseline of current practices, the group made suggestions related to adapted acquisition approaches.

The primary purpose of this session was to discuss methods for the beneficial acquisition of existing bioenergy data sets. Participant contributions suggested that a current gap in the industry could be filled by developing a platform that hosts a central database of available data sets and facilitates data transactions. This role might involve vetting data quality and providing mechanisms that support license agreements between data suppliers and potential data users. Further, a data exchange platform host would also be responsible for defining the rules of engagement, monitoring activity, stimulating activity, and providing certain value-added services such as connecting suppliers with potential users.

Data Monetization and Valuation

Overview

The Data Valuation session was designed to gather contributions from workshop attendees related to identifying relevant factors and potential approaches to assigning monetary value to stranded data sets. Participants were asked to determine the relative importance of the variables involved with establishing the value of a given data set, propose methods for quantifying each variable, and propose strategies to validate the value of data.

Workshop organizers asked breakout session participants to assume the perspective of a data purchaser. This hypothetical data purchaser was calculating a value to offer the data owner (possibly at the beginning of a negotiation process). In this scenario, the data quality/usability had already been confirmed, meaning that the hypothetical data may have some gaps but is trustworthy overall (as developed in the Data Quality breakout session). Further, this scenario assumed that there is a willing seller and at least one user identified, that mechanisms for acquisition are in place, and that the data can be legally acquired (as developed in the Data Acquisition breakout session). The overall focus of the session was establishing an approach for assigning monetary value to the data.

Discussion participants included national laboratory and academic scientists, lawyers, data publishers and brokers, and government representatives that oversee research programs. Attendees with significant expertise in assigning value to data sets included a patent attorney who works on merger and acquisition deals as well as a scientist who helped prepare data rooms with valuations and monetary proposals for future funding of a biorefinery company.

Fit for Purpose and Use

One of the suggestions was that the starting point for establishing the value of any given data set is determining its “fit for purpose.” This concept can be described as the set of functionalities offered by a product or service to meet a particular need. In other words, this measure represents functional quality. Similarly, “fit for use” is a term used by vendor making a guarantee that a product or service will meet its agreed-upon requirements. This can relate to nonfunctional requirements of availability, capacity, continuity, and security. This measure represents

nonfunctional quality.[2] The "fit for purpose" concept was also presented in responses to the RFI, where it was noted that many factors relate to data's value, and that the value of data depends strongly on context and the availability of metadata.

Data Set Valuation Examples

During the data valuation breakout session, participants were asked to share information about transaction prices assigned to bioenergy databases in the past. Importantly, participants noted that data are often only part of a transaction. Nonetheless, the few relevant examples provided are listed here:

- $2,000 for supplier market research
- Approximately $50 million for two ethanol plants (data part of transaction)
- $20 million for nonoperating ethanol plant (data part of transaction)
- Possible Shell/CRI sale of GTI-developed IH² process.

Data Valuation Factors

The group identified the following factors that are important considerations during the data valuation process:

- Time and cost to reproduce (sets the price ceiling)
- Data quality
- Metadata and documentation quality
- Scope and urgency of needs addressed by data
- Data uniqueness/novelty
- Number of users (potential and/or committed)
- Degree of data processing
- Age of the data
- Type of data
- Amount of data
- Data producer reputation
- Ease of use
- Exclusive access
- Rights restrictions (licenses, patents)

[2] van Bon, Jan (Editor). 2007. *Foundations of ITIL Volume 3*. Zaltbommel, Netherlands: Van Haren Publishing.

- IP status (pending patents, quality of patents, continuation applications)
- Total funding expended to produce original data.

Respondents to the 2019 RFI indicated that a mix of strategies would be ideal for valuing data.

Data Valuation Strategies

Breakout session participants had a few ideas about how to strategically determine the value of a data set. One suggestion was to utilize independent assessments, or a review board that could incorporate broad industry knowledge. Alternatively, the data provider could conduct market research to find a comparable successful transaction, conduct or use existing market research analysis, and/or solicit hypothetical offers from another potential buyer (demand-side pricing). The group considered that the cost of generating the data may be a useful starting point to set the upper limit (supply-side pricing). Other data points to consider when valuing a data set might include:

- Information about existing alternatives
- How close to or far from commercial viability the data are
- Industrial and process safety (indicative of whether the process could be safely scaled)
- Progress of competitors
- Similarity to or difference from information already available through biomass properties databases.

The group also brainstormed resources that could be helpful for establishing the value of a data set. Their suggestions included the book *Infonomics* by one of the workshop plenary speakers, Doug Laney, as well as utilizing process design software to develop a process for defining data value. The group also suggested using value of information analysis to estimate quantitative values.

An additional question raised by attendees is how data could be valued prior to being contributed as cost-share for a government funding opportunity.

Data Monetization and Valuation Conclusions

The focus of this discussion was in establishing processes for assigning the monetary value of existing bioenergy data. The breakout session discussion assumed that data quality, usefulness, users, and willing sellers are already established. As such, the hypothetical scenario envisioned involved discussions between data suppliers, buyers, and/or brokers with the objective of establishing a suitable purchase price for a given data set.

A few of the group participants had significant experience with data valorization and shared examples of how purchase prices have been set for bioenergy data sets in the past. They explained that in many cases, data have been just one part of a larger transaction that may also

include equipment or other intellectual property. The also noted that there is precedent for potential data users spending money to conduct market research regarding potential data suppliers.

The group compiled a thorough list of the factors that may affect data value, including the time and cost to reproduce (which would set the maximum reasonable purchase price). Additionally, they suggested that there are several different strategies for validating data value, including independent assessments, bidding among buyers, referencing comparable valuations, and estimating the cost to reproduce the data.

Existing Bioenergy Data

The second set of breakout sessions focused on four BETO technology areas: feedstock handling and biorefineries, thermochemical conversion, microorganisms in biotechnology, and algae. For each of these topics, breakout session groups worked to answer the following questions:

- What existing bioenergy data are not currently available to be leveraged for research and development?
- What would be the benefit from making these existing data available?
- How can the data be made available?

For each of the data sets identified, the groups were asked to discuss the potential impact of the existing but unavailable data sets that were identified, and then the likelihood of successfully acquiring the most impactful data sets. Based on these contributions, the groups determined their top-priority data sets to be pursued using an impact and likelihood analysis. For that subset of priority data sets, participants provided ideas related to next steps for data set acquisition. These findings are outlined in the following sections. Although the breakout session attendees discussed some data sets from specific companies, for the purposes of this report, the data sets are discussed in general terms.

Prior to the workshop, registrants of each breakout session were asked to input bioenergy data sets to discuss. Another question asked prior to the workshop to generally gage whether bioenergy stakeholders think there is substantial existing data was: How many data owners do you estimate there are that have potentially acquirable bioenergy data sets? Thirty-six participants responded to this question, with the ability to select more than one answer (Table 3).

Table 3. Participant Estimates of Sources for Potentially Acquirable Bioenergy Data Sets

Number of data owners	Total votes
0	2
25 or less	10
25–50	7

50–75	5
100 or more	11

Feedstock Handling and Biorefineries

Over a dozen representatives from government agencies, national laboratories, universities, and industry participated in a group discussion to identify potential existing data sets related to feedstock handling and biorefineries. Participants contributed insights based on their experience with data transactions from the roles of data user/requestor, data owner/supplier, and general data dissemination/publishing.

The group first ranked the impact of potentially existing data sets that were compiled from RFI input, pre-question input, and additions during the beginning of the breakout session (Table 4). The top-ranked data sets were carried to the next step, which involved assigning scores for impact and likelihood of being acquired. The attendees noted that the impact of all data sets considered was high; however, regarding the likelihood of success in acquiring the data sets, attendees indicated an expectation that biorefinery data are less likely to be obtained.

Table 4. Feedstock Handling and Biorefinery Data Sets

Impact Rank	Datasets
1	Large-scale data from biorefineries
2	Data on how feedstock quality (dry matter, lipid content, moisture content, macromolecular profile, etc.) changes between harvest and conversion to an energy product under different actual conditions at commercial scale.
3	Potential data exists for "late" biomass supply chain logistics activities, often referred to as pre-processing, such as final comminution methods (type, throughput, energy, up-time) screening efficacy (type, motion, deck specs, PSD) hoppers and feed flows, drying or wetting process, etc. in analogous industries.
4	Custom harvesting operations and equipment manufacturers have valuable experience harvesting, packaging, and distributing a range of terrestrial biomass materials throughout the US.
5	Potential data exist for early supply chain logistics activities such as biomass collection, aggregation, transportation, and handling in such analogous industries as feed and forage collection and the timber, pulp, and paper production.
6	Corn stover logistics and preprocessing data
7	SunGrant funded research programs around the country
8	Data from DOE BRDi projects at multiple universities
9	Wood handling and preprocessing data
10	Equipment manufacturers have huge data sets on growth and cost of growth over a huge range of ag crops, but data may be protected by farmer-manufacturer agreements

Potential Impact

The many impacts of access to these data include: increasing the industrial relevance of BETO's techno-economic analyses, focusing research on the principal challenges facing commercial biomass utilization, identifying and solving crosscutting problems within the supply chain using the scientific and technical resources of the national laboratory system, and ultimately reducing start-up delays and feedstock-related challenges seen in first-generation cellulosic ethanol plants.

Participants provided their ideas about the most impactful data types, listed below. They noted that current data for most of these operational parameters exist within BETO programs but are limited by the amount and completeness of the data. Also, much of the data have been collected during research trials and therefore may not accurately represent full-scale industrial operations. In some cases, only averages for these parameters are reported; ranges and distributions are needed to evaluate the impact of process upsets and process improvements on the downstream operations and thus enable DOE research to focus on the most impactful improvements to the supply chain.

One participant stated that access to more complete and more regionally diverse data sets will allow us to: (1) find common problems among a range of operations, (2) link operational challenges to more than one cause and focus on general rather than situation-specific solutions, (3) evaluate our operational models against existing commercial-scale operational measurements, and (4) anticipate the impact that external factors such as climate, soil type, and growing conditions will have on biomass, which is critical for understanding how early supply-chain operations—occurring mostly outdoors throughout the year—will impact downstream preprocessing and conversion operations that take place under more controlled environmental conditions.

Potentially Impactful Data Types

- Field efficiency (ha/hr)
- Fuel consumption (kg/hr)
- Operational efficiency (mg/hr)
- Operational windows (hr/d and d/yr)
- Productivity (productive hr/operational hr)
- Mean time between failure for specific equipment
- Downtime after failure/time to resume normal operations
- Maintenance schedule and costs
- Environmental impacts on productivity (e.g., ambient temperature, soil moisture, slope)
- System-level interactions at industrial scales that affect production and system stability

- Biomass impacts on productivity (e.g., biomass moisture, field density [mg/ha], foreign matter [soil] content)
- Storage-related impacts and interactions of storage conditions and delivered biomass conditions/time-related impacts on quality
- Additional "metadata" such as harvest time, location, and local weather conditions.

Next Steps for Priority Data Sets

For the highest-priority data sets, participants were then asked to list the most important data to obtain, the possible data owners, and ideas about data acquisition strategies. The input received is as follows:

- Data on feedstock quality changes between harvest and conversion to an energy product under different actual conditions at commercial scale
    - For data such as dry matter, lipid content, moisture content, and macromolecular profile, first steps would be to identify specific relevant projects and coordinate with possible data owners.
- Custom harvesting operations that have experience harvesting, packaging, and distributing a range of terrestrial biomass materials throughout the United States
    - Coordinate with companies.
- Preprocessing data from biomass supply chain logistics
    - Coordinate with companies in analogous industries (e.g., pulp and paper). This includes data related to activities such as final comminution (type, throughput, energy, uptime), screening efficacy (e.g., type, motion, deck specs, particle size distribution), hoppers and feed flows, and drying or wetting processes.

Thermochemical Conversion

Ten workshop attendees joined the Thermochemical Conversion breakout session to determine potential thermochemical conversion data sets for acquisition. The majority of participants were data owners or suppliers, with individuals experienced with data valuation, data transactions, and data dissemination also represented.

Several participants reiterated the point that a lack of metadata on some abandoned data sets may make them very difficult to use even if they can be obtained. There was strong support amongst the group for some sort of repository of biofuel property data that follows a standard format.

The group first ranked the impact of potentially existing data sets that were compiled from RFI input, pre-question input, and additions during the beginning of the breakout session (Table 5). The top-ranked data sets were carried to the next step, which involved assigning scores for impact and likelihood of being acquired. The breakout session participants discussed that an

ideal scenario would entail compiling widely ranging data sets that cover feedstocks, intermediates, intermediate processes, and final products.

Table 5. Thermochemical Conversion Data Sets

Impact Rank	Dataset
1	Waste plastic and construction/demolition debris datasets, including proximate, elemental, unreacted compounds, and contaminants
2	Data on feedstock upgrade processes/costs to specific molecules/blends/products for input into LCA/TEA models
3	Data repository of intermediates. Time-on-stream data regarding process stability and identification of key upsets (feeding, catalyst deactivation, etc.)
4	Thermochemical catalyst throughput, conversion rates per conditions (Temperature and Pressure), deactivation rates to specific contaminants (arsenic, mercury, lead, silicon, aluminum, etc.) and compounds (tars, dust)
5	Common-format tables for proximate and ultimate analysis of biomass feedstocks across the full range of resources: specific crops, crop residues (including forest), regional municipal solid waste mixtures, regional sluge/biosolid mixtures, etc.
6	Bio-oil or bio-product characterization data as a function of feedstock, process, and analytical methodology
7	Human-machine interface (HMI) time-stamped data of instrumented systems
8	Catalyst characterization data - experimental or modeling
9	Laboratory notebook documentation of continuous and batch experiments
10	Commercial ventures from established companies that have been sold already
11	Available, but not accessible data due to issues with findability or poor metadata

Another comment by participants was that collaborative research programs such as the Consortium for Computational Physics and Chemistry would use external data for their models if there were industry partners willing to collaborate.

Next Steps for Priority Data Sets

Breakout session participants prioritized the existing but unavailable data based on both impact and likelihood. The attendees ranked the impact of the more general data sets higher (e.g., catalyst throughput and conversion rates for various conditions). They also mentioned that there is commercial value to U.S. Environmental Protection Agency and ASTM certification data generated during the fuel certification process.

Participants expected greater likelihood of success in acquiring the data sets that were developed through publicly funded research and noted that BETO-funded national laboratory programs have already started collecting some of these data.

Microorganisms in Biotechnology

Representatives from national laboratories, universities, government agencies, and private industry participated in the Microorganisms in Biotechnology breakout session with the goal of determining the priority data sets for acquisition in this field.

Participants emphasized that negative data are needed, such as final reports of companies who have closed their operations. The group noted that the value of these data would be to prevent future research and development from encountering the same challenges/mistakes. Negative data are also necessary to paint a more realistic picture of the true state of the art; at present, there is a definite bias in publications and patents toward only publishing "positive" results. Attendees also explained that there may be more value in final technical reports from DOE-funded projects than from patents due to increased level of technical detail. They suggested that sharing negative data may be more feasible if they are anonymized and pooled with other data sources. Further, negative data are critical to training artificial intelligence and machine-learning models, as these approaches require immense amounts of data.

Potentially Acquirable Microorganisms in Biotechnology Data Sets

Breakout session participants listed the following data types that they expected to be available:

- Strain isolation/culturing protocols
- Gene insertion/integration site information for various organisms
- Negative data associated with gene expression or promoters
- Fermentation scale-up performance data
- Negative data associated with organism screening studies
- Sterilization protocols/contamination risks
- Operational stability data
- Media used (rich vs. minimal).

Priority Data Sets

Participants in this breakout session commented that data associated with organisms should have a high likelihood of acquisition, but that scale-up data may be difficult to obtain. The group agreed that these data should be pursued anyway because there is a lot of value in avoiding repeating larger experiments.

The group indicated that the most impactful data sets may be related to fermentation scale-up performance data, final technical reports from DOE-funded projects, and negative data from DOE projects. Participants expected that most of the data sets would be fairly likely to acquire, except for operational stability data, negative data associated with organisms, and negative data associated with gene expression.

In a combined assessment of impact and likelihood, the following data sets were established as the group's shared priorities:

- Negative data from DOE projects
- Final technical reports from DOE
- Fermentation scale-up performance data
- Sterilization protocols/contamination risks
- Strain isolation/culturing protocols (particularly for the most common groups of organisms).

Note: final technical reports from all DOE-funded projects are required to be uploaded to DOE's Office of Scientific and Technical Information public website. All data generated, including negative results, should be included in final technical reports.

Algae

Participants joined this breakout session from a range of affiliations, including academic, national laboratory, industry, and nonprofit organizations. In addition to discussing priority acquisition of existing data sets, the group also considered the need for a clearinghouse for multiple types of data/metadata. They mentioned that attempts have been made to do this in the past, but it has proven to be a difficult effort. One participant also made a general suggestion to develop a template for uploading data so that sharing can be a straightforward and relatively quick process.

Potentially Acquirable Algae Data Sets

This breakout group identified several types of data that are currently unavailable to the research community and suggested various potential sources for each type of data. Only four specific data sources were listed (all from algal biofuel companies).

The group identified two primary challenges that affect the likelihood of acquisition: (1) logistics and (2) motivation. During discussions related to acquisition strategies, the group identified general approaches for overcoming these expected challenges. Ideas included "just ask," make sharing simple (e.g., via database), provide funding for a program to acquire data, and offer co-authorship, payment, or services (e.g., free analysis).

For each type of data listed, the group brainstormed possible sources for the data (Table 6).

Table 6. Types and Sources of Potentially Acquirable Algae Data Sets

Types of Data	Sources of Data
Phenotypic (for wildtypes and mutants)	Algal biofuel companies
Growth patterns	Academic researchers
Growth conditions	National labs
Field cultivation	Industrial research labs
Failed experiments	Folded startups
Algal genomic	Unpublished (e.g., theses)
Transcriptomic	PhycoCosm portal to enable users to both use the data for research and bring new data to share with others
Other -omic	

Priority Data Sets

When ranked, the most impactful data sets included phenotypic (e.g., growth, growth conditions), -omic, and "failed experiment" data from larger algae companies.

Participant commentary related to these priorities provided input related to potentially useful data sets as well as the tools that could support their benefit to the algal science community's research and development efforts. The areas of discussion related to the following topic areas:

- Algal genomics, transcriptomics, and other omics data linked to metadata produced by academic or industrial research labs.
    - *Existing data set(s):* Integrated algal multi-omics data equipped with analysis and modeling tools may enable algal biology understanding and develop framework for strain improvement.
    - *Potential impact:* These data would be valuable, especially if they include microbiome/pest identification, linked to algal growth data and other contextual "metadata."
    - *Potential data sources:* National laboratories, companies, and/or research labs.
        - Companies and laboratories will likely need motivation to share already sequenced genomes, transcriptomes, and metagenomes that are otherwise often kept private.
- Data with respect to the growth pattern and growth conditions of various algal strains.
    - *Existing data set(s):* Essential algae growth data points include temperature, humidity, and light intensity conditions during the algal growth and an average growth rate with respect to the growth curve of the algae.

  - *Potential impact:* Having a set of algae growth reference data could speed up the research.
  - *Potential data sources:* Companies or laboratories working on the same research topic. If these data are made available, collaborative research can be enhanced.

- Metadata on cultivation media and conditions.
  - *Existing data set(s):* Cultivation media factors of interest are nutrient sources, ratios, and concentrations. Important conditions include temperature, pH, light, and light type.
  - *Potential impact:* No specific input provided in this session.
  - *Potential data sources:* These data should be accessible by parsing existing databases (e.g., MycoCosm, KBase Predictive Biology, PhycoCosm, Silva).
- Large algae companies' data sets from research and development to field cultivation.
  - *Existing data set(s):* Production data and process information could be beneficial to researchers and existing producers.
    - Production data: Time series data of cultivated strains with information on pond management.
    - Process information: Standard operating procedures for strain improvement and screenings.
  - *Potential impact:* Industry has recently been repeating work that was previously done at Sapphire to establish best management/cultivation practices for ponds. Sharing data would eliminate the need to reproduce/develop new practices and would potentially allow producers to accelerate their time horizons for reaching production/profit goals.
  - *Potential data sources:* Direct from the company and/or funding agency. It might be hard to determine how much is owned by the company and how much by the government.
- Data from "failed" experiments or preliminary work.
  - *Existing data set(s):* Making these data available will require motivating researchers to report failed experiments.
  - *Potential impact:* This could benefit the research community by preventing repeat work or approaches that have previously failed; additionally, could help guide and refine research questions.
  - *Potential data sources:* National laboratories, companies, and/or research labs.
- Data on phenotypes of natural strains and mutants.

- *Existing data set(s):* There is a wide range of existing data from every sector, especially from algae companies that keep a larger portion confidential.
- *Potential impact:* Publishing phenotypic data of these strains and mutants could provide new insights into the physiology and biochemistry of photosynthesis. In addition, learning which approaches were not successful would prevent researchers from repeating these methods in the future. Researchers would benefit from the data, as collaboration can speed up the research and would enable sustainable research in terms of both use of resources and economics.
- *Potential data sources:* Primarily companies, but other researchers as well.

Breakout Session Report Outs

At the beginning of the third day, the rapporteurs for breakout sessions 1a, 1b, and 1c presented the overviews from their sessions. And at the end of the third day, the rapporteurs for breakout sessions 2a, 2b, 2c, and 2d presented the overviews from their sessions. At the conclusion of the 3-day workshop, the remaining participants who were willing to turn on video attempted a virtual group photo (Figure 7).

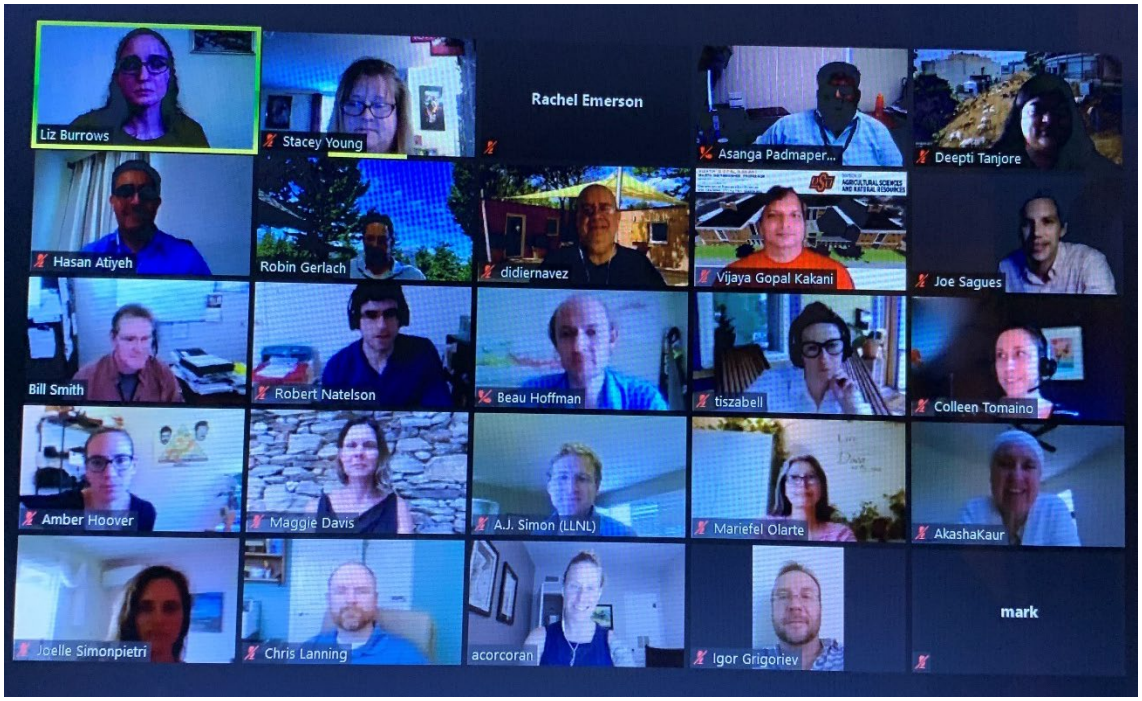


Figure 7. Workshop participant group photo

Summary and Next Steps

Participants at the "Leveraging Existing Bioenergy Data" workshop provided BETO with comprehensive input on all critical aspects needed to successfully acquire high-impact, underused bioenergy data, including:

- Forty-nine existing, potentially acquirable data sets, with highest priority for large-scale industry data
- More than 20 relevant established databases to house the bioenergy data
- Thirty-one data quality metrics with 94 suggested processes for confirming quality/usefulness
- Nine types of data owners and the potential associated legal processes to acquire data from each type of owner
- Multiple ideas for determining the monetary value of data, including independent assessments (e.g., review boards), offers from potential buyers, market analyses, or comparable recent transactions
- Discussion of the benefits of a data exchange platform.

Participants were generally optimistic about the potential to acquire at least certain subsets of existing data that would have a large impact on the field. Ultimately, it was clear that although there is potential for success, each component of a successful data transaction is highly case-specific; the data user would have specific quality needs and the data supplier would have specific conditions that must be met, legal processes that must be followed, and data price ranges that would be acceptable. Workshop participants overall had more bioenergy subject matter expertise, including insight on quantity and quality. In order for the effort to be fully brought to fruition, implementers would need to seek more in-depth legal and valuation expertise.

Although virtual, many connections were made at the workshop, and various researchers have come together on a variety of projects, including compiling a journal special issue on failed experiments, collaborating on database logistics, and sharing common experiences working for biorefineries.

In response to the promising results of this workshop, BETO has funded a small project to perform an initial 1-year proof of concept on the ability to acquire existing data. BETO ran a lab call to develop and implement a process to collect and valorize underused data sets and associated knowledge and make this information available on existing public databases. The process aims to establish:

1. A mechanism for users to submit requests for data as well as a mechanism for suppliers to propose valuable data (e.g., methods, operating parameters, innovations, market analyses, resource assessments)

2. A review system to ensure that data are high-quality, high-impact, and industrially relevant
3. A fair strategy for data monetization
4. A method for data suppliers to provide data and be compensated
5. An efficient way for data to be uploaded or linked to existing public databases.

The project selected was proposed by Oak Ridge National Laboratory, titled "Accelerating Bioenergy Technology Advancement Through FAIR Data Delivery," and will rely heavily on the progress made at this workshop. The project created a website, https://fair-bioenergy-data.pages.ornl.gov/, and is widely soliciting data requests and data offers.

Overall, there exists a finite amount of existing bioenergy data that would be useful to the community, and the amount is expected to slowly increase over time as companies pivot, fail, and/or decide to share certain data. At a minimum, for an effort like the one discussed at this workshop to be considered successful, at least one data set should be acquired and made public, at a cost (and speed) lower than it would take to generate the data again. At a maximum, this could become a well-known and permanent effort, creating new opportunities for data users and data suppliers alike at very low cost. The funding could come from future projects, including funding in budgets to purchase existing data, or the funding could potentially come from user fees to access data. In addition, if a price infrastructure is established through this effort, future projects could include data as part of their cost-share requirements. Lastly, the concepts developed at this workshop are generally applicable to any agency or organization that funds research. Learnings from this effort could be used to establish similar efforts across the federal government.

Appendix A: Agenda

Agenda: Leveraging Existing Bioenergy Data Virtual Workshop U.S. Department of Energy (DOE), Bioenergy Technologies Office (BETO) July 21–23, 2020		
Tuesday, July 21, 2020		
Time (EDT)	**Agenda Item**	**Speaker**
12:30 p.m. – 12:45 p.m.	Login and Networking Activities	
12:45 p.m. – 1:00 p.m.	Welcome and Introduction	Liz Burrows, Technology Manager, DOE BETO
1:00 p.m. – 1:15 p.m.	Introduction to Workshop Software	Lauren Illing, Lead Analyst, BCS LLC
1:15 p.m. – 2:45 p.m.	U.S. Department of Energy Legal Perspective	• Julia Moody, Deputy Chief Counsel for Intellectual Property at DOE • Kim Graber, Legal Counsel at DOE
	Perspectives on the Life Cycle of Critical Data Assets in Technology Development and Commercialization	John Ellersick, President of Next Rung Technology
	The National Microbiome Data Collaborative (NMDC): Building a FAIR Data Resource	Kjiersten Fagnan, Chief Informatics Officer and Data Science and Informatics Leader at the DOE Joint Genome Institute
	Stranded Assets: Considerations of Trade Secret Law	Charles Tait Graves, Partner at Wilson Sonsini Goodrich & Rosati
2:45 p.m. – 3:00 p.m.	Break	
3:00 p.m. – 4:00 p.m.	Leveraging the Value of Data Assets in the Bioeconomy Through Better Data Circulation and Monetization	Didier Navez, Vice President of Strategy & Alliances at Dawex
	Infonomics: The New Economics of Information	Doug Laney, Principal Data Strategist at Caserta
	Energy Data Management, Access, and Analysis	Debbie Brodt-Giles, Group Manager-Data, Analytics, Tools, and Applications at National Renewable Energy Laboratory
4:00 p.m. – 5:00 p.m.	**Open Forum Presentations (3x5)** • Intellectual Property: Types, Eligibility, and Protection • Stranded Data from KiOR • Scale-Up Data: A Hidden Asset • Knowledge Representation to Capture Lessons Learned in Bioprocessing	 • Charles Naggar, Alston & Bird LLP • Bruce Adkins, Oak Ridge National Laboratory • Joe Sagues, North Carolina State University • Deepti Tanjore, Lawrence Berkeley National Laboratory

	• Data Qualification Framework • Multi-omics Data for Fungi and Algae • Computational Catalyst Property Database and Catalyst Deactivation • Time and the Value of Data • Generating and Transferring Technology to Fill Knowledge Gaps	• Rachel Emerson, Idaho National Laboratory • Igor Grigoriev, DOE Joint Genome Institute • Carrie Farberow, National Renewable Energy Laboratory • Bruce Wilson, Oak Ridge National Laboratory • Vijaya Gopal Kakani, Oklahoma State University
Wednesday, July 22, 2020		
Time (EDT)	**Agenda Item**	**Speaker**
10:00 a.m. – 12:00 p.m.	Breakout Session 1a: Data Quality	Moderator: Liz Burrows, BETO
12:00 p.m. – 12:30 p.m.	Break	
12:30 p.m. – 2:30 p.m.	Breakout Session 1b: Data Acquisition	Moderator: Beau Hoffman, BETO
2:30 p.m. – 3:00 p.m.	Break	
3:00 p.m. – 5:00 p.m.	Breakout Session 1c: Data Monetization & Valuation	Moderator: Liz Burrows, BETO
Thursday, July 23, 2020		
Time (EDT)	**Agenda Item**	**Speaker**
12:30 p.m. – 12:45 p.m.	Login and Networking Activities	
12:45 p.m. – 1:30 p.m.	Day 1 Breakout Session Report Outs and Group Discussion	Rapporteurs – one volunteer from each breakout
1:30 p.m. – 3:45 p.m.	Breakout Sessions: • 2a: Feedstock Handling and Biorefineries • 2b: Thermochemical Conversion • 2c: Microorganisms in Biotechnology • 2d: Algae	• Moderator: Mark Shmorhun, BETO • Moderator: Andrea Bailey, BETO • Moderator: Beau Hoffman, BETO • Moderator: Daniel Fishman, BETO
3:45 p.m. – 4:00 p.m.	Break	
4:00 p.m. – 4:45 p.m.	Day 2 Breakout Session Report Outs and Group Discussion	Rapporteurs – one volunteer from each breakout
4:45 p.m. – 5:00 p.m.	Summary Conclusions	BETO

Appendix B: Registrant List

The following table lists registrants and organizers who gave permission to be included in the report.

Last Name	First Name	Job Title	Affiliation
Abdullah	Zia	Biomass Program Manager	National Renewable Energy Laboratory
Acedo	Margarita	Postdoctoral Research Associate	University of Arizona
Adkins	Bruce	Senior Scientist	Oak Ridge National Laboratory
Alexander	Leticia	CEO	z SofTech Solutions
Alward	Gregory	Senior Research Scientist	University of Idaho
Amouri	Mohammed		Centre de Développement des Energies Renouvelables, Algiers
Atiyeh	Hasan	Professor	Oklahoma State University
Bailey	Andrea	Technology Manager	Department of Energy Bioenergy Technologies Office
Barré	Michael	Industrial Technology Advisor	National Research Council of Canada Industrial Research Assistance Program (NRC IRAP)
Bason	Roger	CEO	Atlantic Ocean Aquaculture
Bell	Tisza	Postdoctoral Research Fellow	University of Montana
Bhayani	Bhavin	Consultant	Avatar Sustainable Technology
Brodt-Giles	Debbie	Group Manager - Data, Analytics, Tools, & Applications	National Renewable Energy Laboratory
Burkitt	Adam	Co-Founder	International Waste Petroleum, Inc.
Burli	Pralhad	Economist	Idaho National Laboratory
Burrows	Elizabeth	Technology Manager	U.S. Department of Energy - Bioenergy Technologies Office
Cacho	Jules	Postdoctoral Appointee	Argonne National Laboratory
Cassidy	Chris	National Renewable Energy Coordinator	U.S. Department of Agriculture
Castillo	Krystel	Associate Professor and Director	The University of Texas at San Antonio
Chang	Jeffrey	Research Assistant	University of Delaware
Christensen	Earl	Analytical Chemist	National Renewable Energy Laboratory

Last Name	First Name	Job Title	Affiliation
Cogliani	Leland	Senior Principal	Lewis-Burke Associates
Coleman	Andre	Senior Research Scientist	Pacific Northwest National Laboratory
Collett	Jim	Senior Scientist	Pacific Northwest National Laboratory
Comesana	Ana	Scientific Engineering Associate	Lawrence Berkeley National Laboratory
Contreras	Zariah		None
Corcoran	Alina	Research Scientist	New Mexico Consortium
Crowley	Michael	Center Director/Principal Scientist	National Renewable Energy Laboratory
Davis	Travis	Owner	MalMar LLC.
Davis	Maggie	Natural Resource Data Scientist	Oak Ridge National Laboratory
Dees	John	Ph.D. Student Researcher	University of California, Berkeley
Demirel	Yasar	Professor	University of Nebraska–Lincoln
Dhanasekar	Ashwin	Research Manager	The Water Research Foundation
Dollinger	Caroline	Senior Engineer	Energetics
Donohoe	Bryon	Senior Scientist	National Renewable Energy Laboratory
Dorgan	John	Professor	Michigan State University
Dou	Chang	Scientific Engineering Associate	Lawrence Berkeley National Laboratory
Drennan	Corinne	Sector Lead, Bioenergy Technologies	Pacific Northwest National Laboratory
Dunham	Barbara	Public Educator	Bessemer City Schools
Duoss	Eric	Group Leader	Lawrence Livermore National Laboratory
Dupuis	Eric	CEO	Artona Digital Group Technology
Ebers	Anna	Economist	Tetra Tech
Eldredge	Zachary	Technology Manager	Solar Energy Technologies Office
Ellersick	John	President	Next Rung Technology
Emerson	Rachel	Research Scientist	Idaho National Laboratory
Epperson	Christopher	Technology Consultant	Self Employed
Fagnan	Kjiersten	CIO	U.S. Department of Energy Joint Genome Institute

Last Name	First Name	Job Title	Affiliation
Farberow	Carrie	Group Manager, Researcher	National Renewable Energy Lab
Feng	Maoqi	CEO	Polykala Technologies LLC
Fisher	Aaron	Technology and Innovation Manager	Water Research Foundation
Flowers	Daniel	Program Leader	Lawrence Livermore National Laboratory
Fortier	Marie-Odile	Assistant Professor	University of California, Merced
Gardner	James	Program Manager	Lawrence Berkeley National Laboratory
Geddes	Kristen	Graduate Research Assistant	University of Idaho
Gerlach	Robin	Professor of Chemical and Biological Engineering	Montana State University
Ghasemi	Shokoofeh	PhD Student	North Dakota State University
Glass	David	President	D. Glass Associates, Inc.
Goelz	Ellen	Student	College of Dupage
Graber	Kimberly	Legal Counsel	U.S. Department of Energy
Graves	Charles	Partner	Wilson Sonsini
Grigoriev	Igor	Fungal Genomics Program Lead	U.S. Department of Energy Joint Genome Institute
Gupta	Rishabh	DNA	Forensic Expert
Gussenhoven	Eugene	Director Utilities and Engineering Services	University of Idaho
Hartley	Damon	Computational Scientist	Idaho National Laboratory
Hawkins	Troy	Senior Analyst and Group Leader	Argonne National Laboratory
Hershey	Robert	Consultant	Robert L. Hershey, P.E.
Hewett	Ali	Senior Analyst	BCS, LLC.
Hoffman	Beau	Technology Manager	U.S. Department of Energy - Bioenergy Technologies Office
Hoover	Amber	Research Scientist	Idaho National Laboratory
Hwang	Herng	Manager	Songya Technology LLC
Illing	Lauren	Lead Analyst	BCS, LLC.
Imerman	Mark	President and Senior Consultant	Regional Strategic, Ltd.

Last Name	First Name	Job Title	Affiliation
Jacobson	Oslo	Process Engineer	Lawrence Berkeley National Lab
Jensen	Rasmus	Biofoundry Manager	LanzaTech
Jha	Kshitij C	Founder	Biena Tech
Jones	Daniela	Research Assistant Professor	North Carolina State University/Idaho National Laboratory
Kabeer	Ahammed	Director	Alenso Energy
Kakani	Vijaya Gopal	Professor	Oklahoma State University
Kelleher	Tom	CEO	Xylome Corporation
Khalsa	Akasha Kaur	Associate Energy Specialist	California Energy Commission
Kyeounghwan	Kim	President	SELNIX
Ladisch	Michael	Distinguished Professor and Director	Purdue University
Laney	Douglas	Principal, Data & Analytics Strategy	Caserta
Langholtz	Matthew	Natural Resource Economist	Oak Ridge National Laboratory
Lanning	Chris	Design Engineer	Forest Concepts, LLC
Lee	Steven	Applied Mathematics Program Manager	U.S. Department of Energy Advanced Scientific Computing Research
Leiby	Paul	Distinguished Research Scientist, Team Ldr	Oak Ridge National Laboratory
Li	Chenlin	Distinguished Research Engineer	Idaho National Laboratory
Lorenzo Llanes	Junior	Ph.D. Student	Pontificia Universidad Catolica de Chile
Maki	Alexander	AAAS Science Policy Fellow	U.S. Department of Energy
Manmadkar	Vinayak	Co-Founder	INVENTCO Inc
Martin	Stanton	Data Scientist	Oak Ridge National Laboratory
McGowen	John	Director of Operations and Program Management	Arizona State University: AzCATI
Meadows	Jamie	AAAS Fellow-Bioenergy Technologies Office	U.S. Department of Energy-Bioenergy Technologies Office
Moody	Julia	Deputy Chief Counsel for Intellectual Property at DOE's Golden Field Office	U.S. Department of Energy

Last Name	First Name	Job Title	Affiliation
Mosheim	John	Engineer	GHG Engineering, LLC
Mueller	Evan	Engineer	BGS
Naggar	Charles	Associate	Alston and Bird LLP
Narani	Akash	Sr. Process Engineer	Lawrence Berkeley National Laboratory
Naranjo	Robert	Senior Vice President	BCS, LLC.
Natelson	Robert	Environmental Engineer	Allegheny Science and Technology
Navez	Didier	SVP Strategy & Alliances	DAWEX
Nguyen	Quang	Researcher	Idaho National Laboratory
Padmaperuma	Asanga	Laboratory Relationship Manager	Pacific Northwest National Laboratory
Pandey	Ramsharan	Graduate research assistant	North Dakota State University
Park	Sunkyu	Associate Professor	North Carolina State University
Pathak	Harsh	Graduate Research Assistant	North Dakota State University
Pavan	Marilene	Scientist	LanzaTech
Rapp	Vi	Research Scientist	Lawrence Berkeley National Laboratory
Rashel	Rakib	Postdoctoral Research Scientist	New Mexico Consortium
Rawson	Linda	President	DynaGrace Enterprises, Inc.
Reimer	Don	President	D.R. Systems Group
Resch	Michael	Bioconversion Specialist	National Renewable Energy Laboratory
Rials	Tim	Associate Dean and Director	University of Tennessee
Ripplinger	David	Associate Professor	North Dakota State University
Rohman	Clayton	Senior Project Engineer	BGS
Sagues	William (Joe)	Assistant Professor, Bio. & Ag. Engineering	North Carolina State University
Sahoo	Kamalakanta	Assistant Research Scientist	Forest Products Laboratory
Sahu	Somesh	Design Engineer	Medors Renewable Energy P Ltd
Sanders	Patricia	Founder	Thirty Seven Broad Company
Schaidle	Joshua	Platform Lead	National Renewable Energy Laboratory

Last Name	First Name	Job Title	Affiliation
Shalaby	Hany	Consultant	SC & ER
Sharma	Pankaj	Managing Director	Purdue University
Shmorhun	Mark	Technology Manager	U.S. Department of Energy
Shrestha	Dev	Professor	University of Idaho
Shukla	Satish	Co-Founder	INVENTCO Inc.
Simon	Ben	Project Monitor/Engineer II	BGS
Simon	A.J.	Group Leader	Lawrence Livermore National Laboratory
Smith	William	Staff Scientist	Idaho National Laboratory
Smith	David	Chief Legal Officer	Homasoft
Smith	Sarah	Sr. Scientific Engineering Associate	Lawrence Berkeley National Laboratory
Smith	Emily	Fellow	Department of Energy
Snowden-Swan	Lesley	Engineer	Pacific Northwest National Laboratory
Staples	Lauren	Consultant	Self
Stright	Dana	Sr. Data Strategist	National Renewable Energy Laboratory/Skye Analytics, Inc.
Sun	Jiayang	Senior Fellow at USDA, Professor/Chair at GMU	U.S. Department of Agriculture/GMU
Swanson	Michael	Principal Engineer	UNDEERC
Szymkowicz	Rebecca	Program Analyst	Redhorse Corporation
Talmadge	Michael	Sr. Research Engineer	National Renewable Energy Laboratory
Tanjore	Deepti	Director, ABPDU	Lawrence Berkeley National Laboratory
Tao	Ling	Sr. Engineer	National Renewable Energy Laboratory
Teymouri	Farzaneh	Research assistant professor	Michigan State university
Theiss	Timothy	Group Leader, Renewable Energy Systems	Oak Ridge National Laboratory
Tomaino	Colleen	Vice President, Clean Energy	BCS, LLC.
Tourigny	Guy	Research Engineer	CanmetNENERGY
Ussery	John	Program Director	Northern New Mexico College

Last Name	First Name	Job Title	Affiliation
van Opstal	Edward	AAAS Science & Technology Policy Fellow	U.S. Department of Defense
Vanzin	Gary	Research Faculty	Colorado School of Mines
Viswanathan	Mothi	Postdoctoral research associate	University of Illinois Urbana Champaign
Volk	Timothy	Senior Research Associate	SUNY ESF
Wang	Michael	Center director	Argonne National Laboratory
Wang	Michael	Manager	Argonne National Laboratory
Ward	Christopher	Assistant Professor	Bowling Green State University
Warren	Quinta	CEO	Energy Research Consulting
Wilson	Bruce	Group Leader	Oak Ridge National Laboratory
Wong	Jennifer	Analyst	National Institutes of Health
Wu	May	Principal Environmental System Scientist	Argonne National Laboratory
Wu	Shenghua	Assistant Professor	University of South Alabama
Wu	Sarah	Assistant Professor	University of Idaho
Xu	Hui	Environmental Analyst	Argonne National Laboratory
Yan	Jipeng	Senior process engineer	Lawrence Berkeley National Laboratory
Yang	Bin	Associate Professor	Washington State University
Yang	Minliang	Postdoctoral Scholar	Lawrence Berkeley National Laboratory
Young	Stacey	Senior Conference Manager	The Building People
Yu	Fei	Professor	Mississippi State University
Zhong	Jia	Graduate Assistant	University of Illinois Urbana Champaign

U.S. DEPARTMENT OF ENERGY | *Office of* ENERGY EFFICIENCY & RENEWABLE ENERGY

BIOENERGY TECHNOLOGIES OFFICE

For more information, visit: energy.gov/eere/bioenergy

DOE/EE–2330 · March 2021

Cover photos from iStock 490265597, 1068391222, and 1177522959.