CICompass

Funded by the National Science Foundation, Grant #2127548

NEON's first data releases: challenges and opportunities to provide FAIR data

By Christine Laney, Steve Jacobs, Dawn Lenz, and Claire Lunch

Making the Major Facilities Data Lifecycle FAIR to provide AI-Ready Data

March 1, 2022

CI Compass Cyberinfrastructure for NSF Major Facilities Workshop

USC Viterbi School of Engineering Information Sciences Institute
INDIANA UNIVERSITY
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
TEXAS TECH UNIVERSITY
renci

Christine Laney, Steve Jacobs, Dawn Lenz, Claire Lunch
March 1, 2022

NEON's first data releases: challenges and opportunities to provide FAIR data

1 This material is based upon work supported by NSF's National Ecological Observatory Network which is a major facility fully funded by the National Science Foundation

About NEON

- Funded by US NSF, operated by Battelle (non-profit)
- 81 freshwater and terrestrial sites
- > 4B data points/day, currently transforming into 165 data products that are freely available to the public (17 more being developed), ranging from
  - genomics (organismal and environmental)
  - population dynamics
  - surface-atmosphere fluxes
  - remote sensing, and more
- ~1.5 PB data currently in storage + ~400 TB/yr

>100,000 samples sent to ASU Biorepository per year

Data Types

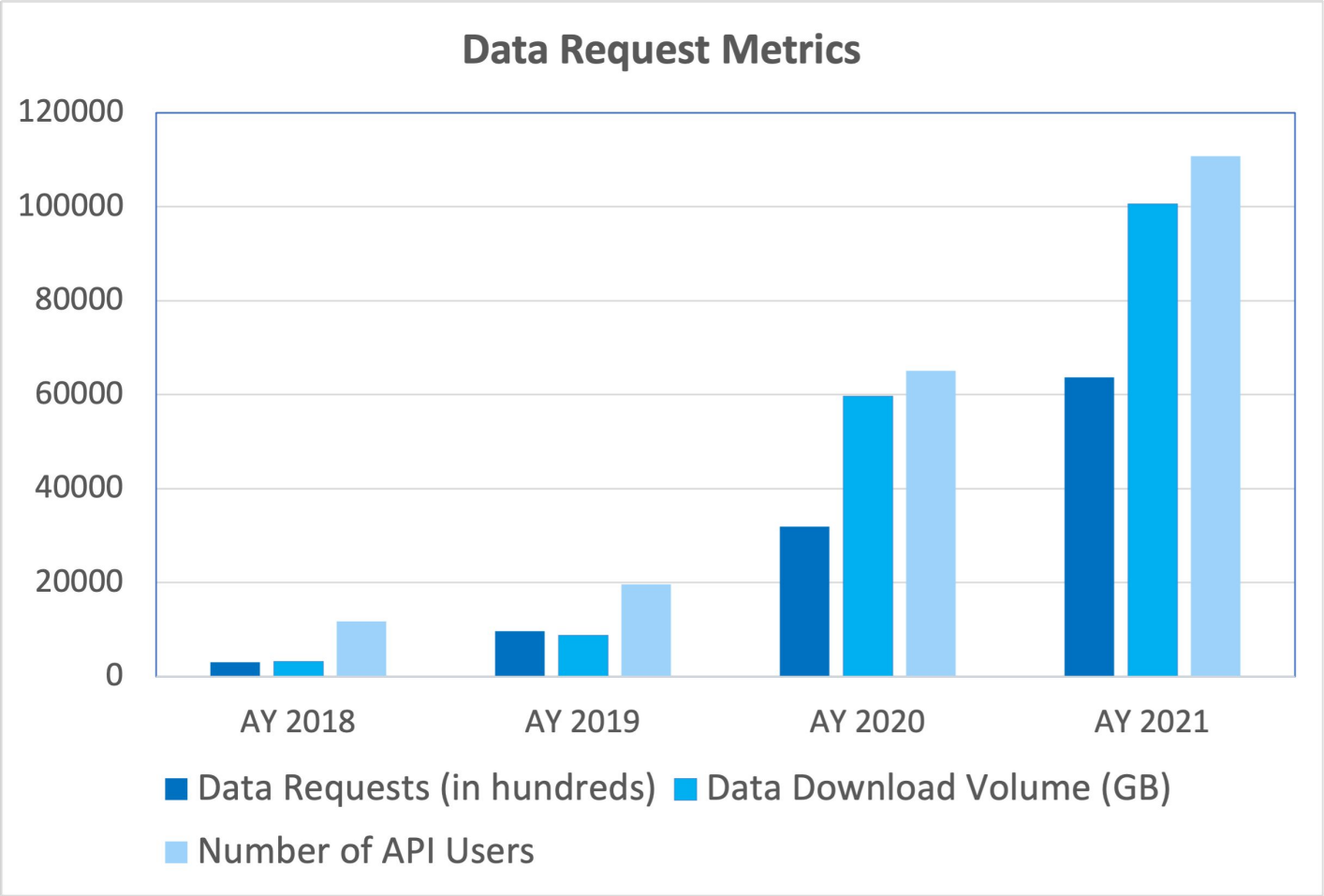
Three major divisions of data as viewed by processing pipelines

- Sensor data streamed into headquarters
  - Raw data points as frequent as 40Hz, published data as frequent as 1 minute.
- Observational data
  - Manual data or sample collection, can be multiple times per year or (rarely) once per multiple years
- Remote sensing
  - Flight season runs March – October. Hard drives are shipped back to HQ for processing

NSF neon Operated by Battelle

The Findable & Accessible

- Data are findable and freely accessible via Data Portal GUI and multiple endpoints of NEON's API
  - Also findable via other third party aggregators, e.g., GBIF, BOLD, Ameriflux, DataONE
- The success of the F & A is demonstrated in the increasing number of data requests, download volumes, and number of API "users" (unique IP addresses)

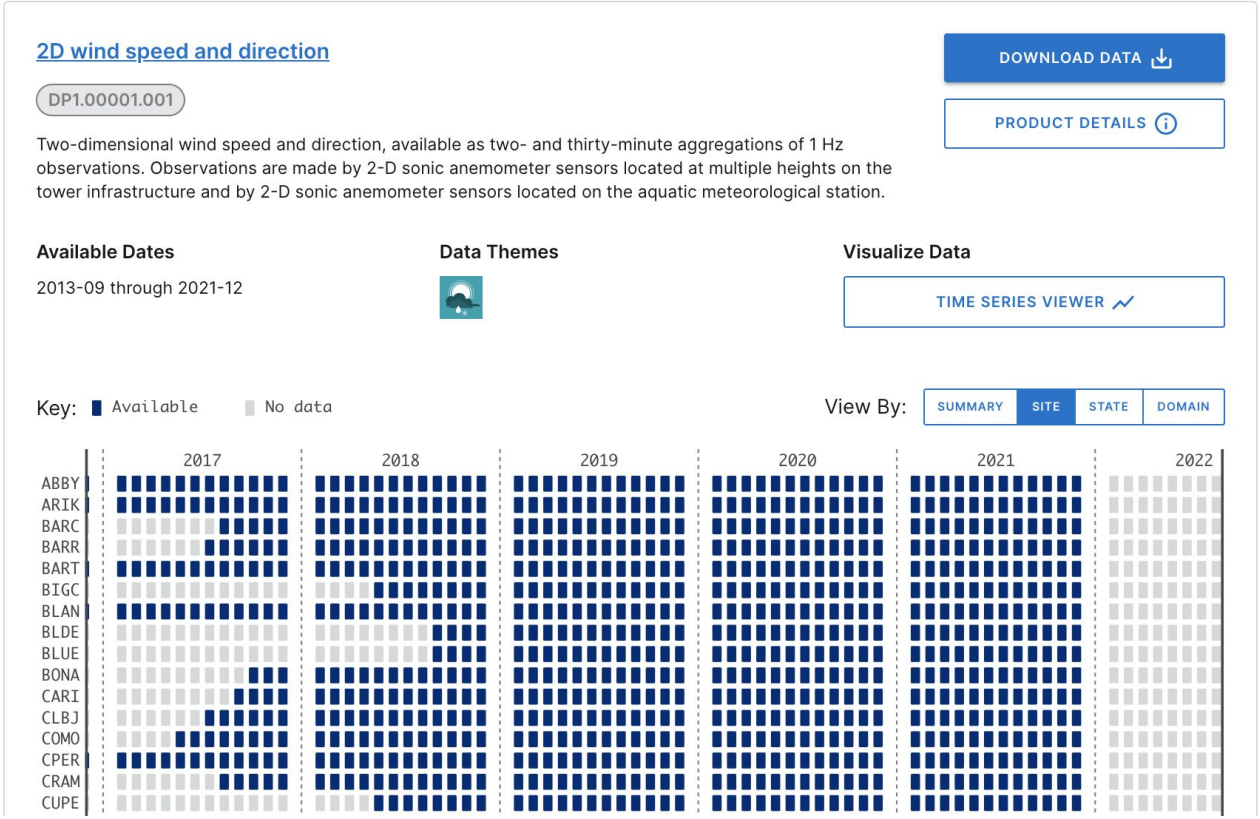


The R – Reusable (and Reproducible and Citable)

- Getting data out quickly is important, as is data quality
- First, provide Provisional data - during Construction & Initial Ops, then rolling
- **NEON's 2021 data release** was the first reusable, reproducible set of data products available to the public and represented a minimum of 1 year of data for most data products.
- DOIs make data products fully citable
- Creating a Release takes months of cross-team coordination

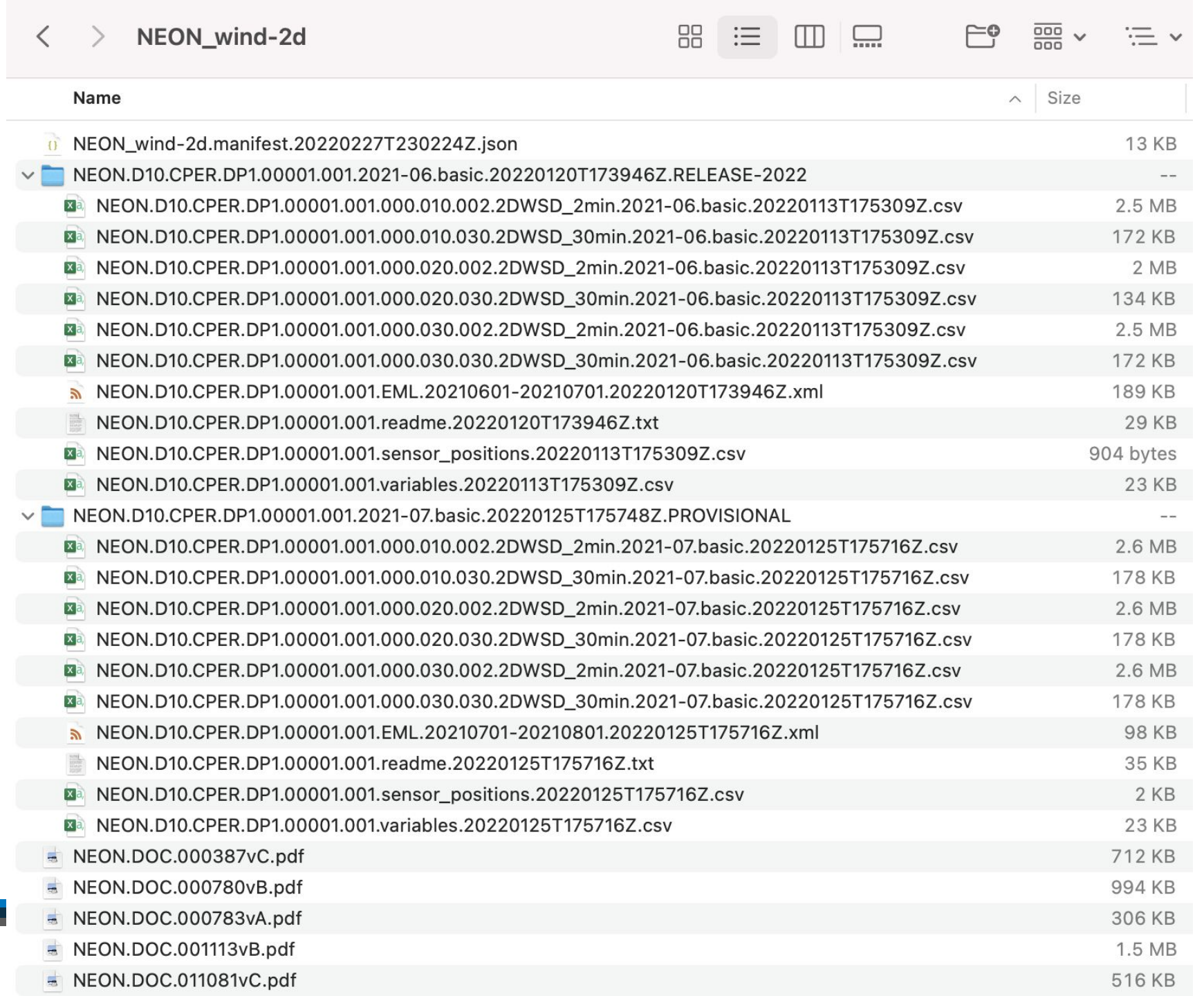
Data Granularity

- Stored files are indexed; the database stores metadata about each file.
- Files are clustered by Data Product, Field Site, and Year Month. Each DP x S x M is documented with a readme and an EML (XML) file.

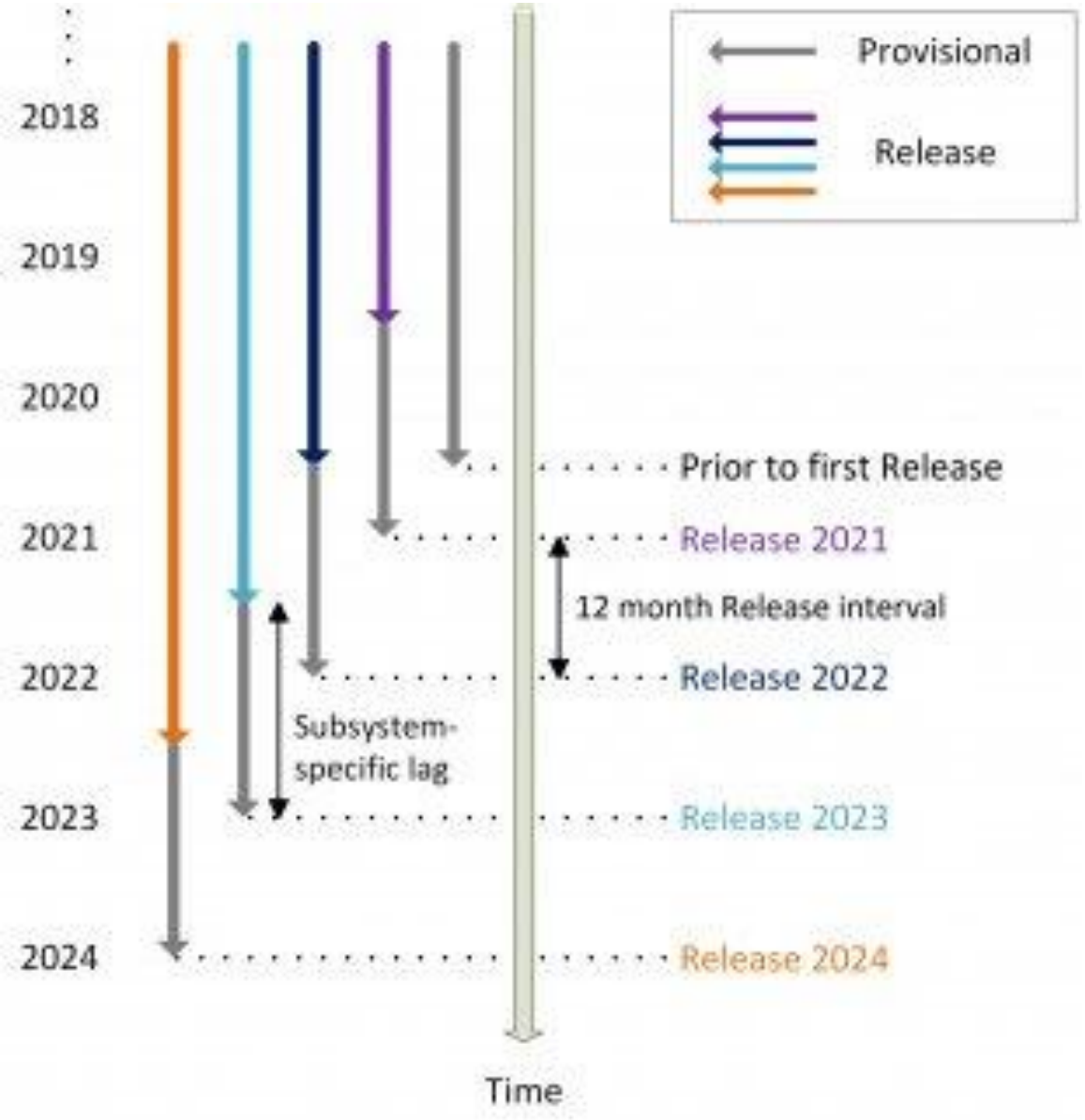


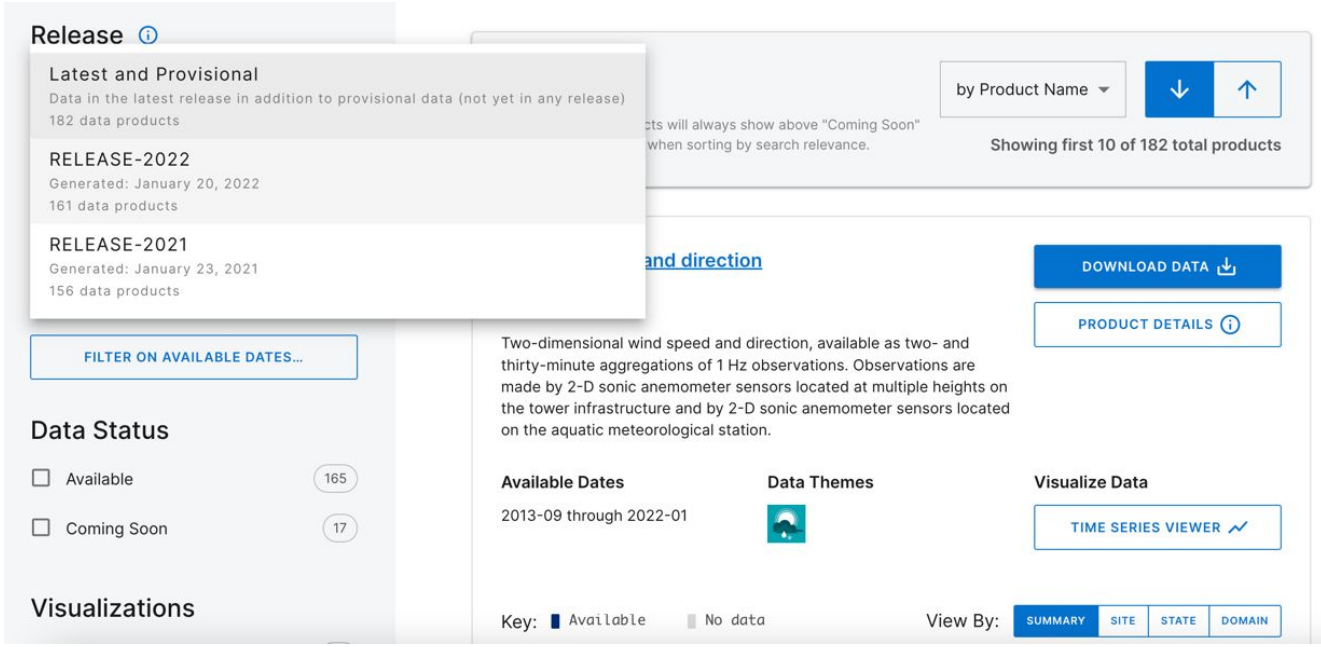
Example of downloaded monthly packages

- 2D Wind Speed and Direction
- June-July 2021, at site CPER
- 1 month of data tagged as RELEASE-2022, 1 month data tagged as Provisional.
- The included manifest (as a JSON file) shows the structure of these packages within the releases.



Data Release Mechanics and Discovery





Striving for R in exceptional cases

- Remote Sensing (AOP) data products: only the most recent release is available due to storage constraints.
- Samples: IGSNs and updated DarwinCore records – these are managed by NEON's Biorepository
- Provisional data: Partnership with the Environmental Data Initiative (EDI) – DOIs on provisional data as included in a submitted workflow including outputs.

Documentation of Changes Between Releases

Changes between RELEASE-2021 and RELEASE-2022

DATA PRODUCT ID	DATA PRODUCT NAME	SUMMARY OF CHANGES BETWEEN RELEASE-2021 AND RELEASE-2022
DP1.00001.001	2D wind speed and direction	All data were reprocessed with an updated algorithm to resolve errors in the computation of wind direction statistics (mean, variance, expanded uncertainty, and standard error of the mean) and wind speed expanded uncertainty. Sensor orientations and resultant wind directions were corrected for select sites and time periods. See data product issue log for details.
DP1.20059.001	Windspeed and direction above water on-buoy	This is the initial release of this data product. Data from 2020 and on have been withheld from RELEASE-2022 as NEON works to correct wind direction data.
DP1.20288.001	Water quality	Data were reprocessed using an updated algorithm to add seaLevelDissolvedOxygenSat (dissolved oxygen saturation corrected to sea level pressure) and localDissolvedOxygenSat (dissolved oxygen saturation as measured at local pressure) and remove dissolvedOxygenSaturation, which was the same as the new seaLevelDissolvedOxygenSat. Additional minor changes and/or corrections were made to quality control thresholds, sensor installation periods, and manual quality flags.

https://www.neonscience.org/data-samples/data-management/data-revisions-releases/release-2022

Conclusion

Successes

- Achieving R for the bulk of NEON's data products
- Reuse is intertwined with Reproducibility

Challenges

- (As many have noted before) The R of FAIR is difficult to achieve
- Provisional data still needs to be preserved by users

Thank you

Operated by Battelle

720.746.4844 | neonscience@battelleecology.org | neonscience.org