Tackling the Challenges of Earth Science Data Synthesis: Insights from (meta)data standards approaches

Valerie C Hendrix[1,1], Danielle S Christianson[1,1], Charuleka Varadharajan[1,1], Madison Burrus[1,1], Shreyas Cholia[1,1], You-Wei Cheah[1,1], Housen Chu[1,1], Robert Crystal-Ornelas[1,1], Joan E Damerow[1,1], Zarine Kakalia[1,1], Fianna O'Brien[1,1], Gilberto Pastorello[1,1], Emily Robles[1,1], and Deb Agarwal[1,1]

[1]Lawrence Berkeley National Laboratory

November 30, 2022

Abstract

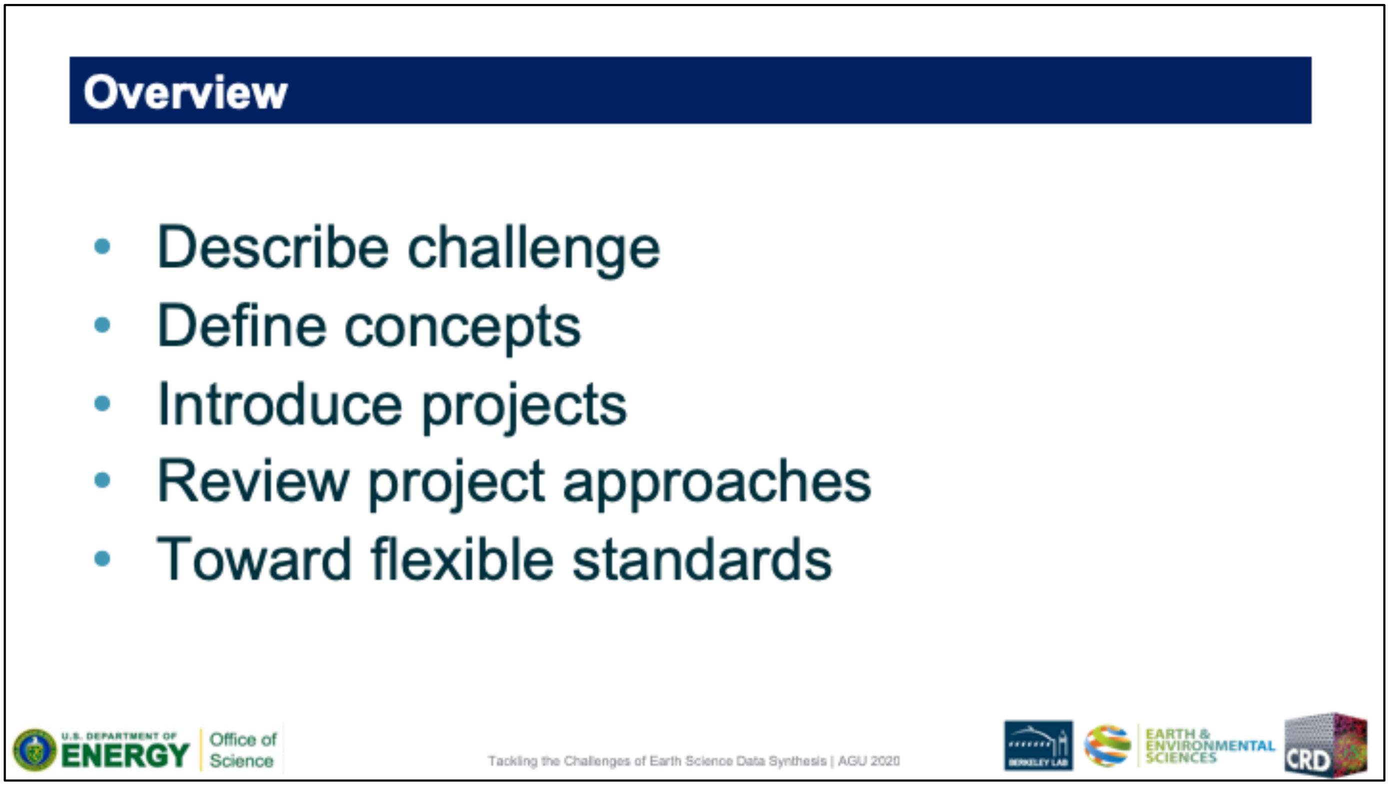
Diverse, complex data are a significant component of Earth Science's "big data" challenge. Some earth science data, like remote sensing observations, are well understood, are uniformly structured, and have well-developed standards that are adopted broadly within the scientific community. Unfortunately, for other types of Earth Science data, like ecological, geochemical and hydrological observations, few standards exist and their adoption is limited. The synthesis challenge is compounded in interdisciplinary projects in which many disciplines, each with their own cultures, must synthesize data to solve cutting edge research questions. Data synthesis for research analysis is a common, resource intensive bottleneck in data management workflows. We have faced this challenge in several U.S. Department of Energy research projects in which data synthesis is essential to addressing the science. These projects include AmeriFlux, Next Generation Ecosystem Experiment (NGEE) - Tropics, Watershed Function Science Focus Area, Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), and a DOE Early Career project using data-driven approaches to predict water quality. In these projects, we have taken a range of approaches to support (meta)data synthesis. At one end of the spectrum, data providers apply well-defined standards or reporting formats before sharing their data, and at the other, data users apply standards after data acquisition. As these projects continue to evolve, we have gained insights from these experiences, including advantages and disadvantages, how project history and resources led to choice of approach, and enabled data harmonization. In this talk, we discuss the pros and cons of the various approaches, and also present flexible applications of standards to support diverse needs when dealing with complex data.

Tackling the Challenges of Earth Science Data Synthesis

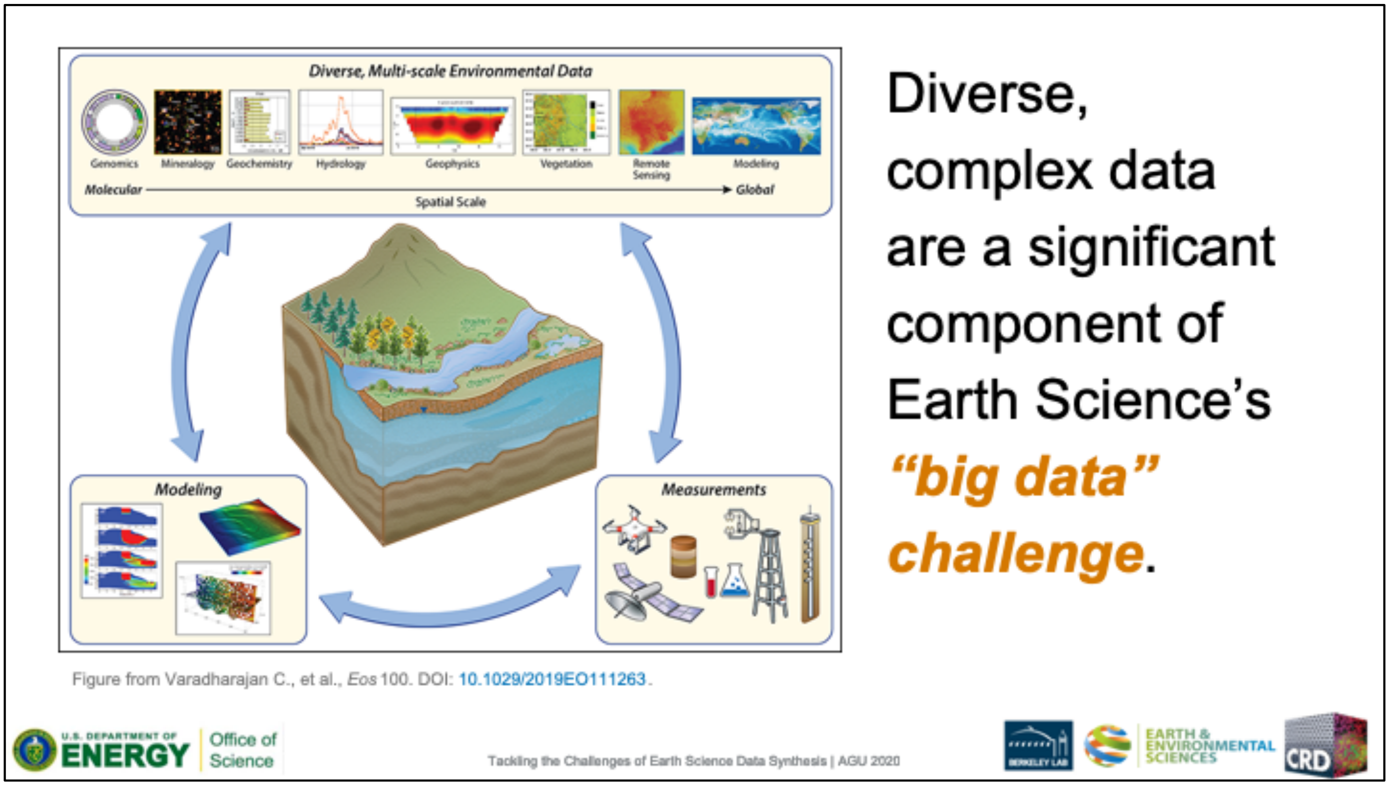
Insights from (meta)data standards approaches

Valerie C Hendrix, *Danielle S Christianson, Charuleka Varadharajan, Madison Burrus, Shreyas Cholia, You-Wei Cheah, Housen Chu, Rob Crystal-Ornelas, Joan E Damerow, Zarine Kakalia, Fianna O'Brien, Gilberto Pastorello, Emily Robles, Deb Agarwal*

My name is Valerie Hendrix. I am a computer systems engineer from Berkeley lab. Today I will be sharing with you how myself and my co-authors tackle the challenges of Earth Science data synthesis by providing the insights we have gained from our different approaches to applying data and metadata standards to earth science data.

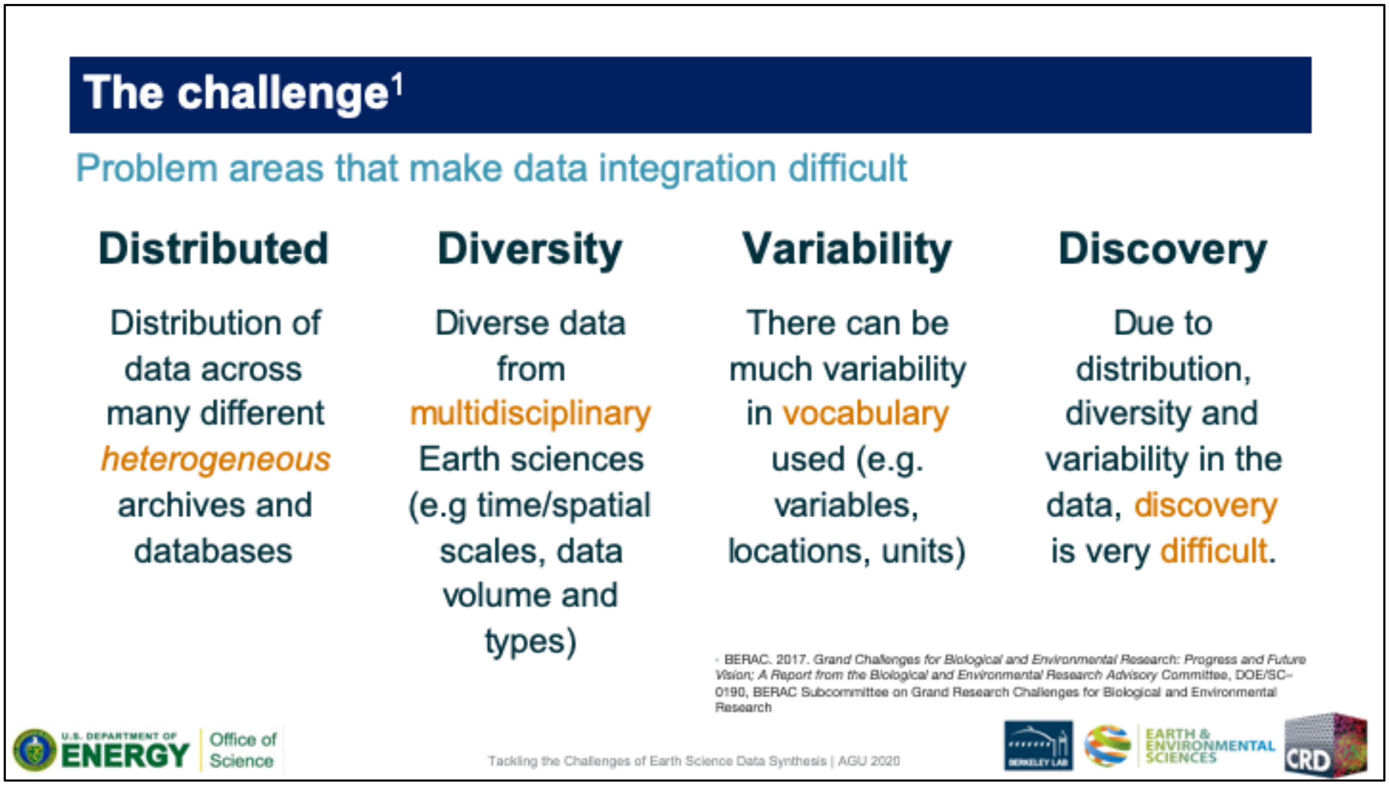


This presentation will first describe the challenge to data synthesis for earth science projects and define some concepts to get us all on the same page. I will then introduce several Department of Energy projects whose approaches to data standards application I will discuss. Finally I will discuss some future work that will build on these project experiences and insights.

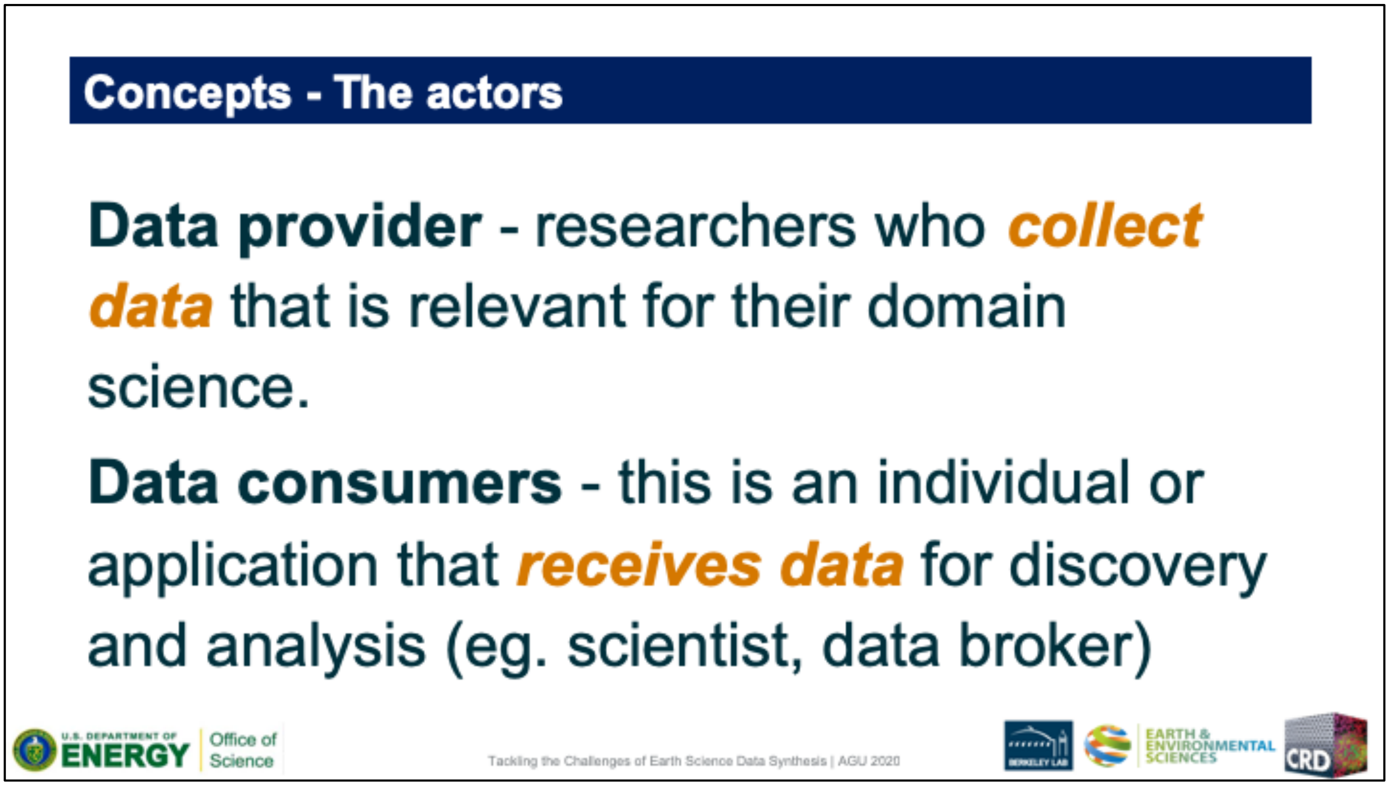


Earth science data sets are generated by an iterative methodology where models guide measurements and vice versa . This approach uses and generates diverse data from multidisciplinary Earth sciences, including hydrology, ecology, climate, geology, geophysics, geochemistry, and microbiology.

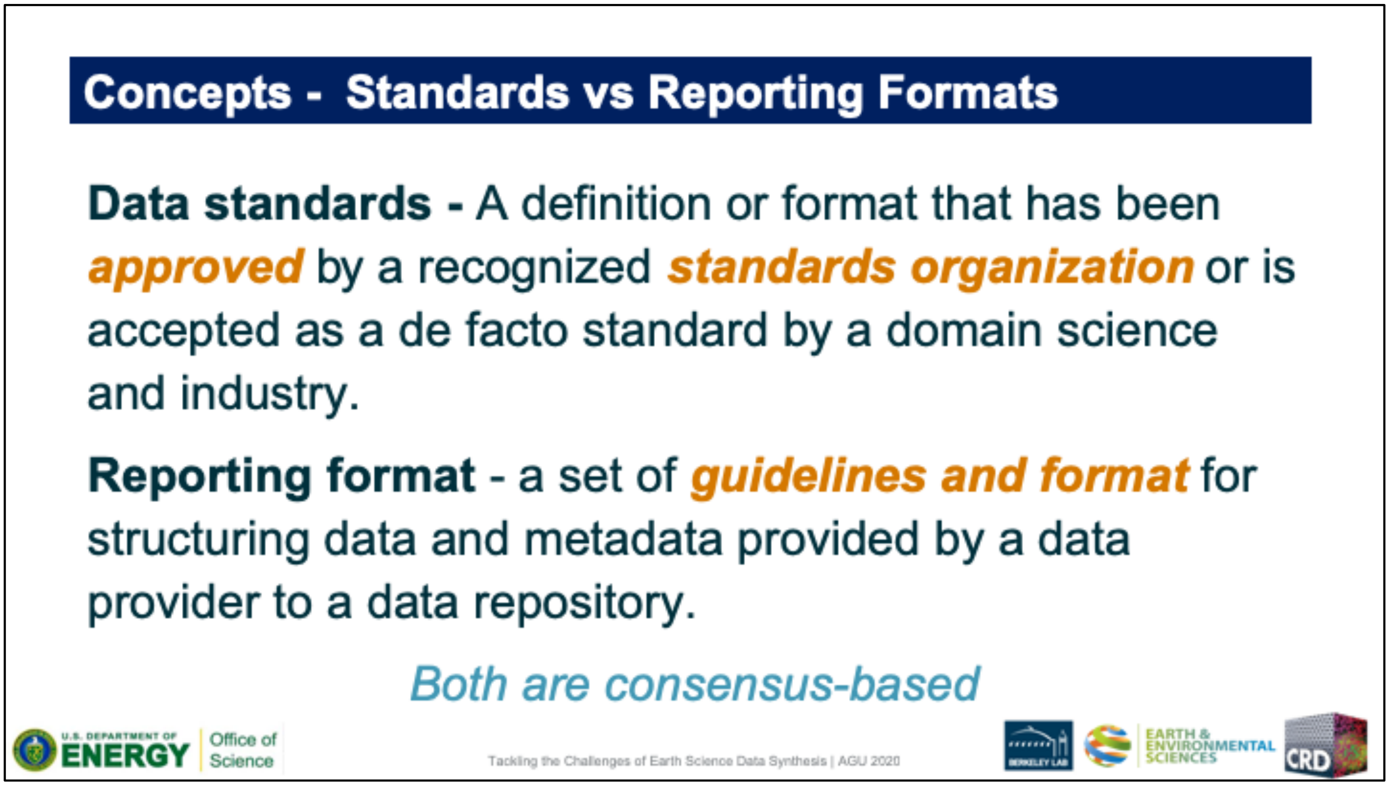
These data are diverse and complex and are a significant component of earth science's big data challenge.



The challenge falls in these problem areas which makes integration of the diverse, complex earth science data difficult thus hindering scientific discovery. Data is distributed across different, heterogeneous archives and databases. Data is diverse, coming from multidisciplinary earth sciences and can be in different scales, volumes and data types. There also is a lot of variability in vocabulary used such as variable names, measurement units used and how locations are described. Finally, due to the distributed, diverse and variable nature of the data discovery of data for scientific analysis is very difficult.



Now that I have discuss earth science data challenges, I want to define some concepts that are important to understand for this talk.
Data provider are researcher who collect data that is relevant for their domain science.
Data consumers can be person or software that receives data for discovery and analysis (eg. scientist, data broker)



When talking about the structure of data we use standards or reporting formats. Both are designed by consensus. A data standard is a definition or format that has been approved by a recognized standards organization can be developed over decades. Standards not only exist for data formats but also for programming languages, operating systems, communications protocols, and electrical interfaces

A reporting format is a set of guidelines and format for structuring data and metadata provided by a data provider to a data repository. These are community-driven data guidelines developed to promote well structured data where no standard is available.

For the purpose of this talk. I will use standards to mean both reporting formats and standards.

Concepts - Metadata and Data

Standards can ***define the structure*** of both data and metadata. (e.g. variable names, units, requirements, file formats, etc)

- ***Metadata*** is information about the data and context of the data.
- ***Data*** are measurements or observations (time-series, images, genomes …)

Standards can be used to define the structure of both data and metadata. Metadata is information about the data and the context of the data. Data are measurements and observation such as time-series, image, genomics ...

My co-authors and I are on the data management teams of the U.S. Department of Energy research projects listed on this slide. We have faced this data challenge in which data synthesis is essential to addressing the science. I will be discussing four approaches to using standards taken by the following 5 projects

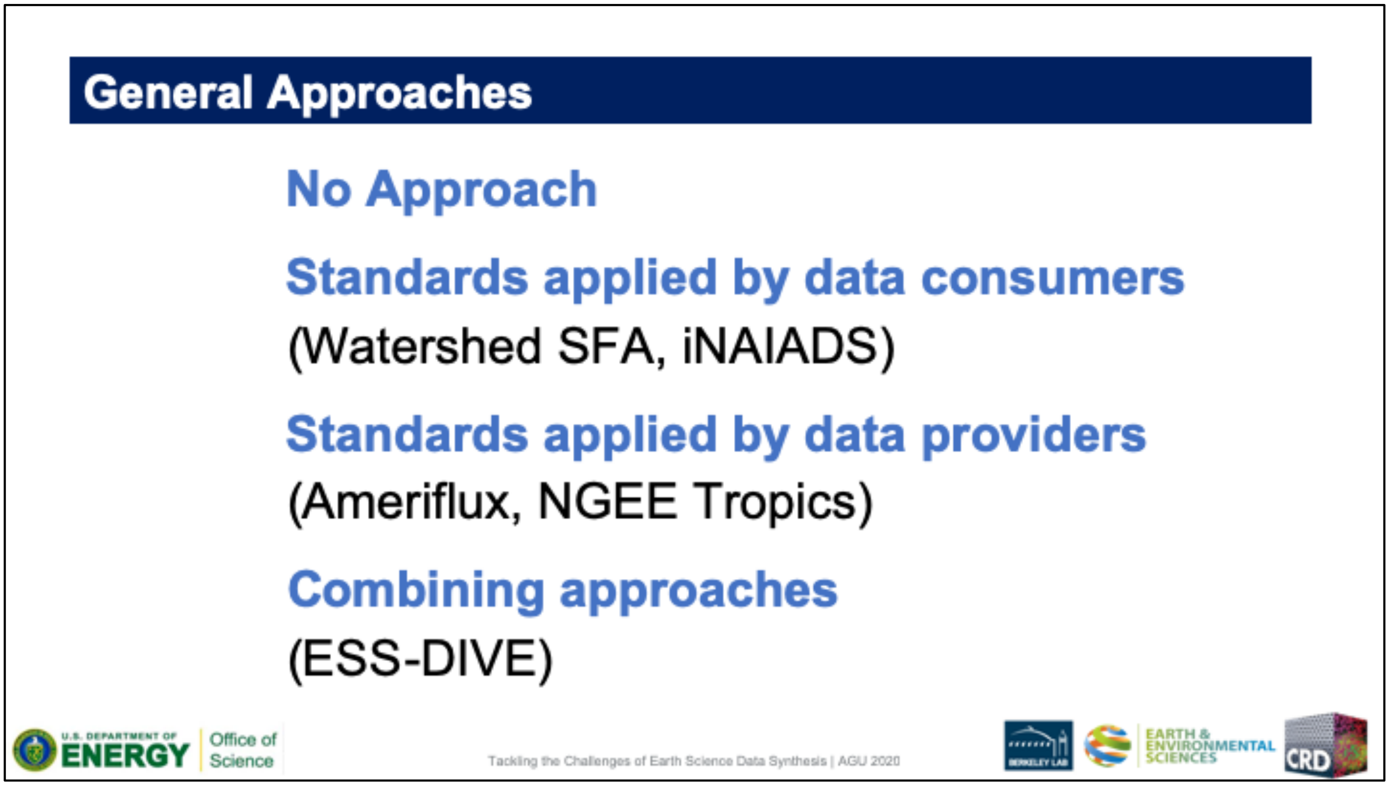
NGEE Tropics: A team science project to project carbon cycle in the tropics under future climate. Models inform observations across multiple tropical sites.

AmeriFlux is a network of independent researchers studying carbon, water, and energy exchange between the atmosphere and land surface in diverse systems across the Americas.

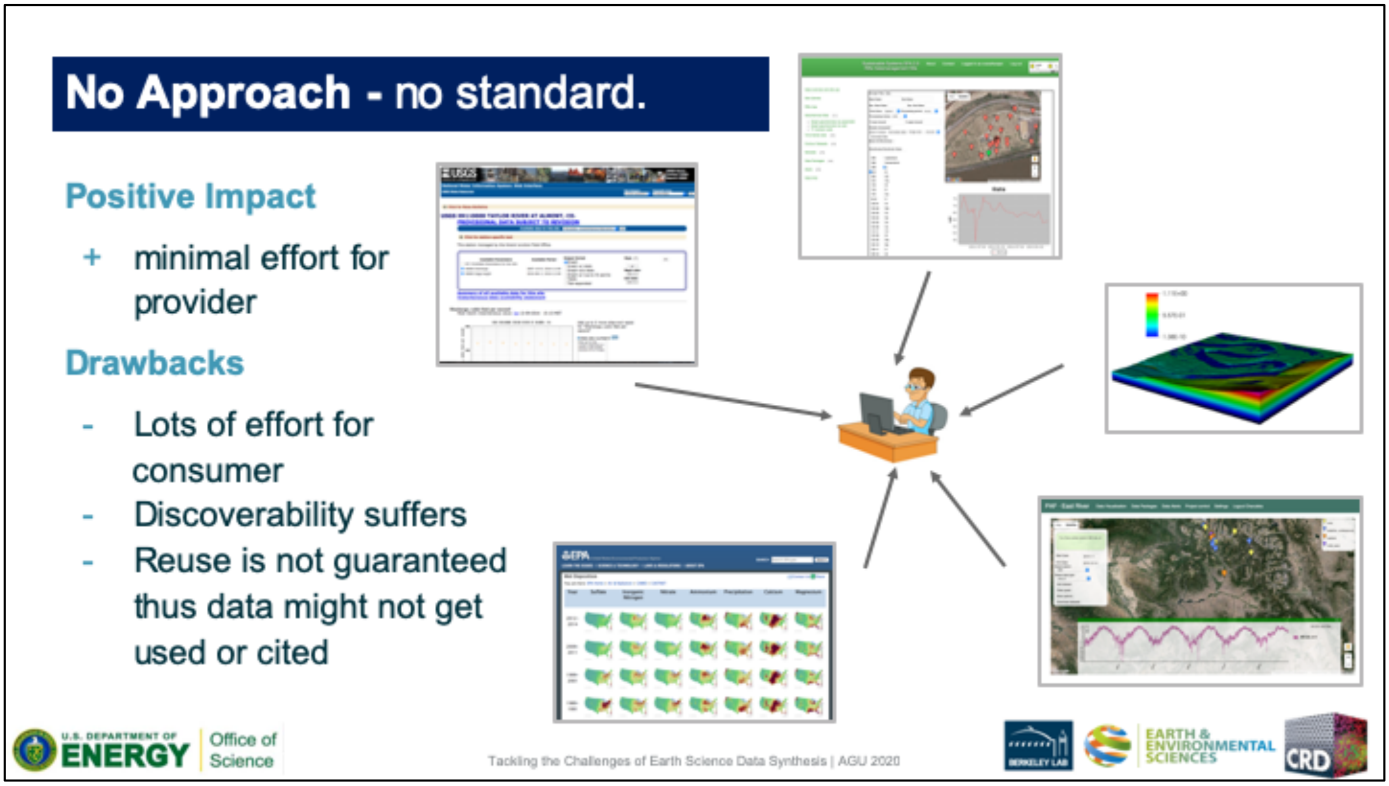
iNAIADS is a **Department of Energy Early Career Research project** using data-driven approaches to predict water quality.

Watershed Function SFA project is developing understanding and tools to measure and predict how droughts, early snowmelt, and other perturbations impact downstream water availability and biogeochemical cycling at episodic to decadal timescales.

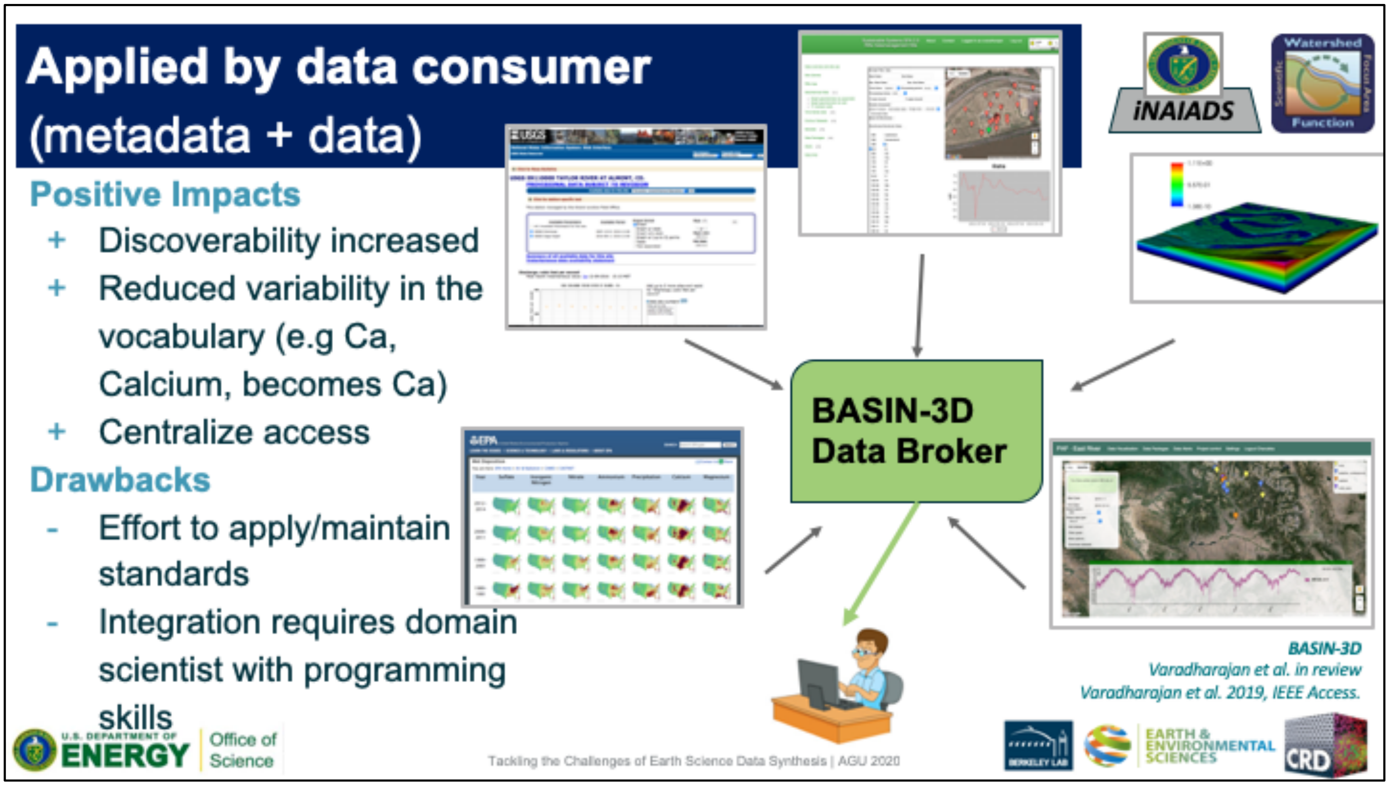
ESS-DIVE is a data repository for Earth and environmental science data. ESS-DIVE stores and publicly distributes data from observational, experimental, and modeling research funded by the DOE's Office of Science



Now I will discuss four approaches to standards. no approach, standards applied by the data consumer, standards applied by the data provider and finally I will discuss a combined approach that is inspired by the previous mentioned approaches.



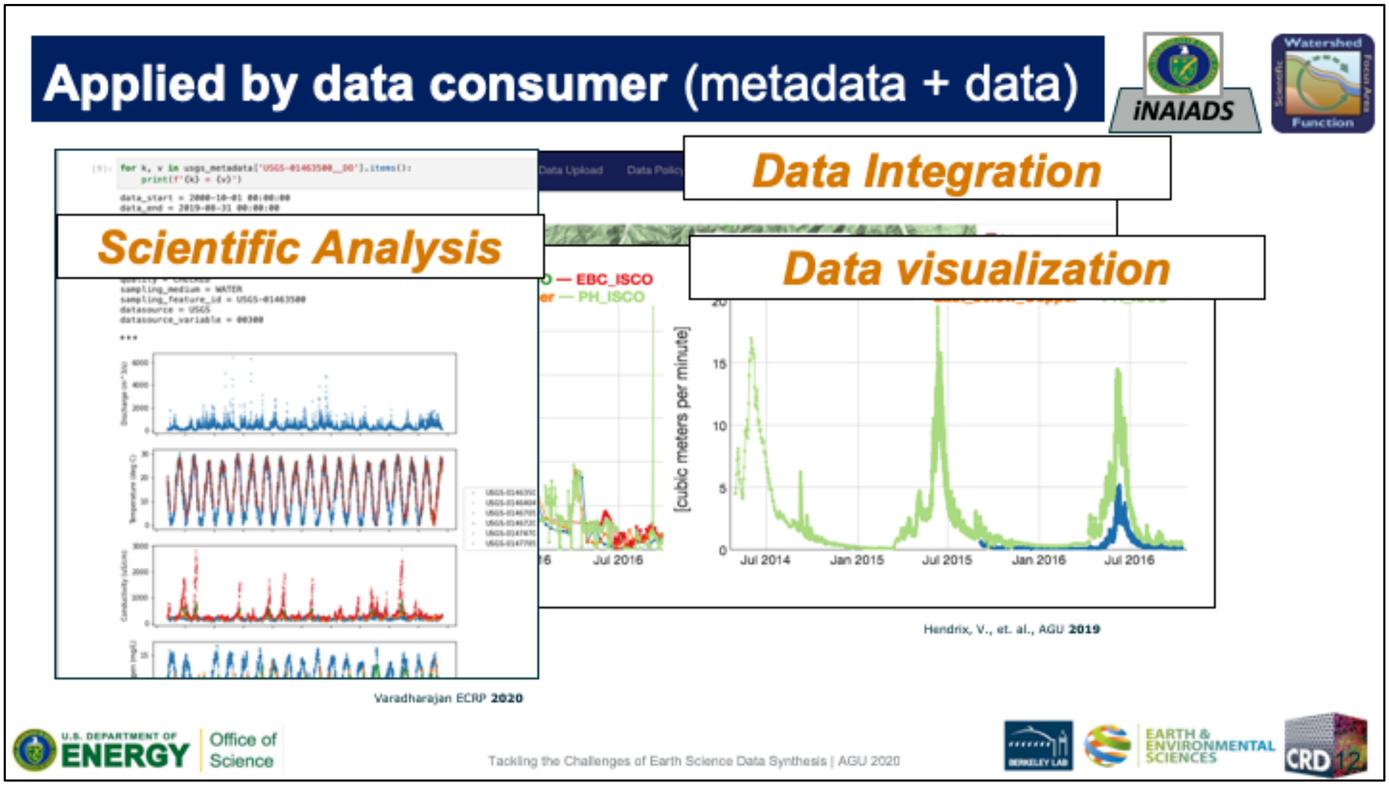
No Approach - no standard. Data consumer accesses each data repository separately and transforms the data to custom structure of their choosing. On the positive side this takes not extra effort on the part of the data provider but it is lots of effort for the consumer. The discoverability of each dataset suffers which means it might not get used or cited and thus reduces the potential scientific impact.



In this next approach by Watershed Function SFA and ***iNAIADS*** a data brokering framework was developed called BASIN-3D that integrates data and metadata and data from disparate, heterogeneous data sources and applies a standard to the metadata data for use by downstream users as you can see from this graphic.

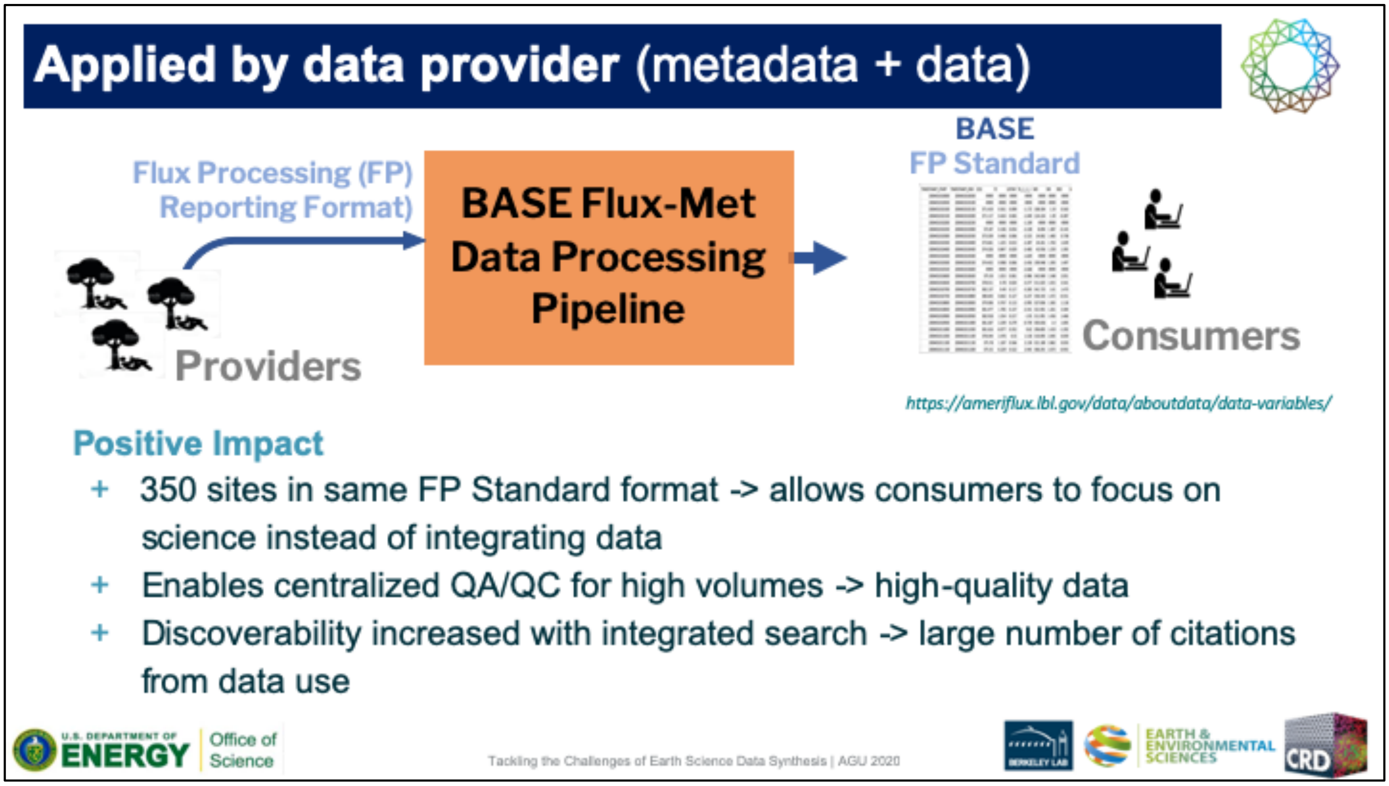
For downstream users this increases the discoverability of the data through integrated search across data sources, Removes the variability in the vocabulary such as variable names and provides centralized access to download the data

The downside is that it takes considerable effort to apply standards to each data repository and keep up-to-date with changes in data and metadata structure. Additionally integrating new data sources requires efforts by a domain scientist with programming skills.

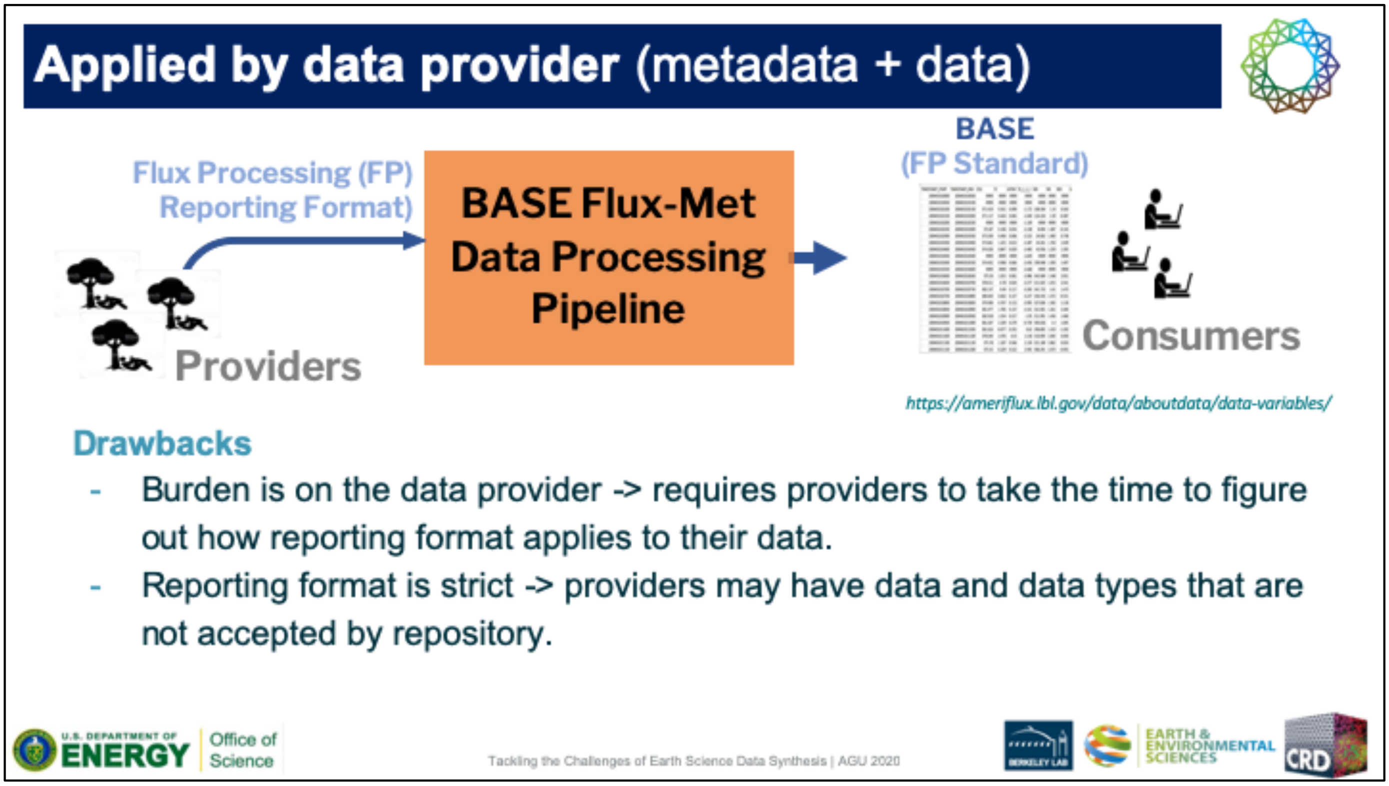


The BASIN-3D technology has enabled data integration where project users can search harmonized data across data repositories.

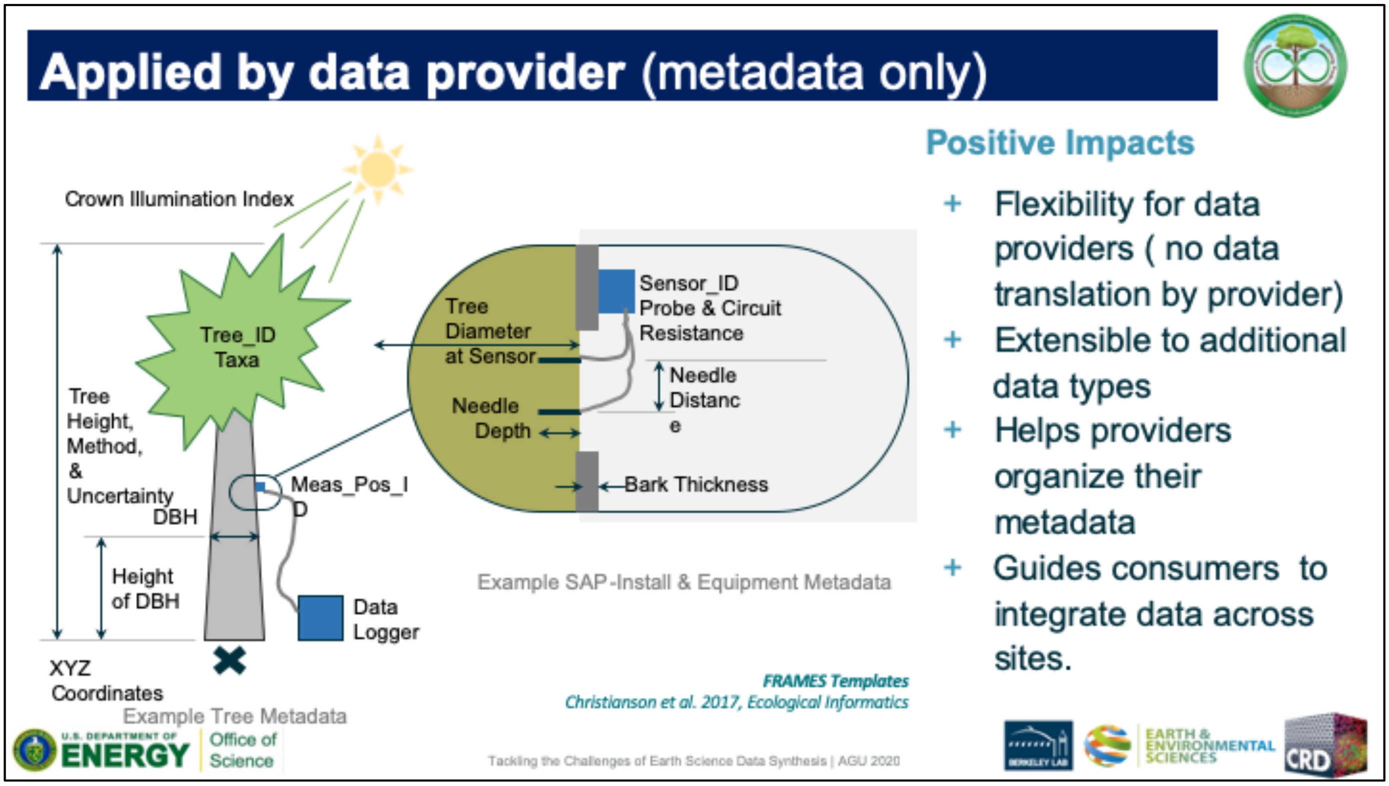
It also enables data visualization where users can preview the data to determine its usefulness to their research before download, and scientific analysis of harmonized data as seen in this Jupyter Notebook.



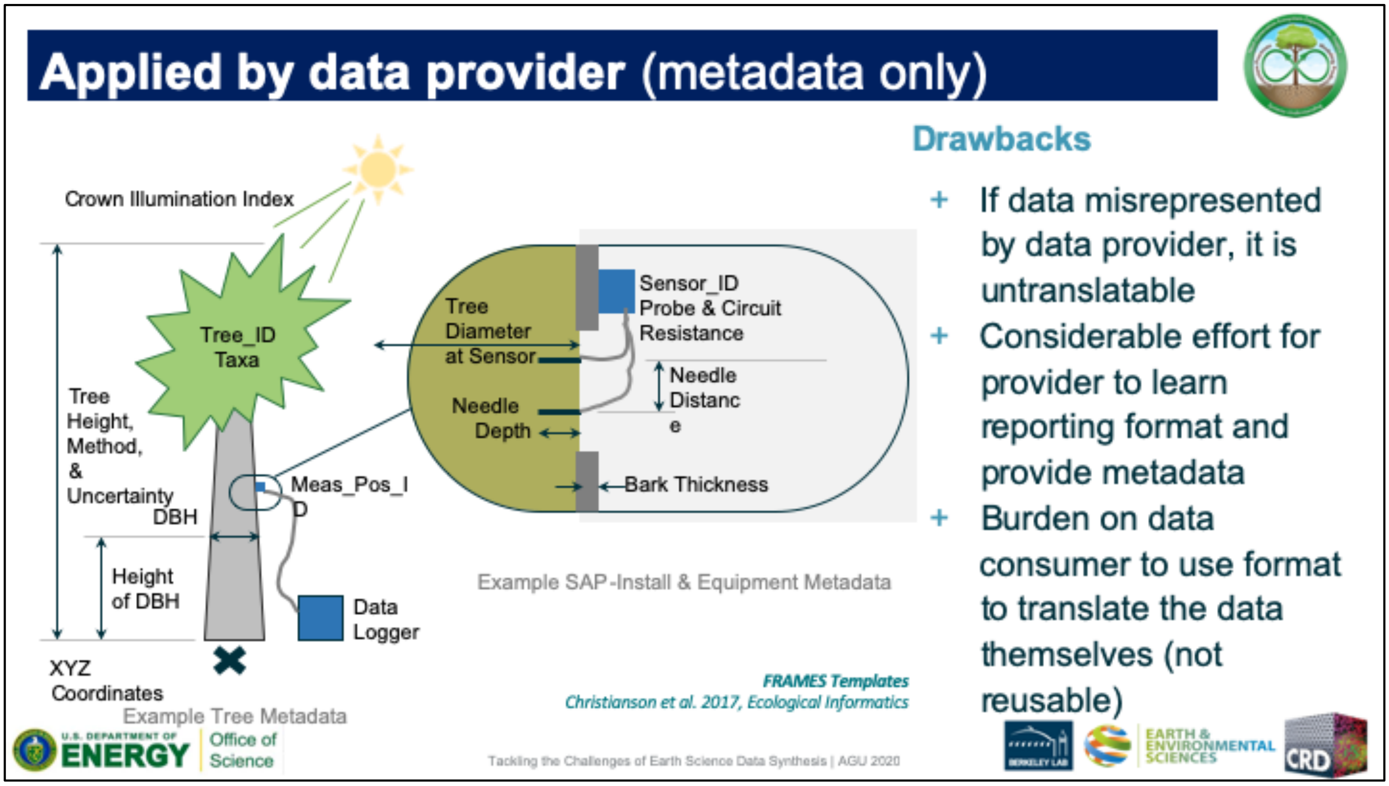
In this approach by Ameriflux standards are applied by the data provider to the metadata and data.
Using this approach Ameriflux is able to provide data and metadata in the same FP standard format which allows scientists to focus on the science instead of data integration. It also enables a Ameriflix to execute a centralized QA/QC pipeline for high volumes of data which results in high quality data. Finally discoverability is increased with and integrated cross-site search which contributed in an increase of citations from data use.



The downsides of this approach are that requires data providers to take the time to figure out how reporting format applies to their data and also the Reporting format is strict. If providers have data and data types that are not defined in the reporting format it is not accepted by repository.



In this approach by NGEE Tropics the metadata only is provided by the data providers. Providing formatted metadata only allows the data providers more flexibility because they can describe their data in the way that they want. This is very extensible and enables new data types to be added when needed. .Some providers viewed this approach as a way to organize and preserve their metadata for the long-term. One consumer, a modeler was able to write a script to use data from these templates to integrate data from sites across various tropical regions.



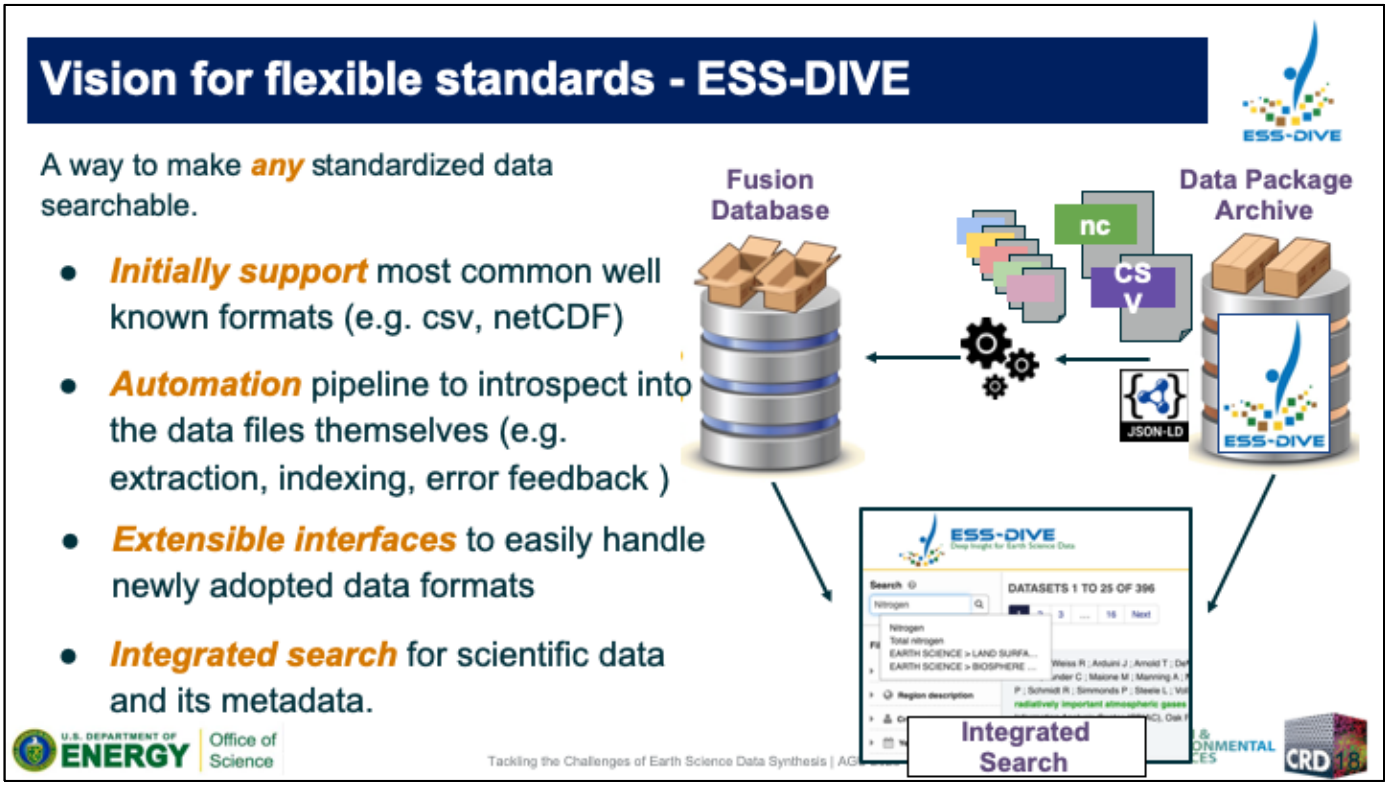
This approach however does not work well if data providers make mistakes in their metadata as the data would be untranslatable. Learning the metadata format standard takes time and effort for the providers to understand and apply to their data. The onus is on each data consumers to use each provided metadata format to translate the data themselves. However it has been demonstrated that a single script could translate any data describe by the metadata format

Vision for flexible standards - ESS-DIVE

Combining approaches

- ***Data standards*** or reporting formats applied by data providers
- ***Metadata standards*** applied by data providers,
- If ***no standards***, I'm feeling lucky approach (i.e., what can we extract from naive parsing algorithms)

As mentioned previously, ESS-DIVE is a data repository that accepts stores and publicly distributes data from observational, experimental, and modeling research. The data in ESS-DIVE is diverse and complex and the approach we are taking aims to reduce the burden on the data providers by working with our ESS community to develop and adopt "standards" (reporting formats) for select ESS data types. ESS-DIVE's approach builds on the insights gained from the previously mentioned approaches.

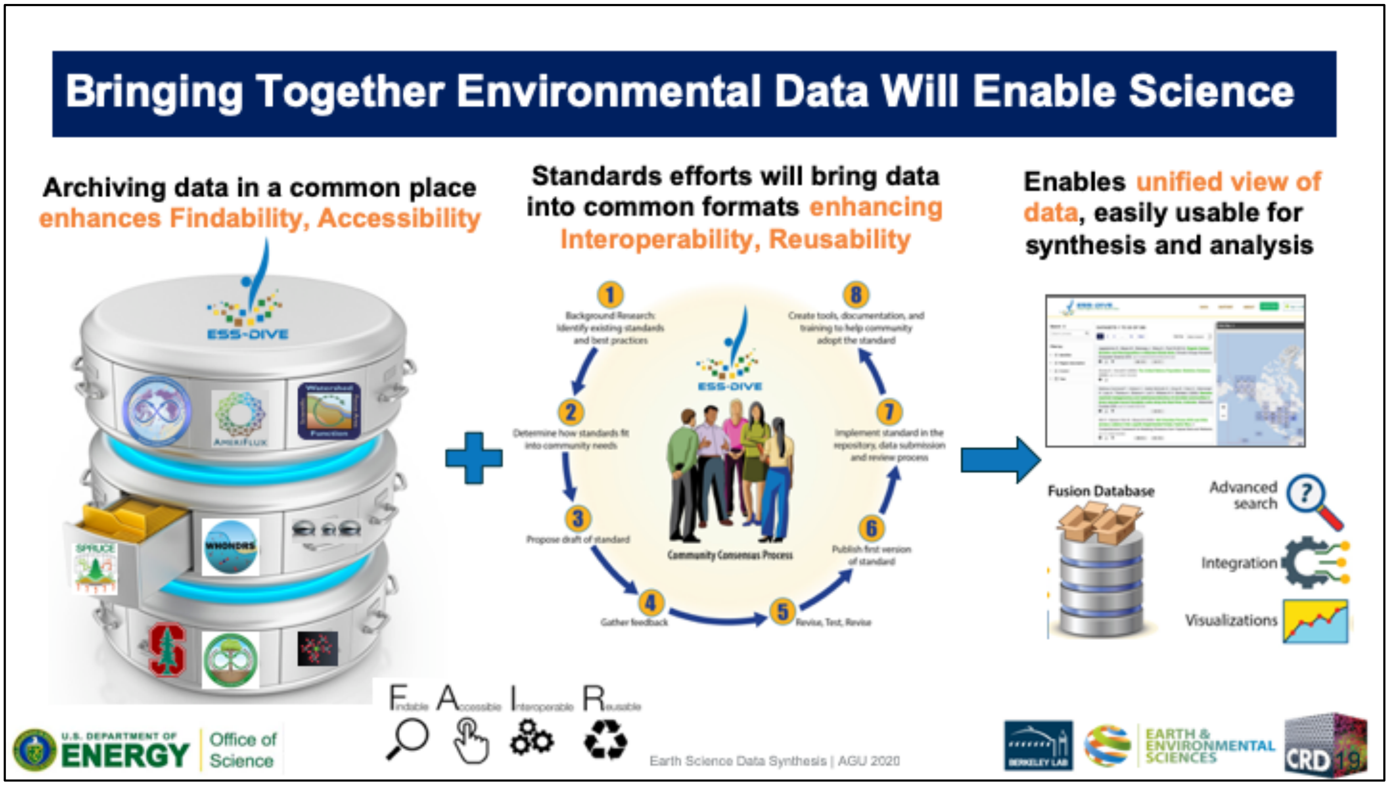


The ESS-DIVE vision for the future is to make use of these standards is by creating what we call a fusion database to enable deeper access.

ESS-DIVE is a data repository that allows data providers to create and publish data packages. A data package consists of some high level metadata (title, abstract authors) plus corresponding data files bundled together in a single package.

The purpose of the ESS-DIVE fusion database is to make any standardized data searchable by extending the search beyond the general data package metadata into the files themselves.

- We will initially support the most common well known formats. Towards this goal, we have already created a limited prototype that extracts data from CSV files and provides summary statistics of the data to show feasibility of this approach
  We envision a parser for each data standard format that can be applied to any data files which adhere it.
- We will develop an automation pipeline of interfaces that allow for automated data extraction, indexing and error handling.
- These interfaces will be extensible allowing us to easily handle new data standard formats as they are adopted.
- All of this will will result in an integrated search for both the package-level metadata and its corresponding scientific data in the ESS-DIVE data portal.



Scientists should care about getting involved in how data repositories are being designed and built since it can enable your science. In the case of ESS-DIVE, what we hope that in bringing many of our datasets into one place, we will enhance their findability and accessibility

Through our standards efforts to bring data into common formats, we want to enhance their Interoperability and Reusability.

What this will then enable for the science is first a unified view of the data where you can find the data needed more easily for synthesis and analysis.

Use Case	Standard Type		Standard Applied By		Standard Application
	Data	Metadata	Provider	Consumer	(strict/loose)
No Approach				?	N/A
WF-SFA, iNAIADS (BASIN-3D)	X	X		X	strict
AmeriFlux	X	X	X		strict
NGEE Tropics		X	X		strict
ESS-DIVE	*X*	*X*	*X*	*?*	*loose*

X Standard - includes both reporting formats and standards
? Custom format - most likely not consensus-based

***Data integration** is important*

*Standards are **essential***

*Still **more work** ahead*

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB EARTH & ENVIRONMENTAL SCIENCES CRD

This table summarizes approaches of each use case and how they use standards and reporting formats to aid integration of data. The last row shows the ESS-DIVE use case which as I mentioned is a vision for the future.

The Xs indicate what standards are applied and who applies them. The question marks indicate that there is some structure but that it is most likely not consensus based like standards and reporting formats,

While none of the approaches are perfect we know from our experiences that:

+ Data integration is important for analysis synthesis, machine learning and AI
+ Standards are essential in whatever the form and whoever applies them. We have already seen the impact to our projects in increased usage and citation
+ This is still a work in progress and we need to make is easier for providers to submit well structured data and for consumers to discover and use data in their science.

Thanks for your time. The projects mentioned in this presentation are funded by the Department of Energy's Office of Science BER Program.

If you have any questions, contact me at the email listed.