OAK RIDGE National Laboratory

DOE Bioenergy Technologies Office (BETO) 2021 Project Peer Review

Accelerating Bioenergy Technology Advancement Through FAIR Data Delivery

March 12, 2021
Data Modeling & Analysis Session

Bruce E. Wilson
Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Background: Leveraging Existing Bioenergy Data

While this is a new project, it builds on BETO work with the community to assess and refine our understanding of the need.

- BETO 2019 Request for Information (RFI)
- BETO Workshop July 21-23, 2020 (COVID delay from spring 2020)
  - https://www.energy.gov/eere/bioenergy/events/leveraging-existing-bioenergy-data-workshop
  - Wide range of participants:
    - Data producers, data valuation, data publication/preservation, data users
    - Legal, industry, lab, academia
- Draft report from RFI and Workshop shared with project team
- Identified data and sources of potentially broad interest
  - Process scale-up generally seen as the most valuable
  - Data from "failed" experiments could be useful, especially in aggregate
  - Some leads are quite specific, others more general

Background: FAIR data

https://go-fair.org
https://doi.org/10.1038/sdata.2016.18

FAIR is a useful framework for assessing the (re)usability of data and information.

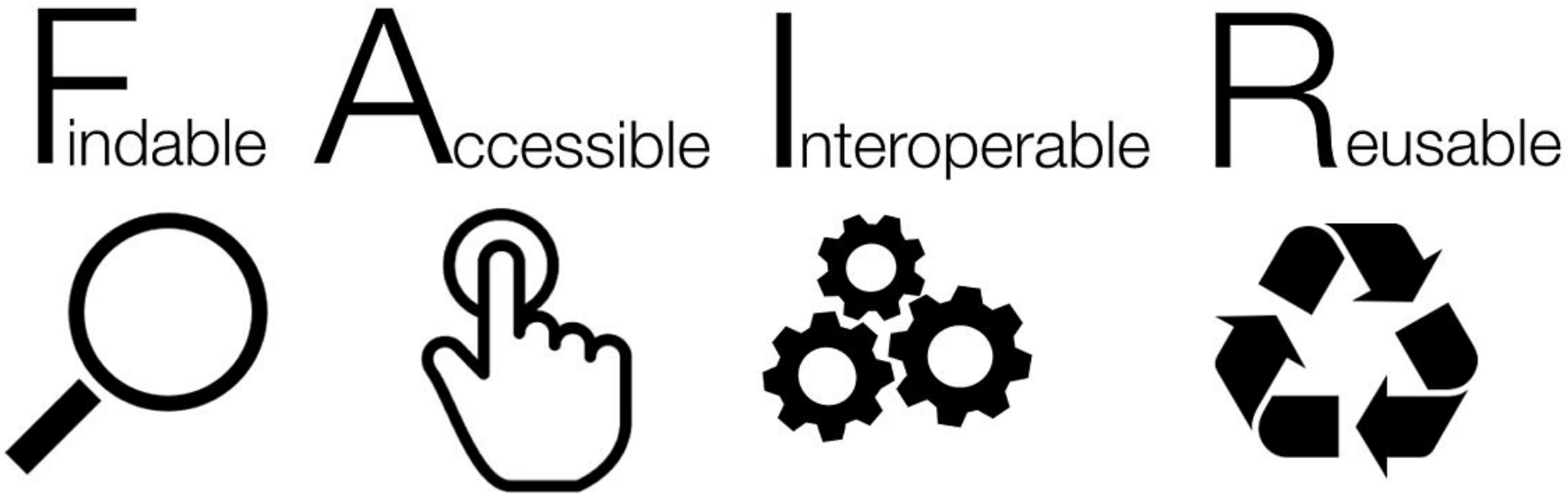


Image via Wikimedia, CC-BY 3.0

Deficiencies in any of these can lead to data cemeteries

Not all data warrants the same degree of investment in FAIRness

Project Overview

- **Goal:** Demonstrate the value of improving the FAIRness of ***selected*** bioenergy technology data for advancing the bioenergy economy
- New project, started in mid-November 2020
- **Underlying hypothesis:** Data exists which would be valuable to bioenergy technology researchers, but that data is not sufficiently FAIR
  - Failed experiment data often not published or shared
  - Proprietary data may be available for sale or cost share
  - Some data only exists on researchers' hard drives
  - Good metadata is essential for Findability
  - Good documentation is essential for Reusability

1 – Management

The ORNL team, working with BETO, has the necessary domain and data expertise, as well as community connections.

- People:
  - Bruce Wilson (PI): process analytical chemist with experience in cellulose functionalization and data management
    - Current primary role: Manager for NASA archive of record for terrestrial ecology data
    - Previously worked on instrumenting and running pilot-scale equipment
  - Jess Welch: Data scientist and ecologist, with substantial experience analyzing, publishing, and documenting data
  - Rocio Uria Martinez, Paul Leiby, Matt Langholtz: ORNL Environmental Risk and Energy Analysis Group; substantial expertise in resource economic analysis
  - BETO: Input from 2020 data workshop, overall guidance, connections into the community
  - Laboratory Resource Managers: Cross-laboratory connections, information on potential data, and help with valuation assessment

1 – Management

Based on the 2020 workshop, we are focusing on scale-up data. The delay in starting the project has been handled.

- Primarily seeking biomass processing data, particularly scale-up
  - Likely most valuable, based on July 2020 workshop
  - Data and reports from failed/pivoted companies (not Accessible)
  - Data and reports that have no metadata (not Findable, likely not Reusable)
  - Data and reports squirreled away on computers/hard drives (not Accessible)
  - Data potentially available "upon request" (not Findable, or clearly Accessible)
- Timeline and milestones
  - 10/1/20: Project start (delayed until mid-November due to funding)
  - 2/28/21: Identify 3-6 potential datasets and complete initial assessment (was 1/31)
  - 4/30/21: Select 1-2 datasets; start acquisition/improvement work (was 3/31)
  - 7/15/21: Target dataset published; start collecting feedback (was 6/30)
  - 9/30/21: Project complete

1 – Management

Good communication is particularly important for this project. Key risks have been identified and mitigated.

- Communication and Collaboration
  - ORNL-internal Microsoft Teams environment and RESolution Project tracking
  - Annual Operating Plan (AOP) and regular reporting; Regular check-ins with BETO
  - Web site: https://fair-bioenergy-data.pages.ornl.gov/ (see slide 14)
  - Generic email: bioenergy_valorize@ornl.gov
  - Will distribute data through Bioenergy KDF and other channels
- Key Risks
  - Information disclosure/contamination risks if working with proprietary data
    - Mitigation: Follow established proprietary data procedures; firewall project area
  - Assessing value of data/information is a hard problem
    - Mitigation: Extensive ORNL experience; don't let the perfect be the enemy of the good
  - No appropriate data available
    - Mitigation: July 2020 workshop suggest this risk is low; multiple options already identified
  - DOE can't pay twice for rights to data

    Mitigation: Make sure we clearly understand how the data was funded

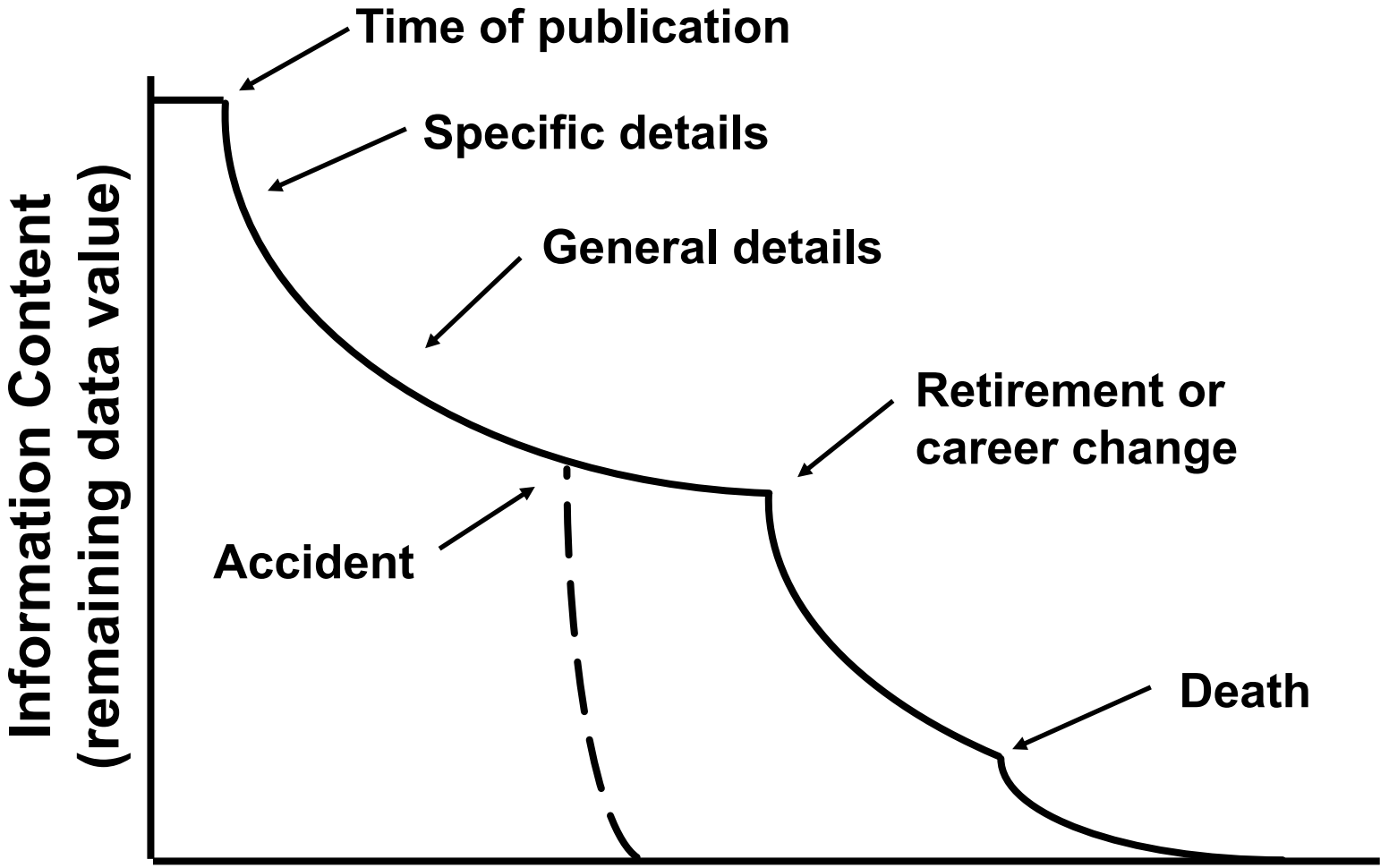
2 – Approach

Our approach uses a well-understood data publication model.

- Identify potential datasets
  - Workshop potential datasets listing
  - Work with LRMs, BETO network, community
  - https://fair-bioenergy-data.pages.ornl.gov/
- Evaluate potential value and expected costs
  - Use experts across the DOE community for qualitative assessment
- Select (go/no go) 1 or 2 best candidates
  - Do we have at least one group that clearly will use the data?
- Leverage ORNL data publication expertise
  - Soon and usable now is more important than long-term reusability
- Promote through multiple community channels, including internet search
- Assess impact and determine next steps

2 – Approach: Elements of assessing value

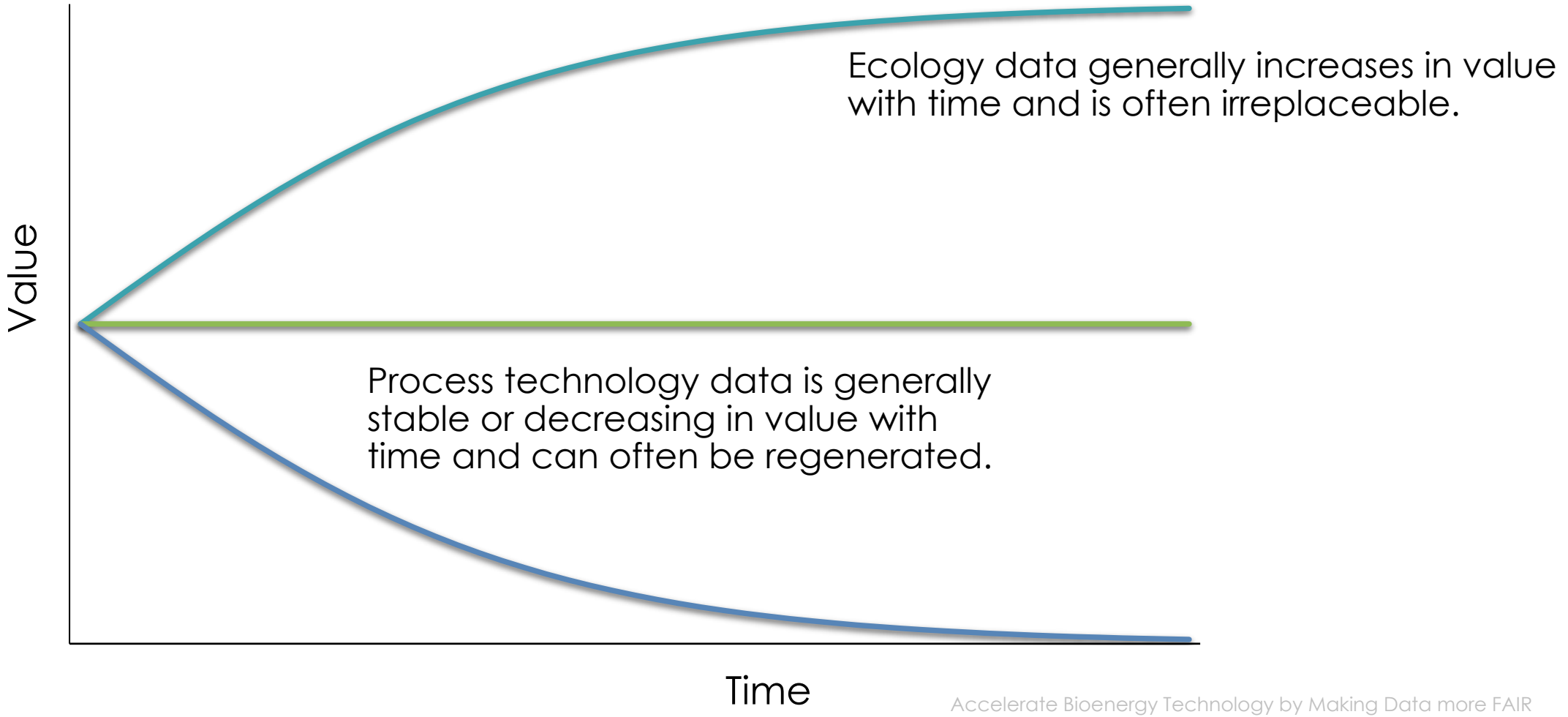
The earlier we can work with a data provider, the more residual value we can maintain from the data.



Michener et al (1997) doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

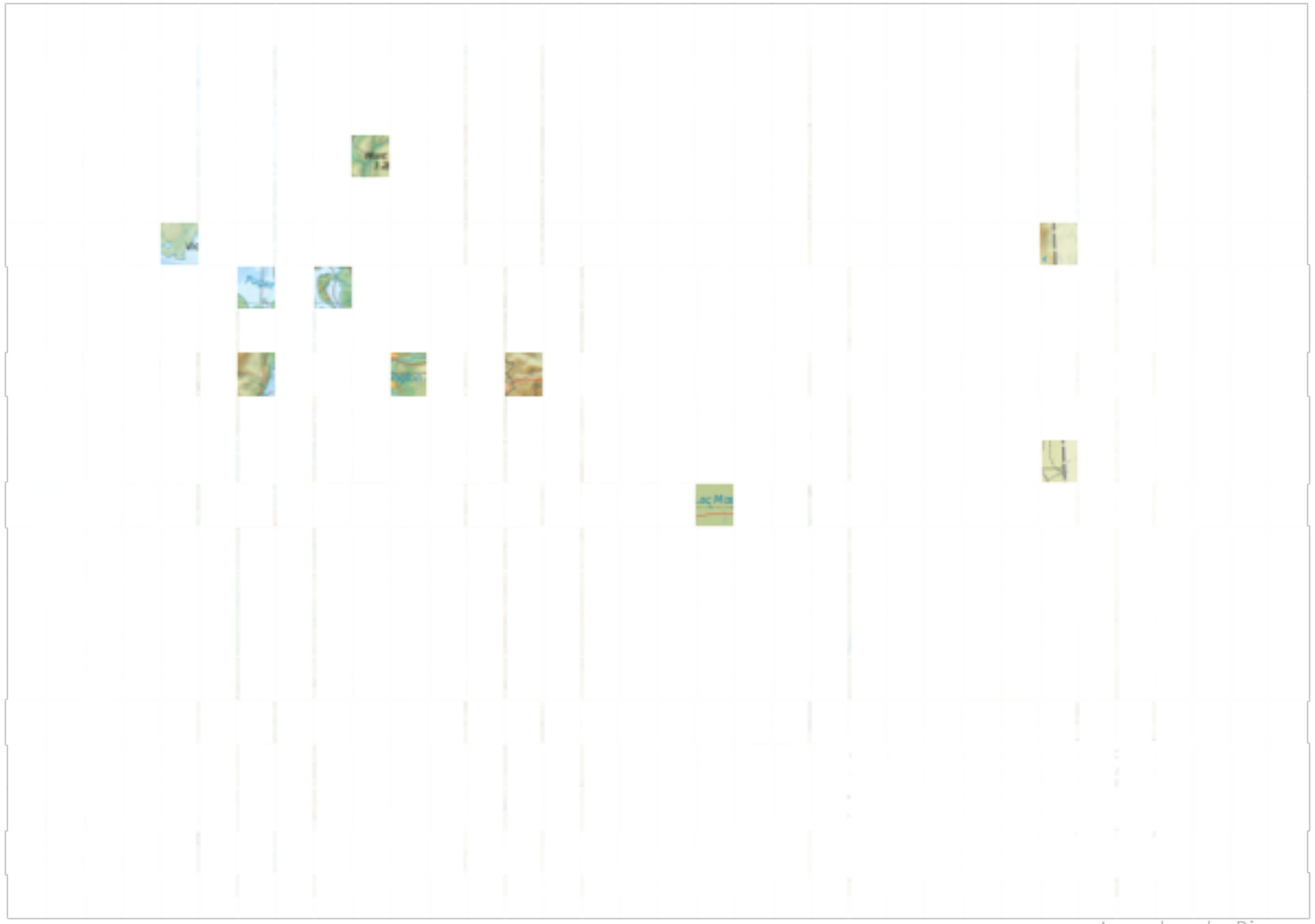
2 – Approach: Elements of assessing value

Understanding data value versus time further supports our focus: relatively recent data.



2 – Approach: Elements of assessing value

Data valuation is context-dependent. With isolated data we can't see much of the big picture.



2 – Approach: Elements of assessing value

Data valuation is context-dependent. With more of the picture, we can be more strategic filling in holes.

The community picture indicates scale-up is the most critical.

3 – Impact

Making data FAIR has clear value for accelerating industry and advancing science.

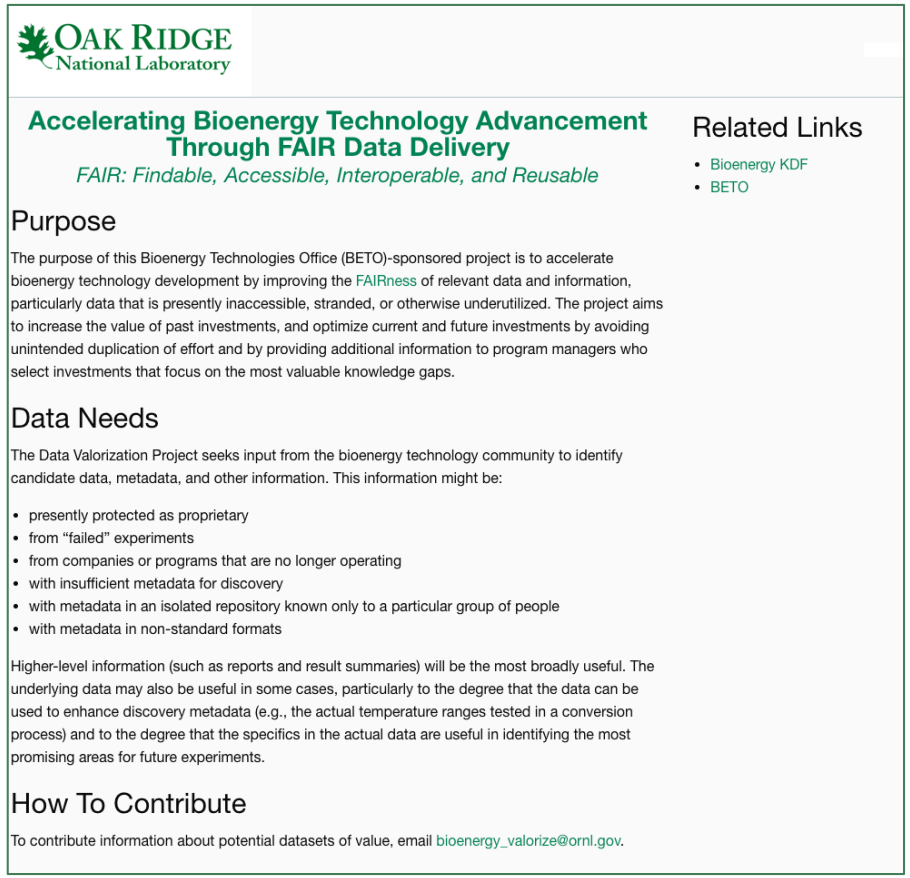
- Avoiding inadvertent duplication of effort improves efficiency
- Data and information to assess why things didn't work is key to finding paths that do work
- There are examples of the buying data to accelerate research
  - Industry does this regularly (via acquisition)
  - Data has been used as in kind/cost share contribution on projects
- There are multiple examples of data rehabilitation creating value
  - Particularly where very old data cannot be reproduced
- The data someone can find & use often depends on their network
- FAIR data doesn't have to be open: Access rules just need to be clear
  - Public discovery metadata is critical for driving value

[1] A European Strategy for Data https://ec.europa.eu/digital-single-market/en/european-strategy-data (accessed 2021-02-13)
[2] M.S. Ries and R. Kennett, AIChE Journal (2018) https://doi.org/10.1002/aic.16203
[3] USGS, Economic Valuation of Landsat Imagry (2019) https://doi.org/10.3133/ofr20191112

4 – Progress and Outcomes

Based on preliminary results, the question is not whether there is valuable data, but rather how much is worth pursuing.

- Started in mid-November
- Basic project website & contact email address
- Multiple promising leads from workshop and community suggestions
- Making data reusable will require some investment, but well within project budget
- A key deliverable will be an assessment of the range of available data and estimated costs



OAK RIDGE National Laboratory

Accelerating Bioenergy Technology Advancement Through FAIR Data Delivery

FAIR: Findable, Accessible, Interoperable, and Reusable

Related Links
- Bioenergy KDF
- BETO

Purpose

The purpose of this Bioenergy Technologies Office (BETO)-sponsored project is to accelerate bioenergy technology development by improving the FAIRness of relevant data and information, particularly data that is presently inaccessible, stranded, or otherwise underutilized. The project aims to increase the value of past investments, and optimize current and future investments by avoiding unintended duplication of effort and by providing additional information to program managers who select investments that focus on the most valuable knowledge gaps.

Data Needs

The Data Valorization Project seeks input from the bioenergy technology community to identify candidate data, metadata, and other information. This information might be:

- presently protected as proprietary
- from "failed" experiments
- from companies or programs that are no longer operating
- with insufficient metadata for discovery
- with metadata in an isolated repository known only to a particular group of people
- with metadata in non-standard formats

Higher-level information (such as reports and result summaries) will be the most broadly useful. The underlying data may also be useful in some cases, particularly to the degree that the data can be used to enhance discovery metadata (e.g., the actual temperature ranges tested in a conversion process) and to the degree that the specifics in the actual data are useful in identifying the most promising areas for future experiments.

How To Contribute

To contribute information about potential datasets of value, email bioenergy_valorize@ornl.gov.

Summary

- While this is a new project, it builds on BETO work with the community to assess and refine our understanding of the need.
- FAIR is a useful framework for assessing the (re)usability of data and information. Making data FAIR has clear value for accelerating industry and advancing science.
- The ORNL team has the necessary domain and data expertise, as well as community connections.
  - Based on the 2020 workshop, we are focusing on scale-up data.
  - Our approach uses a well-understood data publication model and well-understood approaches to data curation.
- Based on preliminary results, the question is not whether there is valuable data, but rather how much is worth pursuing.

Quad Chart Overview

Timeline

- Start: mid-November 2020
- End: September 30, 2021

	FY20	Active Project
DOE Funding	(10/01/2019 – 9/30/2020)	$250K ($50K data buy budget w/overhead)

Performers

- Oak Ridge National Laboratory

Barriers addressed

- Negative results are often not published
- Making data FAIR accelerates science
- Not all data are worth preserving
- Value of technology data often diminishes with time

Project Goal

Accelerate bioenergy technology development by improving the Findability, Accessibility, Interoperability, and Reusability (FAIRness) of relevant data and information that is presently stranded, inaccessible, or otherwise underutilized.

End of Project Milestone

Complete first cycle through the Identify/Evaluate/Select/ Publish/Promote/Reassess process, with Reassess including a retrospective session with BETO to refine procedures for a potential (presently not funded) Year 2.

Funding Mechanism

1-year seed AOP