NASA’s data traceability study

Rahul Ramachandran, **Manil Maskey**, Chris Lynnes
NASA

Arun John, T. Mukherjee
University of Alabama in Huntsville

CEOS WGISS-53
March 22-24, 2022

Summary: NASA Earth Observation data traceability study

Scope:
- Investigate problems that arise due to duplication and mirroring from a data user's perspective
- Explore a range of technical solutions to address this data traceability problem
- Focus only on addressing a set of specified use cases

Study Facets:
- Data Integrity Checks
  - Authenticity
  - Data Deletion
  - Data Modification
- Non-repudiation
- Ease of Implementation
- Performance

Approaches:
- Filename lookup
- Data upload and Verify
- Hash lookup
- Provable data possession
- Signed hashes

Approach	Data Integrity			Non-repudiation		Ease of Implementation
	Deletion	Modification	Authenticity	Non-Repudiation-without data transfer	Verification mechanism	
Filename Lookup	YES	NO	None	NO	Lookup	Low - current setup allows this
Data upload+verify	YES	YES	100%	NO	Data transfer	Low - but HIGH cost to maintain
Hash Lookup	YES	YES	100%	NO	Lookup	Low - current setup can be modified
Provable Data Possession	YES	YES	Probabilistic	YES	Challenge-Response	High
Signed Hash	NO	YES	100%	NO	Lookup	Medium - sign hashes during existing data ingest process and share public key on the product website

Key takeaways:
- No approach addresses all the identified needs, both hash lookup and signed hashes are viable practical approaches to address some of the issues
- Hash lookup is suitable if the data user wants to process the full archive of a data set for any kind of large scale processing need

Basic Definitions

- Authoritative source: an entity that is authorized to develop or manage data for a specific purpose
  - Can also be an actively managed repository of valid or trusted data that is recognized by a community and supports data governance and stewardship practices.
- Trusted source can be a service provider or an organization that publishes data obtained from a number of authoritative sources
  - Are often compilations and subsets of the data from more than one authoritative sources
  - They are “trusted” because there is an “official process and/or agreement” for compiling the data from authoritative sources, data is actively managed, and documented
- Science users typically obtain data from authoritative or trusted sources.

Problem statement

- Cloud computing driving authoritative and trusted sources concept
- A data lake is a core component in cloud-based analysis paradigm providing a flexible data store to address data processing and analysis needs at scale.
    - Ideally there exists only one authoritative or trusted data lake on a cloud platform
- Each agency or data producing organization is using their own cloud platform based on their constraints or some are still using on premise infrastructure
    - Data products are being replicated and mirrored at different cloud platforms by third party actors
- Data duplication by third party actors, without proper data stewardship is threatening to create data swamps.
    - More critically it lacks data curation with little to no active data management throughout the data life cycle and little to no contextual metadata and data governance

Use Case Scenarios

- Ideal
    - all the authoritative data required is available on the same cloud platform
    - user only requires access to the cloud storage and notification mechanism
- Worst case
    - different data products are unavailable on the cloud platform
    - burden falls on the user to replicate the archive from different authoritative sources
    - user now requires a uniform scalable data access mechanism
- Common case
    - some data products may be available on the selected cloud platform, but their provenance is unknown.
    - user now has to not only move the unavailable data but also verify the authenticity of available data
    - in addition to a uniform approach for data access for the missing data, the user needs a uniform approach to verify the authenticity, completeness, correctness and consistency of the data available on the cloud

Actors

- Data Producers/Data Distribution Centers – source of the authoritative or trusted data
- Data Stewards – organizations within authoritative sources charged with the collection and maintenance of authoritative data
- Data Users – Consumers of the data
- Third Party Actors – Individuals or organizations replicating the data
- Verifier - Service or an actor that validates the integrity of the data

Study Facets

- Data Integrity Checks – three types:
    - Authenticity: A data user should be able to determine if the data file they accessed from a third party repository is from an authoritative source. For example, a user should be able to verify that a specific file is genuinely from NASA's data systems.
    - Data Deletion: A data user should be able to check if the file they accessed from a third party resource has been deleted at the authoritative source. For example, the original data product that has been mirrored to a cloud platform has undergone a major version update and the older version is no longer considered fit for use.
    - Data Modification: A data user should be able to check if the file they accessed from a third party resource has been modified or tampered with.
- Non-repudiation – the assurance that someone cannot deny the validity of something. For data users, it is the ability to prove their use of authoritative data despite claims to the contrary.
- Ease of Implementation – a qualitative assessment of the effort data producers and data stewards will need to expend if they adopt a specific technical approach.
- Performance - the overhead of extraneous data movement in or out of the cloud or of computing verification parameters must not outweigh the analysis performance gains.

Approaches

Filename Lookup

- filename is checked, not the contents

Data Upload and Verify

- users upload the files they accessed from a third party source to a verification service provided by the authoritative source
- involves high data movement, making it expensive and time consuming

Hash Lookup

- hash values for files are then stored in an authoritative source's catalog or in a file that is accessible
- user can query the catalog or download the file for the hash and compare it with the hash generated for the file they have accessed from the third party source for verification

Provable Data Possession

- combines public key cryptography and cryptographic hashing. It has three steps: (1) producer key generation, (2) tag generation, and (3) verification

Signed Hashes

- instead of requiring the entire file to be signed, only the hash generated for the content is signed using the private key
- data users can authenticate the hash signature by using the public key made available by the authoritative source and verify the integrity of the file itself by comparing the signed hash against the hash they generated

Result Summary Table

Approach	Data Integrity			Non-repudiation		Ease of Implementation
	Deletion	Modification	Authenticity	Non-Repudiation-without data transfer	Verification mechanism	
Filename Lookup	YES	NO	None	NO	Lookup	Low - current setup allows this
Data upload+verify	YES	YES	100%	NO	Data transfer	Low - but HIGH cost to maintain
Hash Lookup	YES	YES	100%	NO	Lookup	Low - current setup can be modified
Provable Data Possession	YES	YES	Probabilistic	YES	Challenge-Response	High
Signed Hash	NO	YES	100%	NO	Lookup	Medium - sign hashes during existing data ingest process and share public key on the product website

Key Takeaways

- Both hash lookup and signed hashes are viable practical approaches to address some of the issues
- Hash lookup is suitable if the data user wants to process the full archive of a dataset for any kind of large scale processing need. Full listing of the files along with their hash values is required for completeness, correctness, and consistency checks
- Signed hashes are suitable when the data users only want subsets of the archive for analysis and completeness is not an issue. The data users can verify the data authenticity using the public key from the authoritative source and validate the integrity using the hash value
- It is the data user’s responsibility to verify the authenticity, completeness, correctness, and consistency of the data they are using

EXPLORING BLOCKCHAIN TO SUPPORT OPEN SCIENCE PRACTICES

Iksha Gurung[1], Slesa Adhikari[1], Abdelhak Marouane[1], Rajesh Pandey[1], Satkar Dhakal[2], Manil Maskey[3]

University of Alabama in Huntsville
Vanderbilt University
NASA Marshall Space Flight Center

Open Science

“Open-source science is a commitment to the open sharing of software, data, and knowledge (algorithms, papers, documents, ancillary information) as early as possible in the scientific process.”

- *Open-Source Science Initiative, NASA*

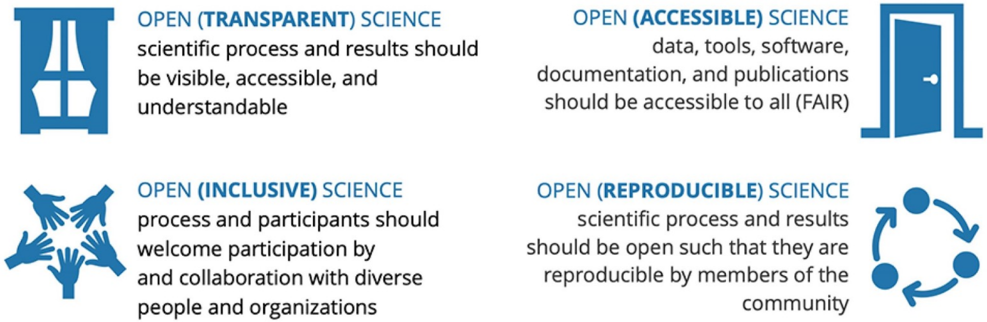


Fig: Characteristics of Open Science, https://science.nasa.gov/open-science-overview

Challenges:

- Management of scientific resources, data
- Data verification
- Proper attribution
- Ensuring transparency throughout the research process

Blockchain for Open Science

Open Science Infrastructure Requirements / Blockchain Characteristics	Decentralization	Cryptographic Hashing	Timestamping	Immutability (Append-Only)	Consensus Mechanism	Access and Governance System
Collaborative Environment	X				X	X
No censorship	X	X		X	X	
Open Data		X				X
Open Access		X				X
Identity and reputation management	X	X			X	X
Extensible system					X	
Incentives for collaboration and sharing	X				X	X
Equality of all participants	X				X	X
Simple workflow integration					X	
Data and content sharing	X	X			X	X
Crowdfunding		X			X	X
Trail of research (objects)	X	X	X	X	X	
Citizen Science	X	X	X			X
Open Source code and tools	X					X
Resource sharing	X				X	X
Metrics	X				X	X
Connected systems						X

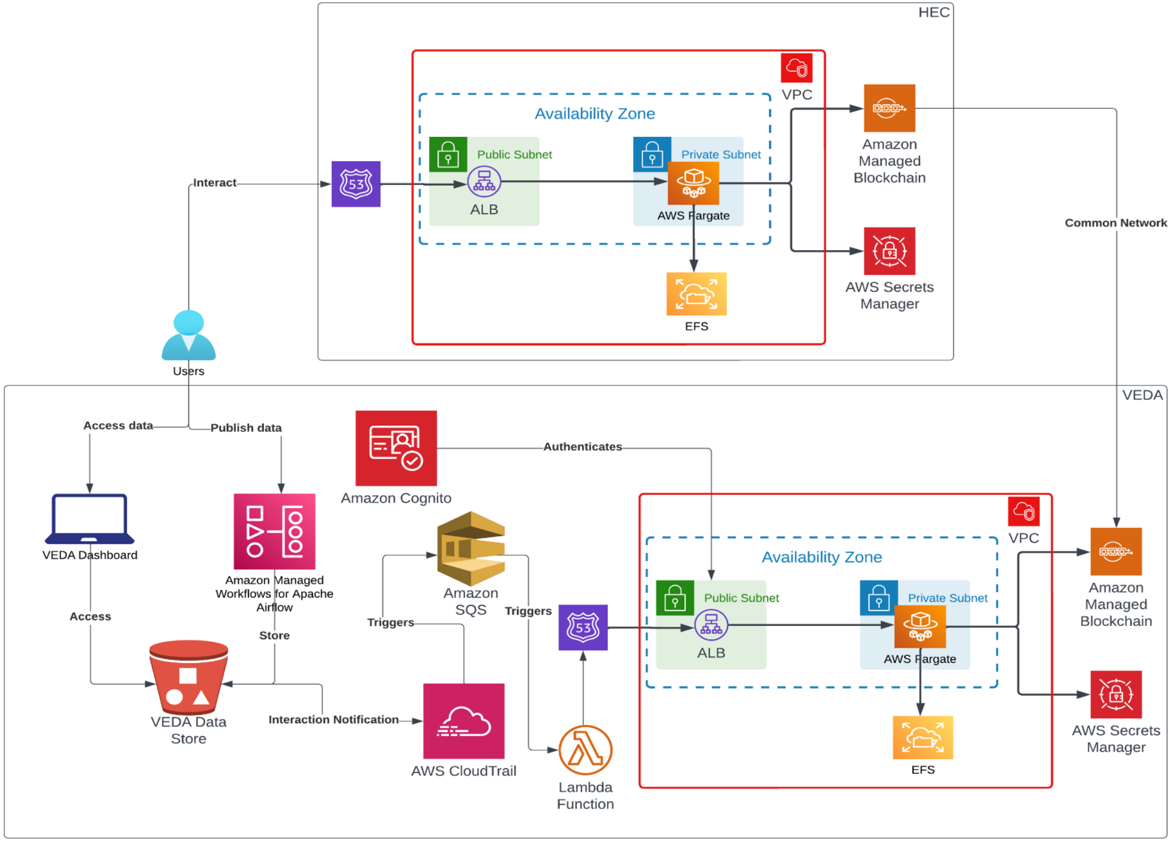
Fig : Open science infrastructure requirements and blockchain technology characteristics that are fulfilling them, Leible et.al

From : https://doi.org/10.3389/fbloc.2019.00016

Open Science Blockchain Use case

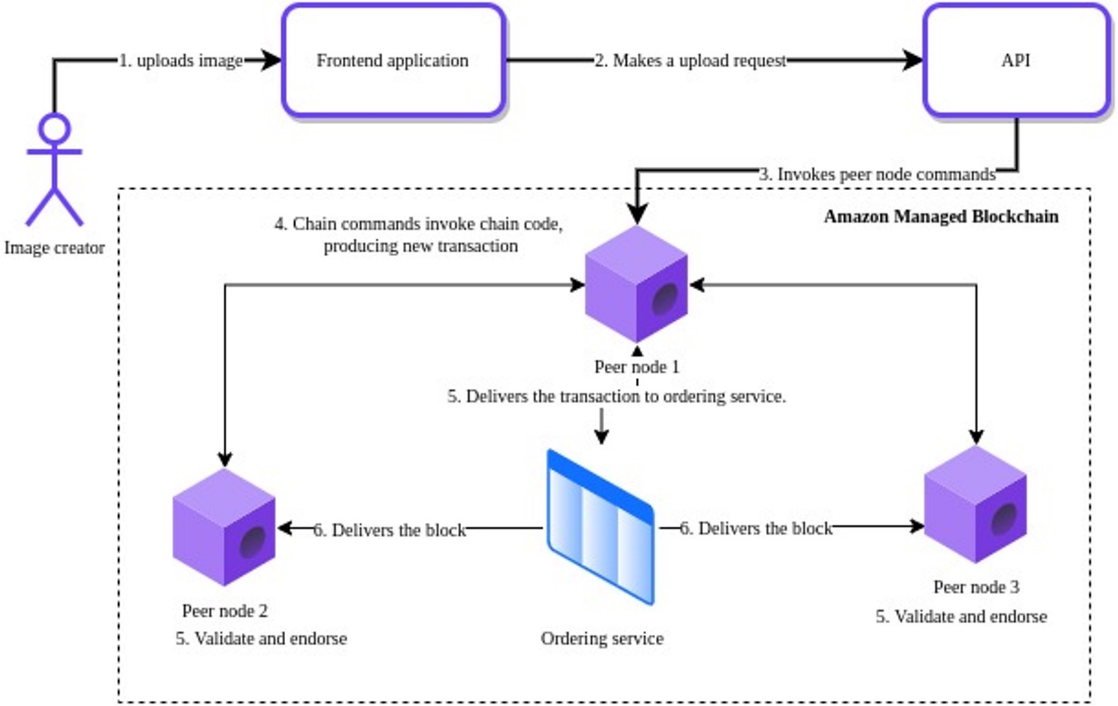
- Utilizing Visualization, Exploration, and Data Analysis (VEDA) project for the proof of concept.
- VEDA
  - utilizes open-source tools and resources to create an open-source science cyberinfrastructure for data processing and geographic information systems (GIS) capabilities
  - allows scientists from different organizations and teams to contribute datasets. These datasets are used to create discoveries or Data Stories to highlight multiple events
- VEDA's collaboration with High-End Computing program (HEC) used to demonstrate how Open Science Blockchain fosters Open Science.

Implementation



Architecture diagram

Implementation



Usecases

Nitrogen Dioxide (NO2) and NO2 difference (NO2 diff) data in VEDA

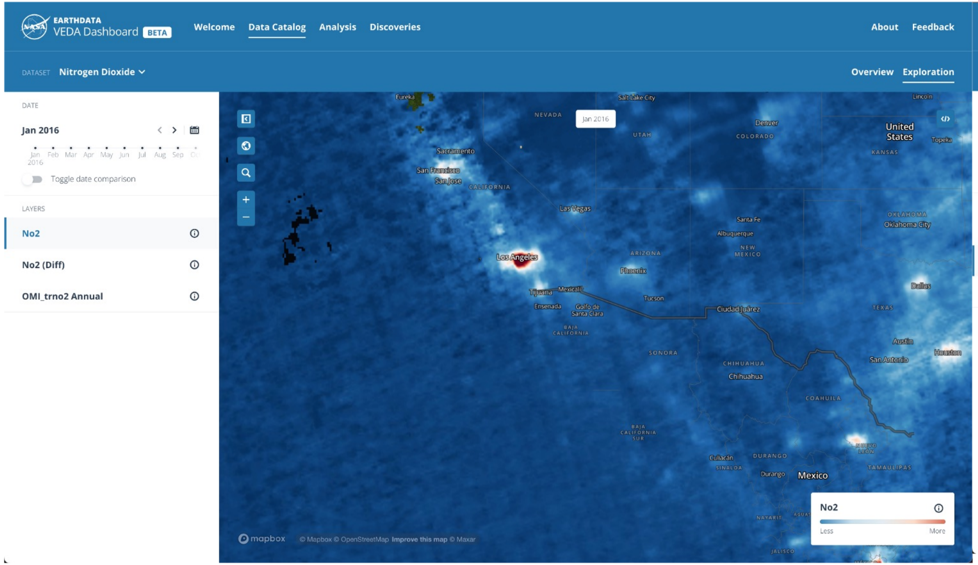


Fig: N02 data visualization in VEDA Dashboard

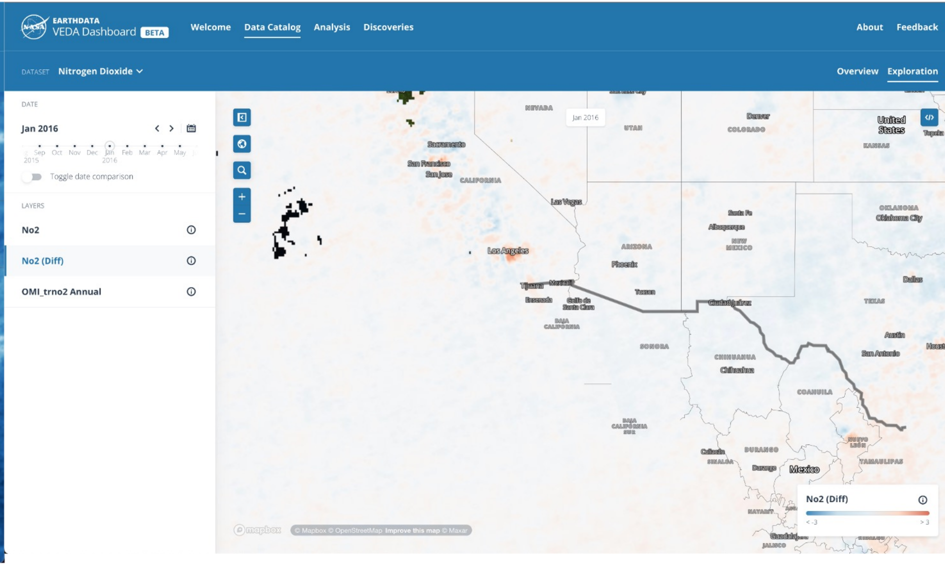
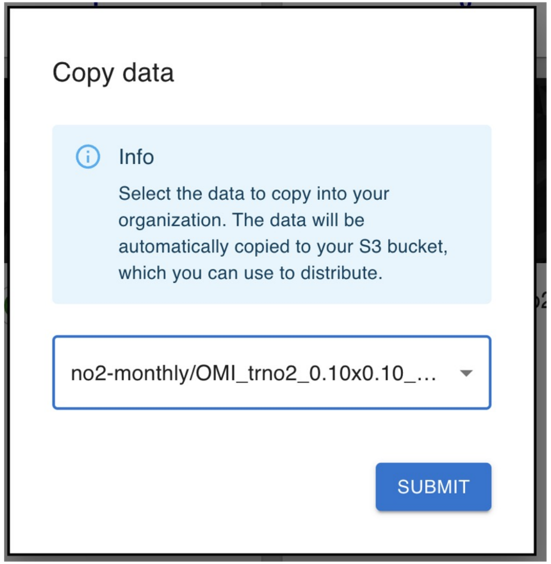


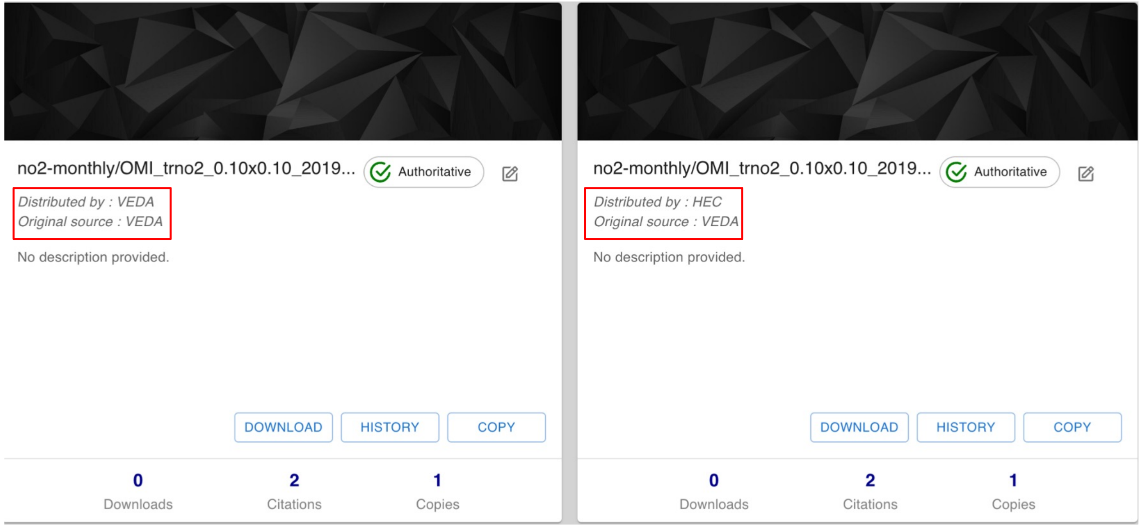
Fig: N02 diff visualization in VEDA Dashboard

Usecases

Use case 0: HEC Copies data from VEDA to become authoritative distributor of the data

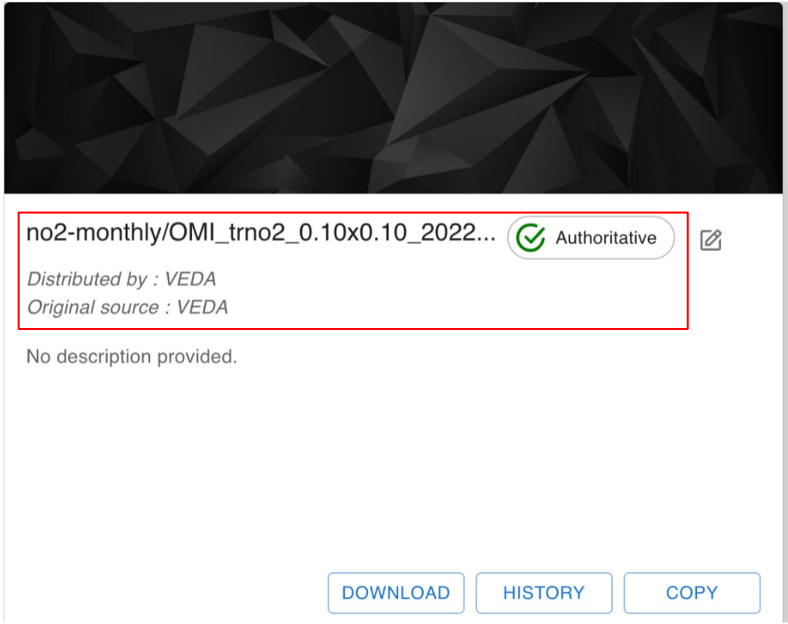


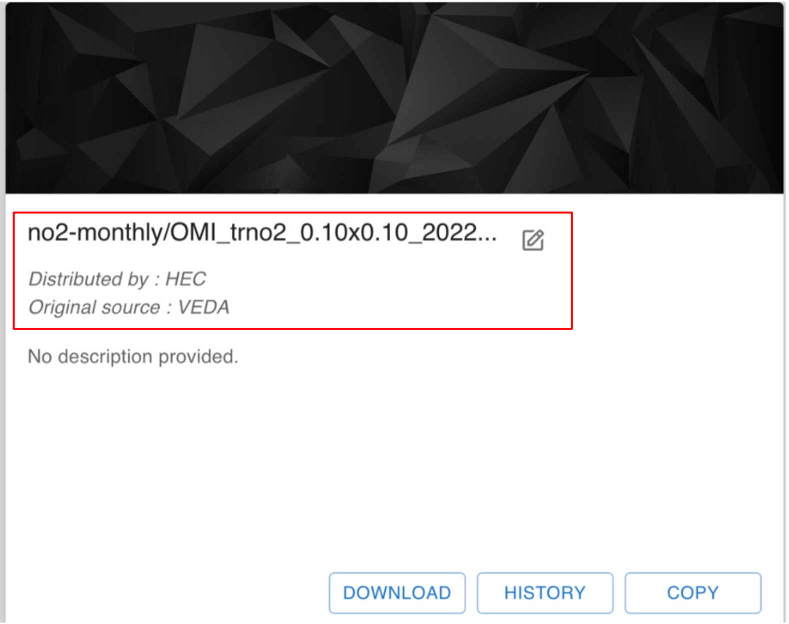
116	no2-monthly/OMI_trno2_0....	VEDA
117	no2-monthly/OMI_trno2_0....	HEC



Use cases

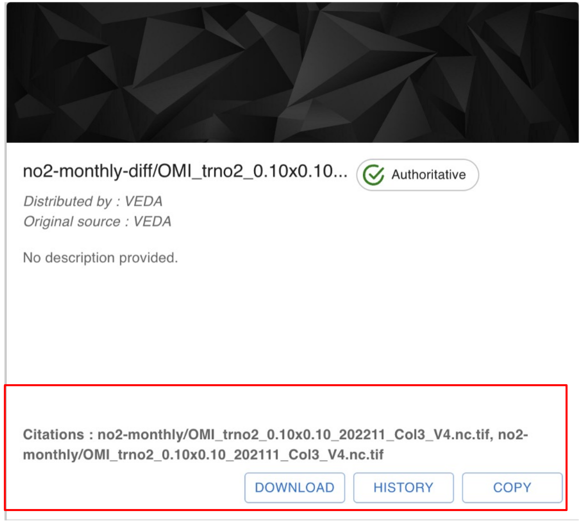
Use case 1: Differentiate between authoritative and non-authoritative data

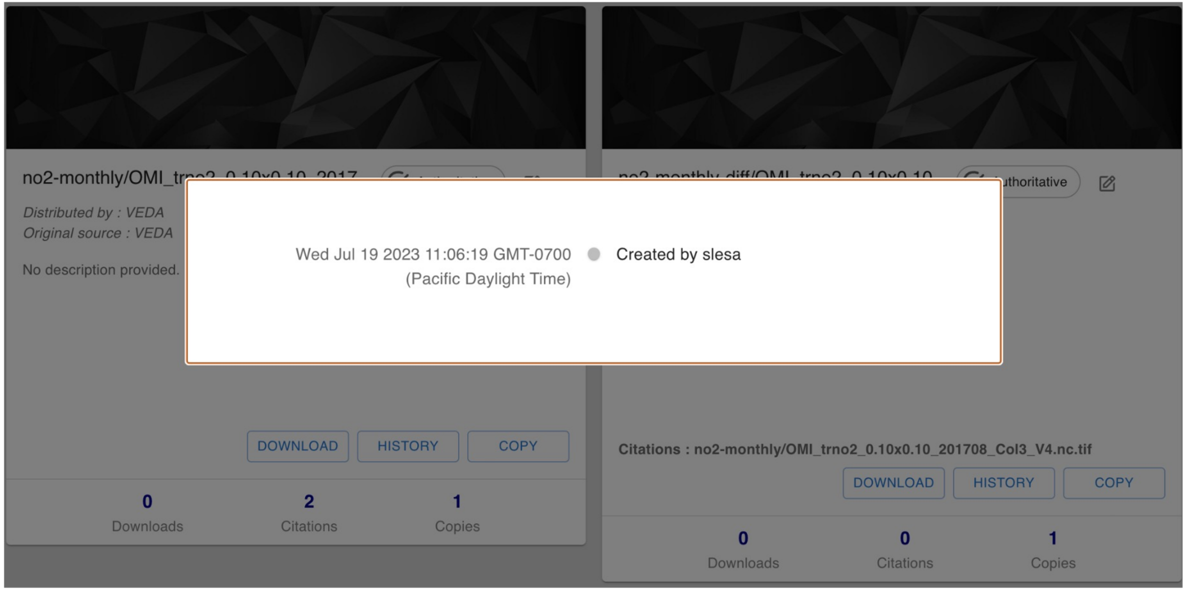




Use cases

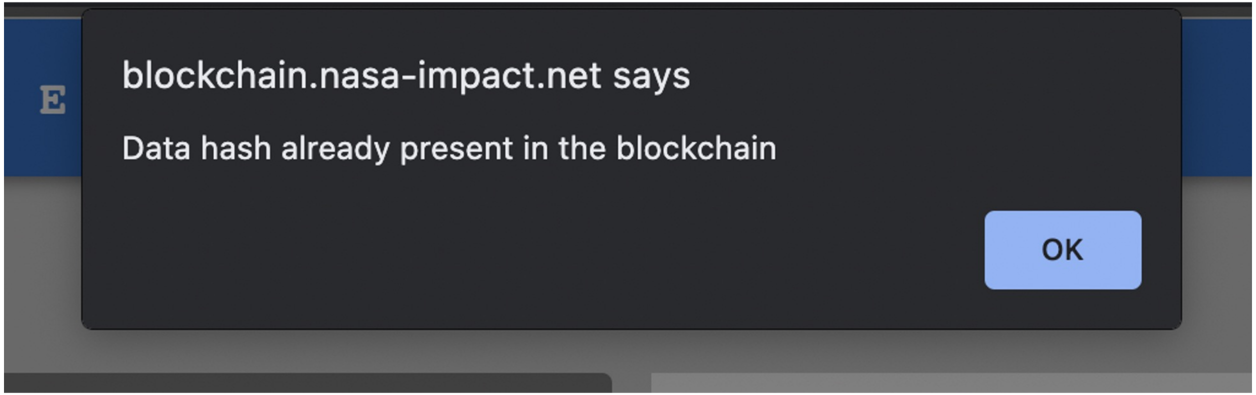
Use case 2: Citation is enabled to support open science





Use cases

Use case 3: Re-uploads to the network are not permitted



Use cases

Use Case 4: Adding a new organization to the network

Invitations (1) Info

View details | Reject invitation | Accept invitation

Network name	Invitation ID	Network ID	Description	Status	Expiration Date	ARN
osblockchain	[redacted]	[redacted]	network-osblockchai...	Accepted	8/30/2022, 6:33:32 ...	arn:aws:managedblo...

Voting Summary

Yes	No	Outstanding
2	0	0

Votes

All

Member name	Member ID	Vote
member-osblockchain	[redacted]	Yes
member2	[redacted]	Yes

Conclusion

- The presented case study demonstrates practical implementation and contributes insights for further research.
- Blockchain technology offers solutions for:
  - Addressing challenges in open science
  - Promoting transparency, reproducibility, and accountability
  - Decentralized and secure sharing of authoritative data
  - Traceability of scientific resources and proper attribution
- Blockchain-based incentives foster collaboration and efficient research practices.
- Adoption of blockchain in open science requires:
  - Policy frameworks
  - Change in data management approaches
  - Cost considerations
- Full adoption of blockchain in open science will take time due to:
  - Policies around data citations and access
  - Additional cost of maintenance and development resources
  - Exploration of decentralized data storage and distribution for proper open science
  - Continued research is essential to unlock blockchain's full potential for advancing open science.

National Aeronautics and
Space Administration

NASA SMD Initiatives to Lower the Boundaries to Science Research: *Earth Science Division Activities*

Dr. Manil Maskey
Earth Science Division
Science Mission Directorate
NASA Headquarters

Reducing research barriers

1 Trilateral EO Dashboard

- Global collaboration
- Education and training
- Data sharing
- Open source software

2 VEDA

- Open source software
- Accessible platform
- Standards & interoperability
- Interdisciplinary science

3 AI Foundation Model

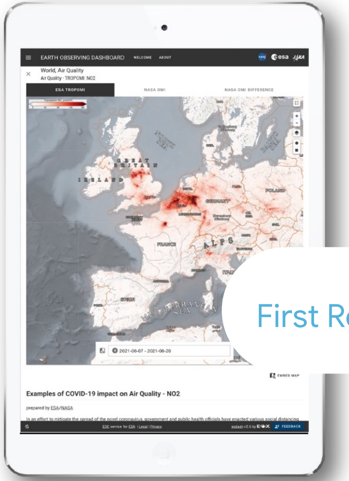
- Partnerships
- Collaboration
- Open large scale AI models
- Reduced cost to build downstream applications

4 Earth Information Center

- Public outreach
- Education
- Promoting scientific insights

Trilateral EO dashboard

Trilateral EO Dashboard: objectives and timeline



First Release 2020

eodashboard.org

Objective 1

- **Demonstrate joint capabilities of NASA-ESA-JAXA to observe COVID-19's environmental and economic impacts from space**

Objective 2

- **Convey indicators to the general public and decision makers; and maximize the value of curated information**

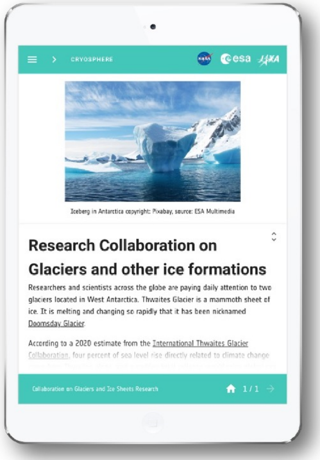
Objective 3

- **Engage the wider public via outreach & competitions, e.g. EO Dashboard Hackathon, SpaceApps**

KEY ACHIEVEMENTS

Rapid release	EO Indicators	Global Presence	Communication
3 months	**12 EO missions**	**146 countries**	**251 websites**
Eodashboard.org released in June 2020	Indicators based on data from 12 EO missions of ESA, NASA, JAXA	Global awareness with accesses across the world	Citations by 251 websites and joint participation in CEOS, AGU, SpaceApps, etc.

Trilateral EO Dashboard: ongoing collaboration



2021 - ongoing

Objective 1

- Broaden scope for 7 thematic areas: atmosphere, oceans, biomass, cryosphere, agriculture, economy, covid-19, using open data and interoperability across agencies

Objective 2

- Communicate scientific findings via advanced visualization exploiting tri-agency EO data

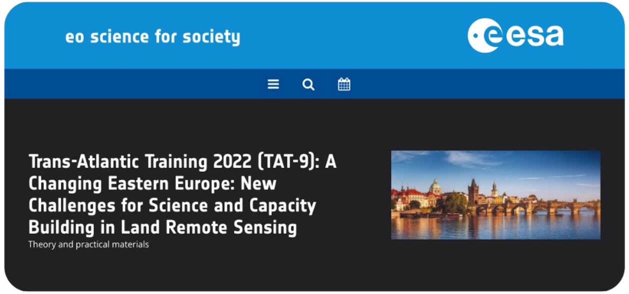
Objective 3

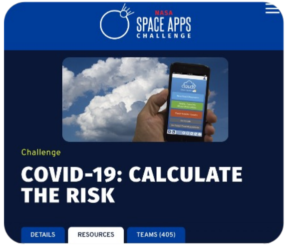
- Promote Open Science best practices and engage with the community

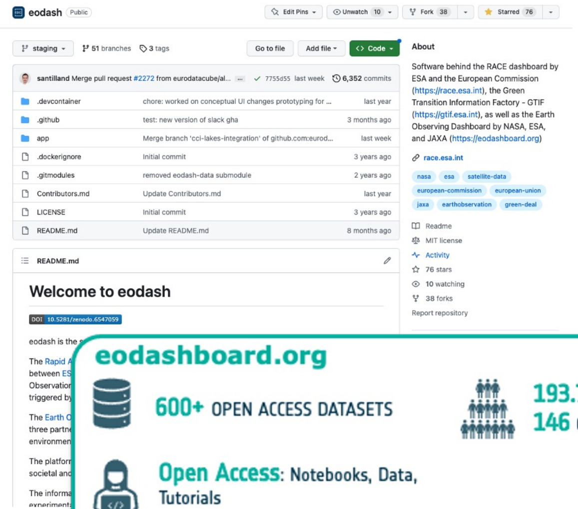
KEY ACHIEVEMENTS

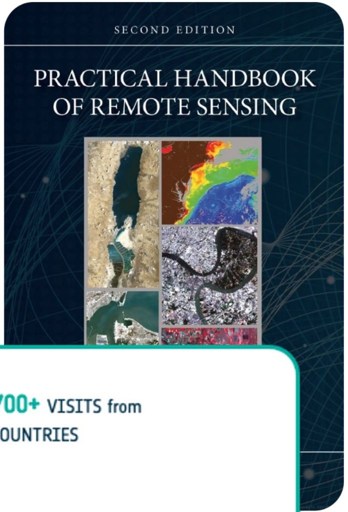
Open Data	**Storytelling with Open Data**	Training and Education and Open Science	Promote Cooperation
25 EO missions	**19 stories**	**5 Workshops**	**IAF Award**
Expanded EO indicators covering 7 new domains	Jointly developed stories to communicate tri-agency scientific findings	IGARSS tutorials and workshops, FOSS4G, LPS, EGU, GEO ODOK, etc.	**GLOC 2023 IAF Award**

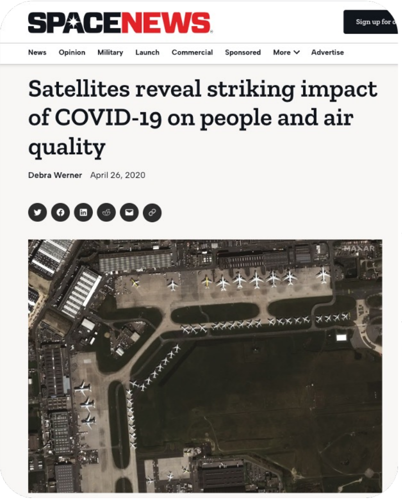
User community

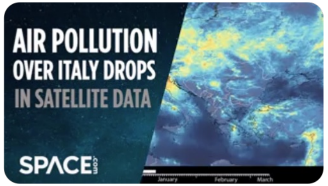












EO DASHBOARD HACKATHON

eodashboard.org

600+ OPEN ACCESS DATASETS

193.700+ VISITS from **146** COUNTRIES

Open Access: Notebooks, Data, Tutorials

Used by

- **Educators** – training in EO and Earth Science
- **Developers** – open-source to build new dashboards such as GTIF
- **Data Scientists** (4 competitions & hackathons e.g. NASA SpaceApps)
- **Nat. Statistical offices** – to introduce EO in their practice
- **Journalists** – to report on observed impacts e.g. of covid, air pollution, etc.

IAF special award on Space for Climate Protection

23 May 2023, Global Space Conference on Climate Change, Oslo, Norway

The **ESA-JAXA-NASA Earth Observing Dashboard** was chosen as: “*the most valid example of the ways in which remote sensing data can support climate protection and allow decision-makers, citizens and the scientific community to easily access information that may be fundamental to protect our planet.*”

NASA's evolution from trilateral dashboard

Contribution to trilateral dashboard:

- Datasets and interoperable services for data transformation, access, and visualization
- Science expertise
- Infrastructure and services
- Coordination and host for data science challenges (SpaceApps)
- Communication and outreach

NASA has utilized components developed for the trilateral dashboard (and other investments) to assemble a platform called VEDA (Visualization, Exploration, and Data Analysis), which:

- Leverages data in the cloud
- Provides interoperable data services for other priorities
- Lowers the barriers to computing platform that is next to data
- Enables advanced visualization and exploration with enriched science communication

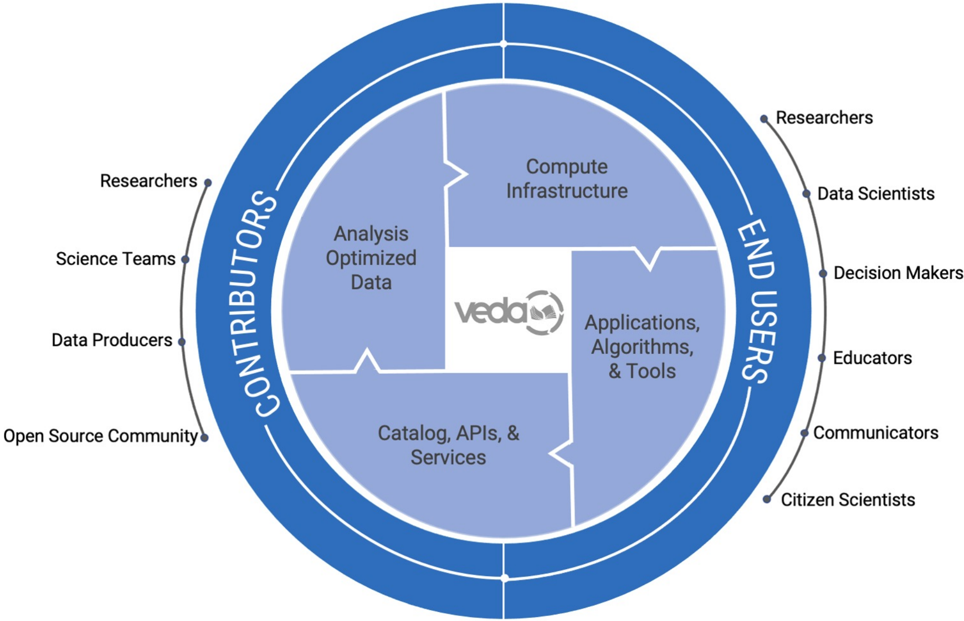
veda

Visualization Exploration & Data Analysis

Visualization, Exploration, and Data Analysis

Why?

- Interdisciplinary science depends on large amount of Earth science data and computational resources
- Working with these datasets is non-trivial
- Big data science requires advanced distributed computing knowledge

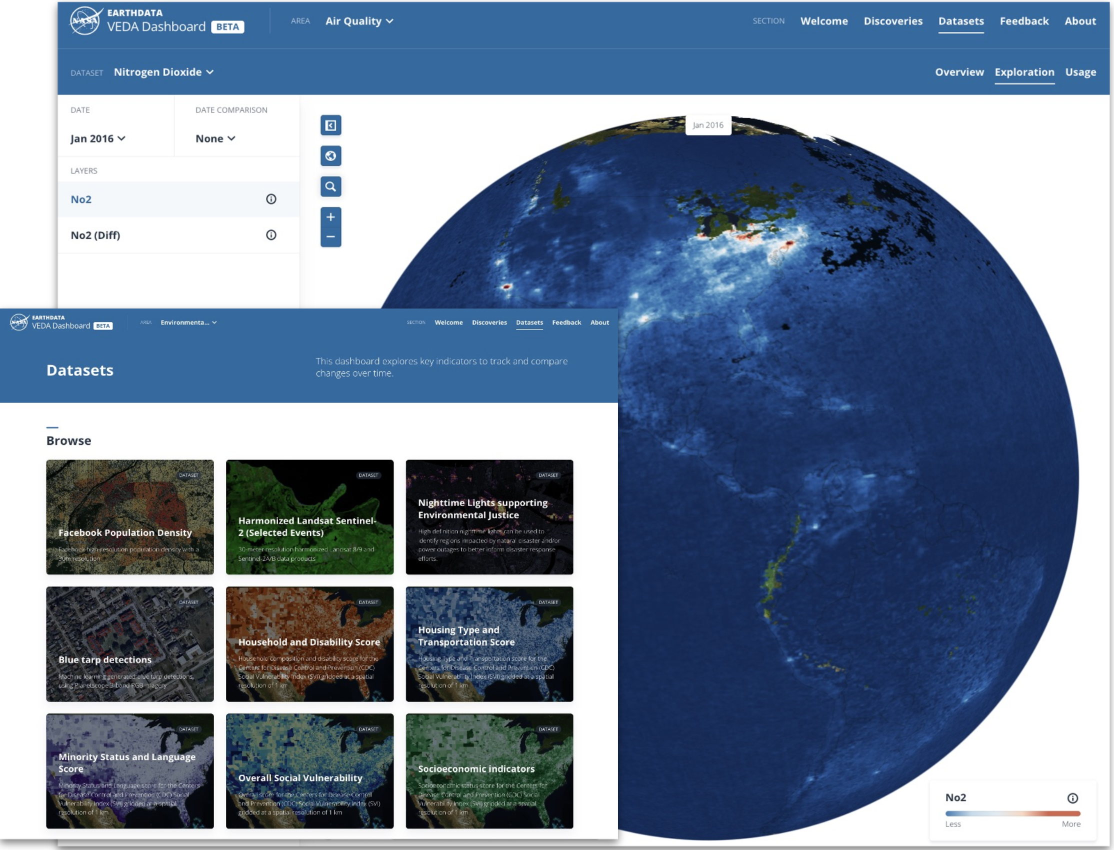


What?

VEDA is an open platform that brings **key Earth science datasets next to open source tools** for data processing, analysis, visualization, and exploration in a managed and **more accessible computing environment**

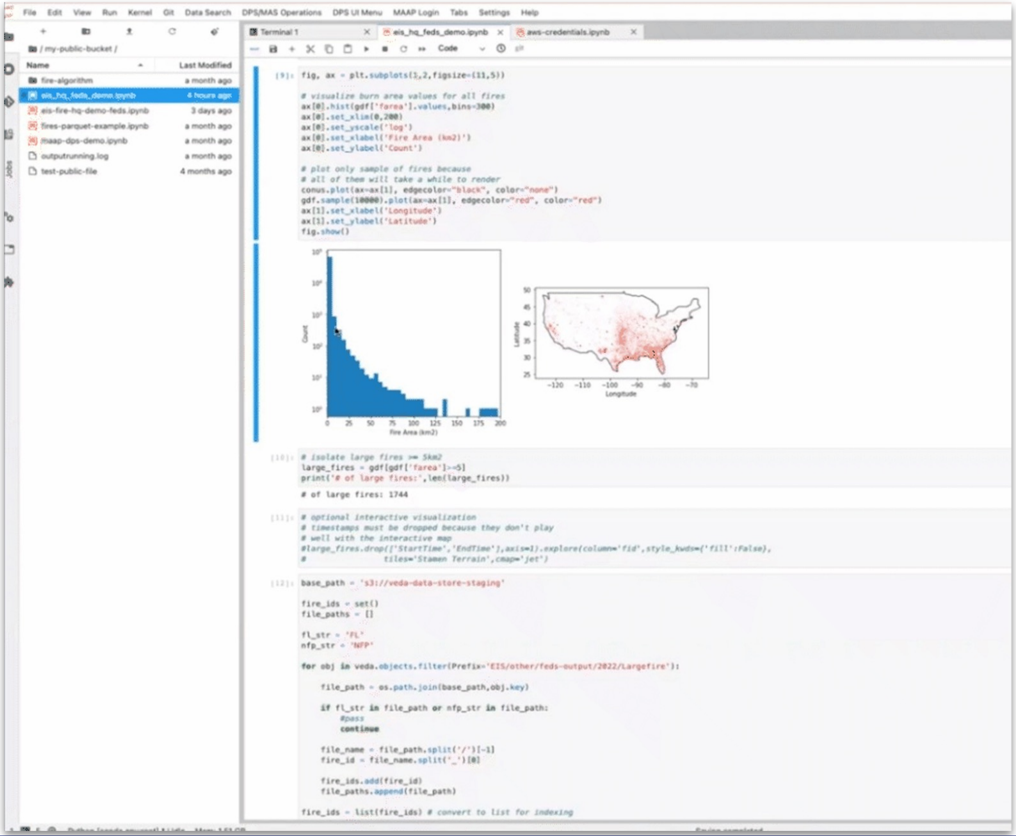
Built using existing investments in open source software

Explore | Analyze | Publish | Communicate



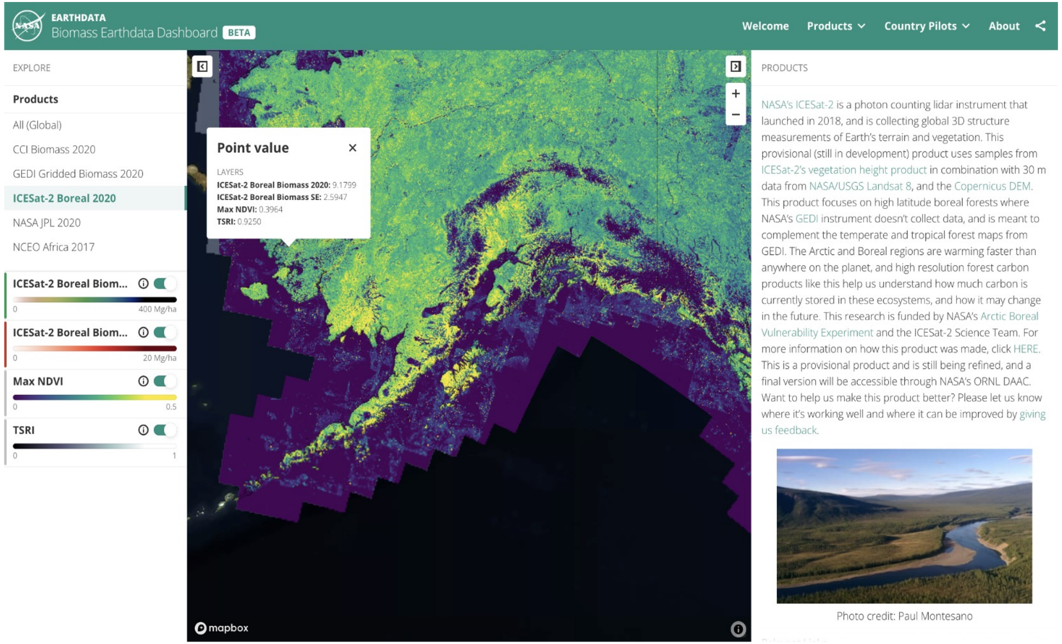
- Finding relevant data products
- Exploring data to identify interesting features

Explore | **Analyze** | Publish | Communicate



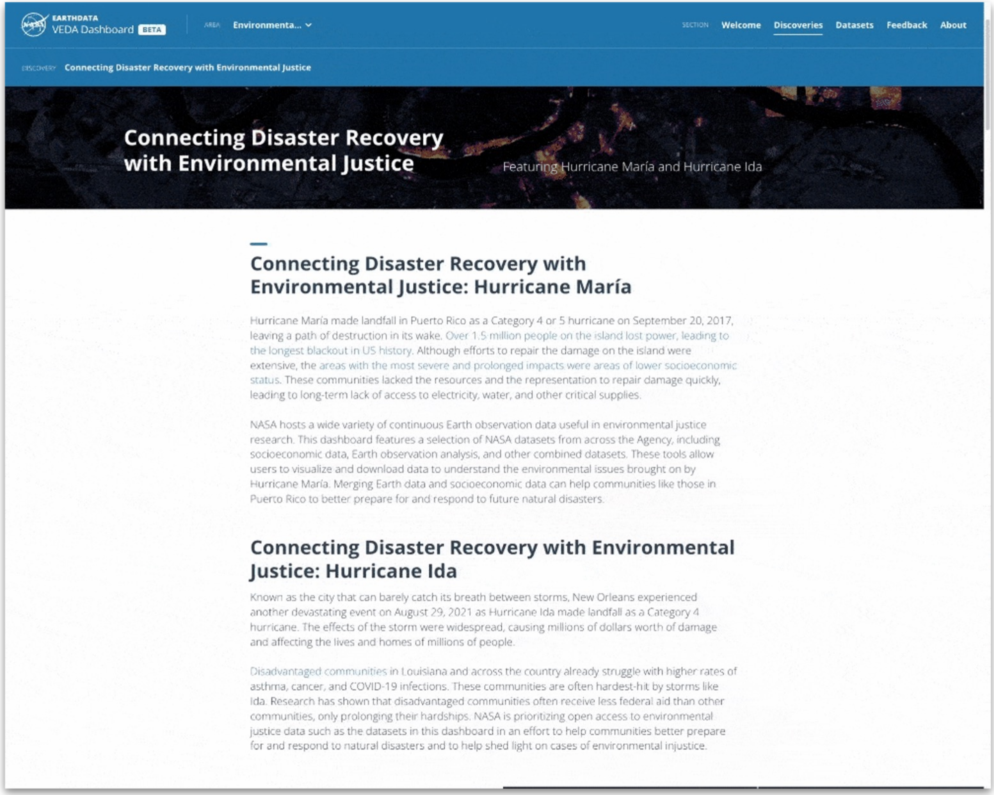
- Developing advanced data products and analysis
- Carrying out calculations "in place" without the need to download data
- Dynamically allocating resources for computationally demanding processing

Explore Analyze **Publish** Communicate



- Conveniently delivering data through existing interfaces
- Providing automatic access to interactive visualization capabilities
- Allowing users to analyze your products within the environment

Explore | Analyze | Publish | **Communicate**



- User friendly data-driven storytelling
- Enrich science and applications narratives with interactive exploration

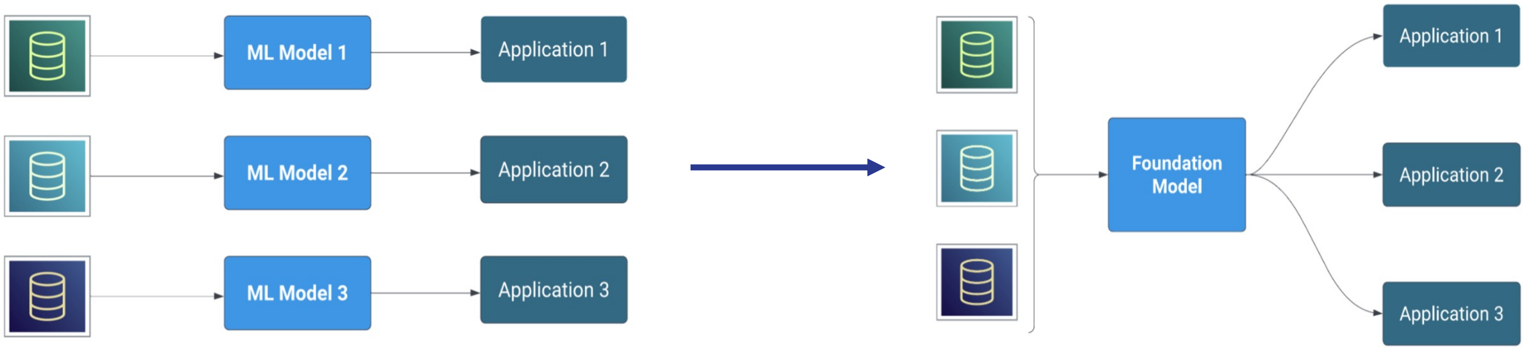
Capabilities supported by VEDA

- Earth Information System
- Environmental Justice Initiative
- US Greenhouse Gas Center
- NASA/ESA/JAXA Trilateral Dashboard

Geospatial AI Foundation Model

AI Foundation Models

- Pre-trained on a comprehensive dataset and used for various downstream tasks
- Substantially reduce the downstream effort for building AI applications, including the need for large labeled training datasets
- Model captures emergent behavior within the data representation
- Creating a foundation model includes data curation, training, validation, and large-scale computation



AI challenges in Earth science

Advancing Application of Machine Learning Tools for NASA's Earth Observation Data

Jan. 21-23, 2020 | Washington, D.C.
Workshop Report

Maskey et al. "Advancing AI for Earth Science: A Data Systems Perspective," AGU Eos 2020

- **Training data** is the main component of supervised machine learning techniques and is increasingly becoming the **main bottleneck to advance applications of machine learning** techniques in Earth science.

- Geoscience models must **generalize across space and time**; however, for supervised learning one needs large training datasets to build generalizable models.

Release of the First Geospatial AI Foundation Model

Collaboration with IBM Research under a Space Act Agreement and NASA IMPACT project at MSFC

Pretrained on NASA Harmonized Landsat Sentinel-2 dataset

Examples of how it can be used:

- Burn scar mapping
- Flood detection
- Multi-temporal crop identification

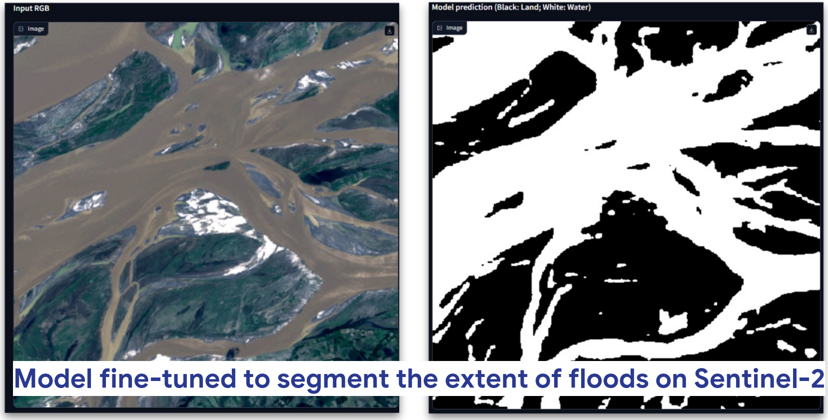
Openly available at Hugging Face including Models, Datasets, and Code.

*Released on **Aug 3rd, 2023***

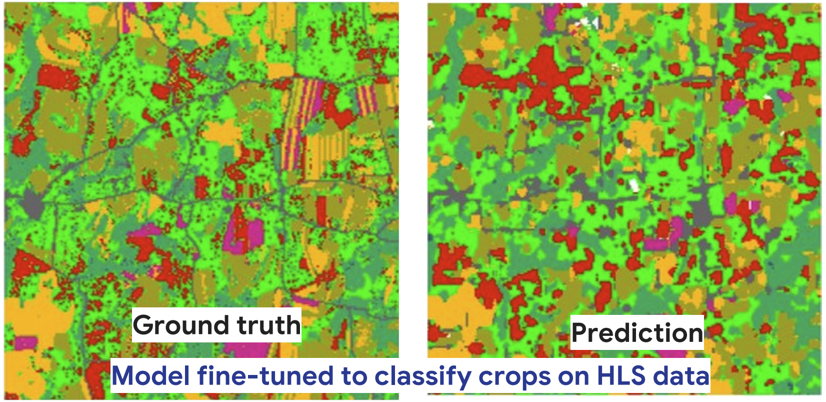
- *75 clones*
- *Several applications have already been built by community*
- *SpaceApps 2023 GeoAI challenge*

Collaboration with like minded partners IBM and Hugging Face

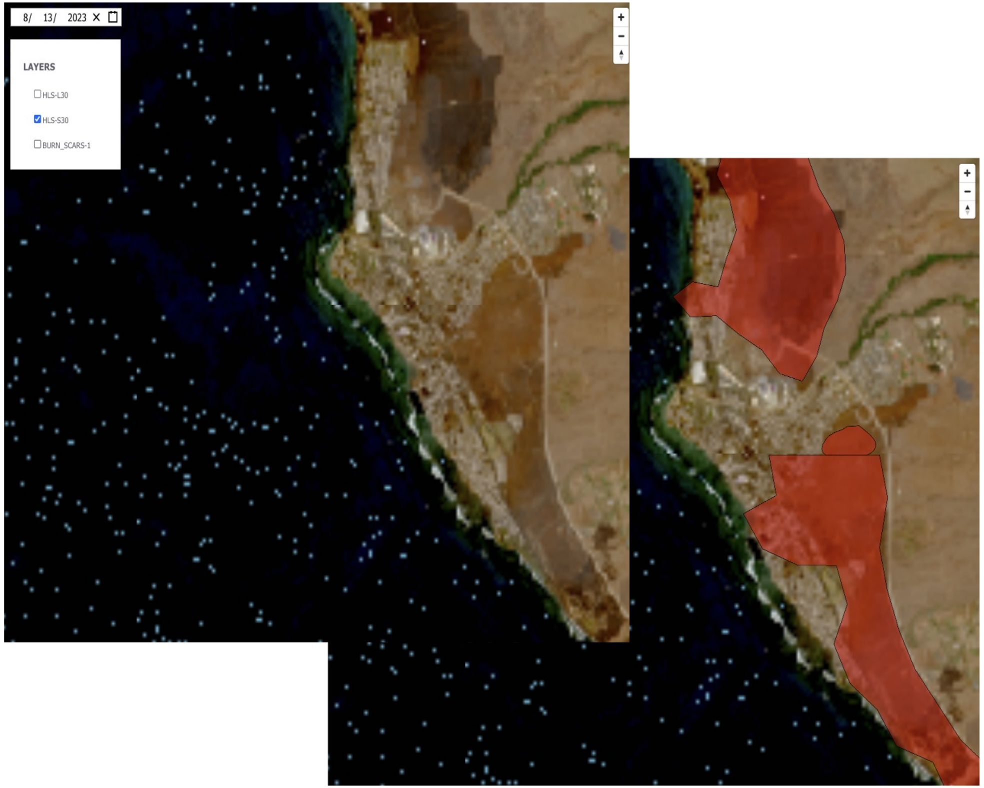
GeoFM : downstream applications



Model fine-tuned to segment the extent of floods on Sentinel-2

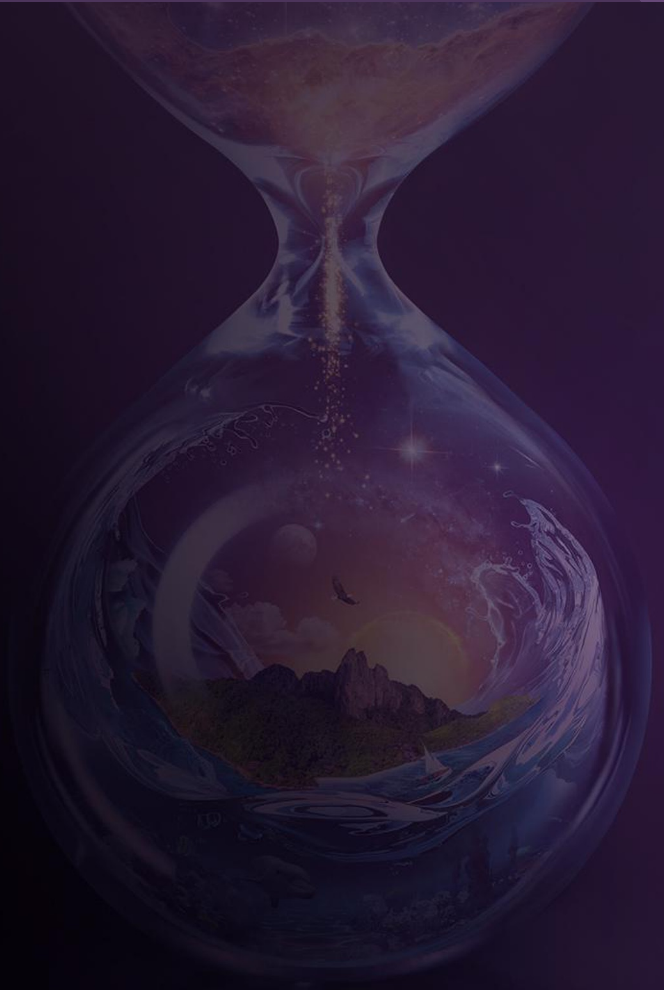


Model fine-tuned to classify crops on HLS data



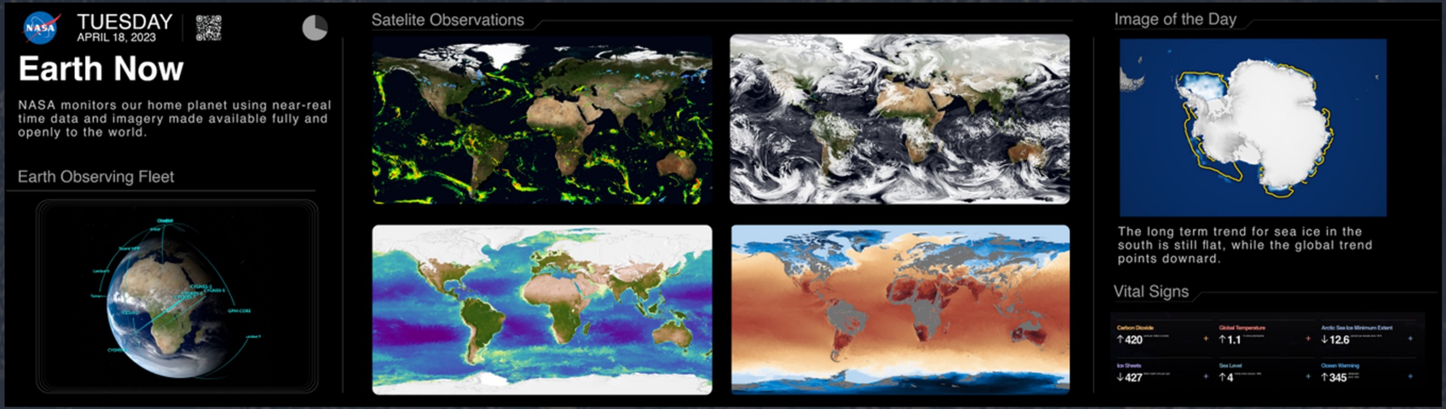
Burn scar mapping: Maui Fire 8.13.2023

Earth Information Center



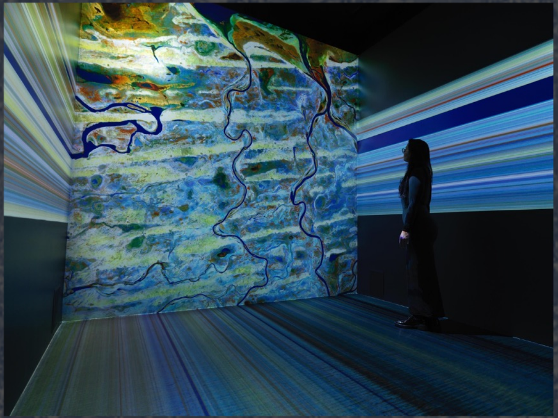
Earth Information Center

A physical and virtual space to engage and amplify impact – *to show people our Earth as we see it*. Visit the virtual site *at go.nasa.gov/eic*

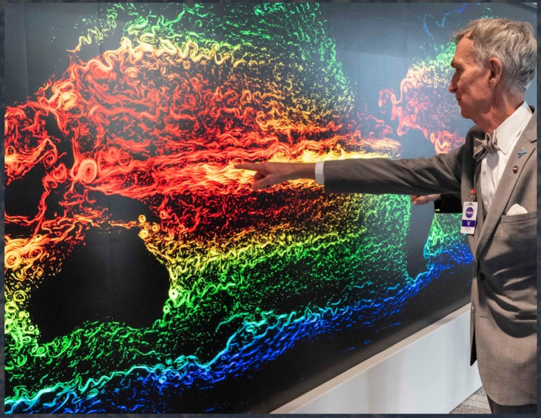


A Hyperwall to highlight real time NASA data and stakeholder stories

Earth Pulse display showing Near Space Network NRT data collection



An immersive installation to allow visitors to go inside the data



Key takeaways

- Open data and platform ease scientific research barriers
- Standards and interoperability enable wider collaboration and maximize data use
- Core set of services based on open-source software offer agility needed for priority initiatives/projects
- Strategic partnerships accelerate research, especially through access to computation and technical expertise
- Community involvement enhances equitable and impactful scientific future

Contact

Iksha Gurung, ig0004@uah.edu