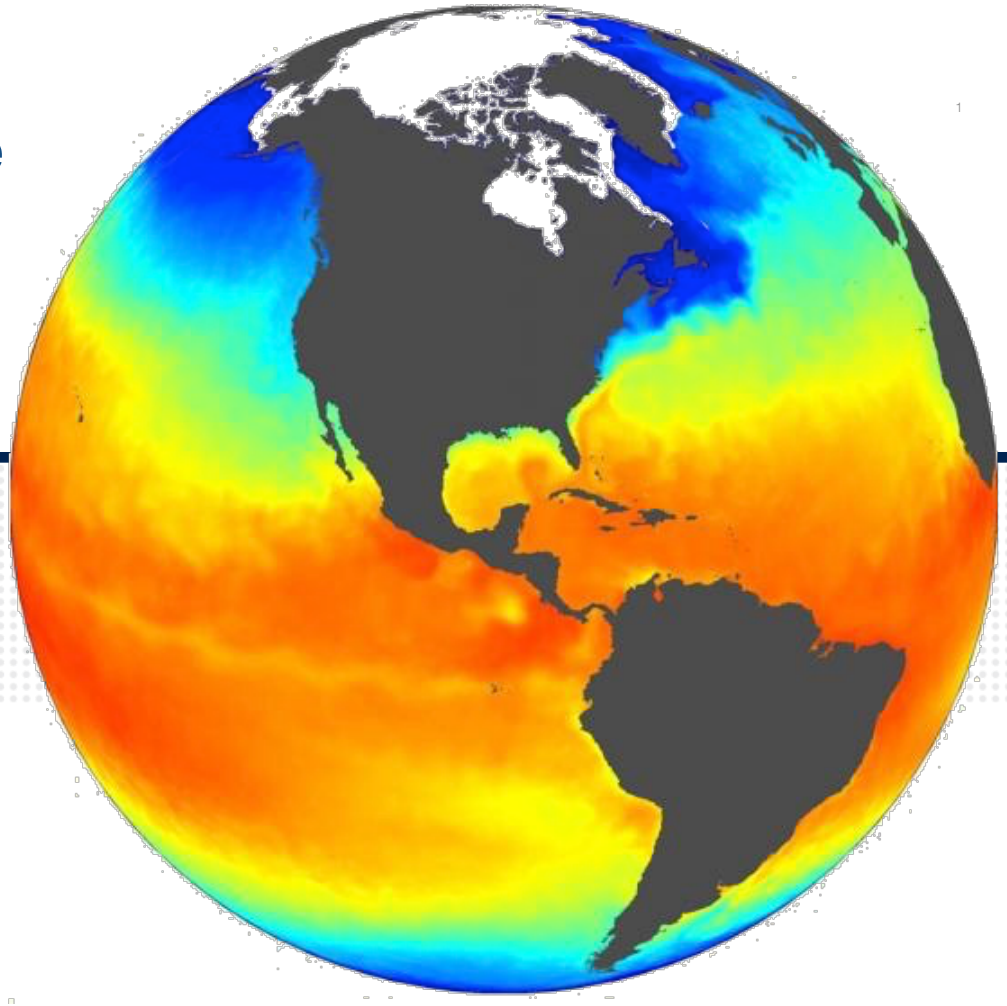
HDF5 – A Foundation for Sustainable Earth Science Data and Software

Ted Habermann
Director of Earth Science and Community Development
thabermann@hdfgroup.org



The HDF Group

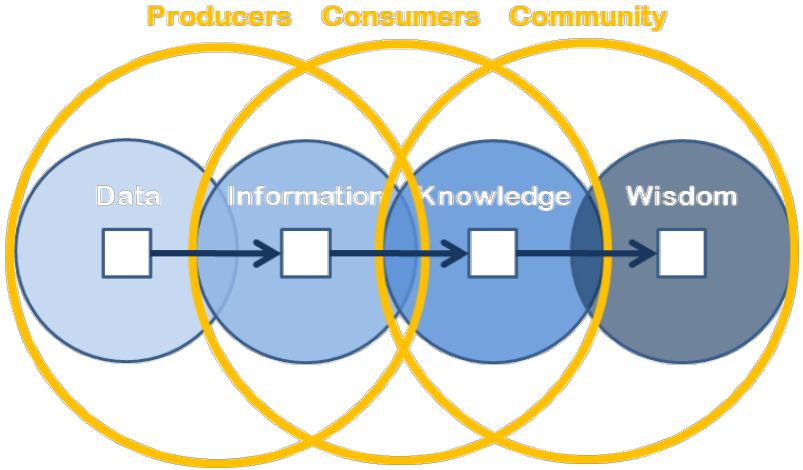
HDF Mission

Mission:

To ensure the sustainable development of HDF (Hierarchical Data Format) technologies and the ongoing accessibility of HDF-stored information.

One Goal:

To ensure long-term availability and support for HDF technologies, and by extension, long-term accessibility of information stored using HDF technologies.



HDF5 Philosophy

- A single platform with multiple uses
- One general data model
- One general format
- One library
- Adaptable for almost any kind of data
- Works on almost any architecture
- Supports compression and chunking
- Supports parallel I/O
- Ability to interact well with other technologies (3rd party filters, APIs etc)
- Attention to past, present, future compatibility

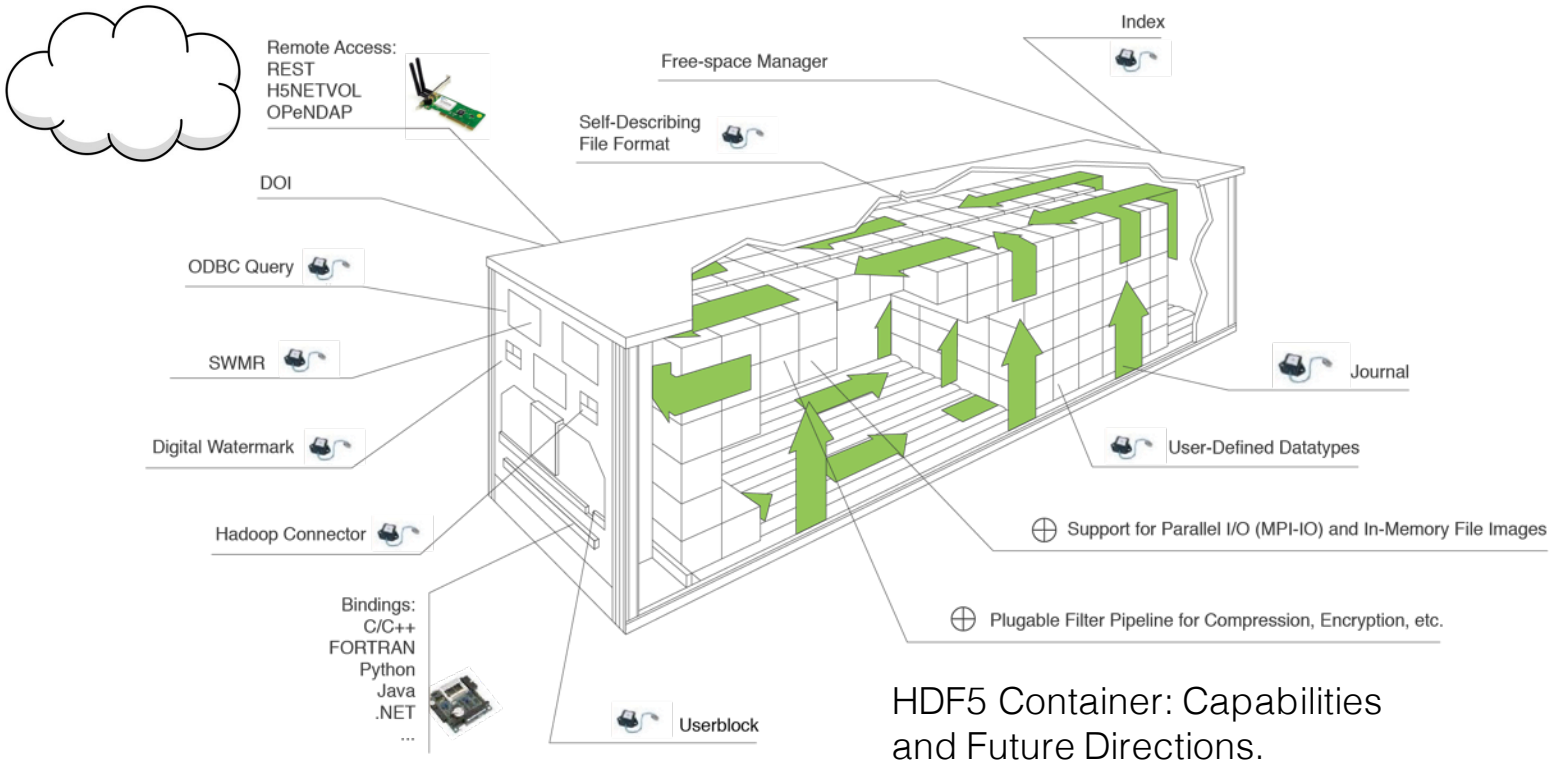




Conventions – Domain Data Types

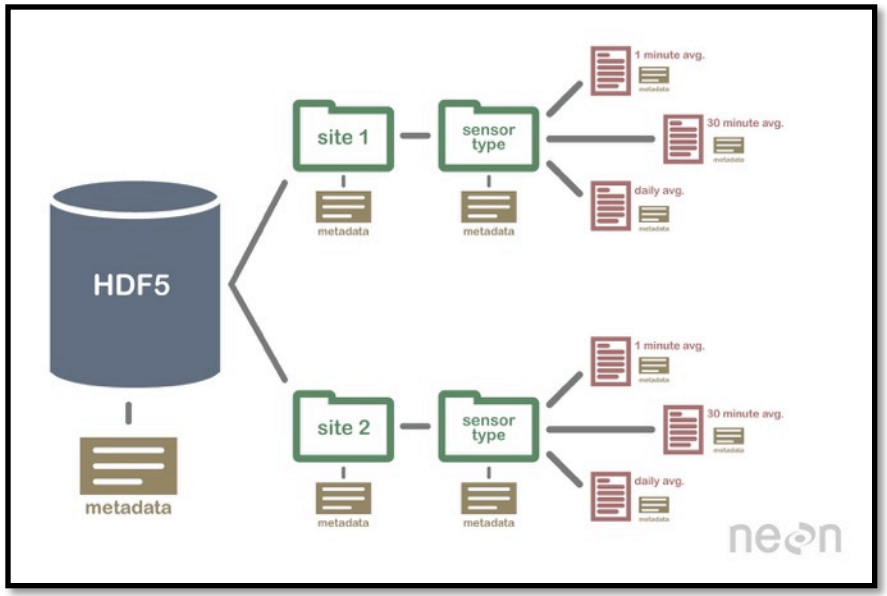
HDF – Computer Science Data Types

HDF – Smart Containers for Earth Science Information

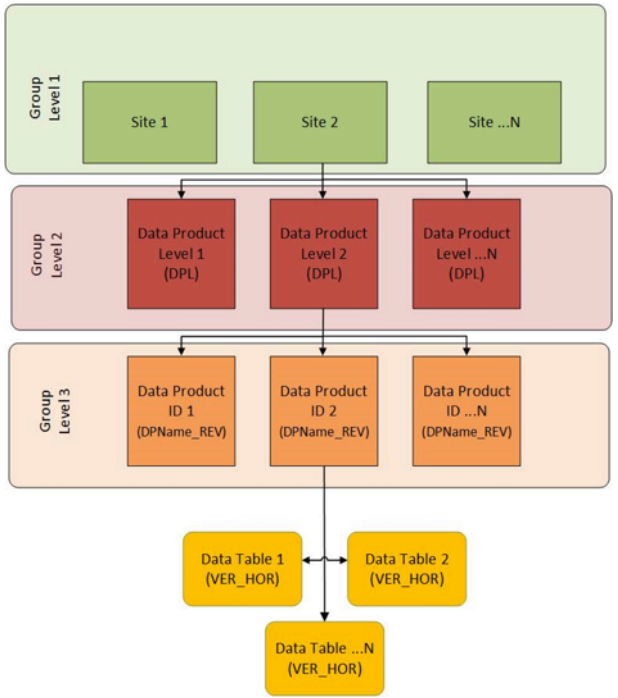


HDF5 Container: Capabilities and Future Directions.

NEON HDF5 file structure

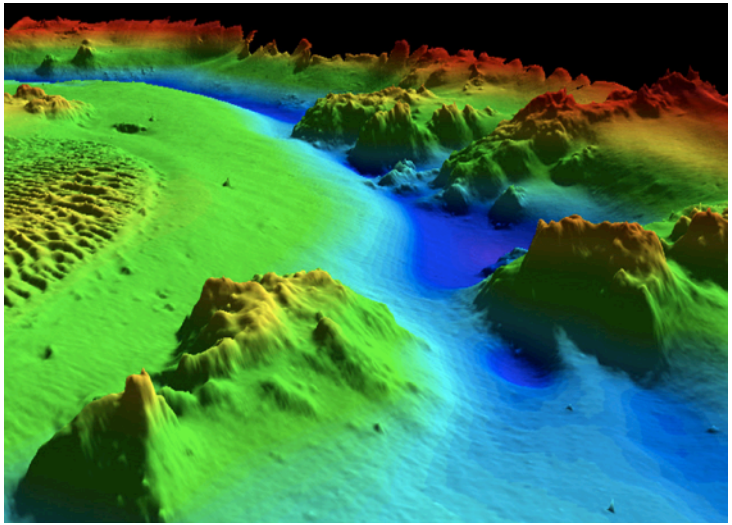


http://www.neonscience.org



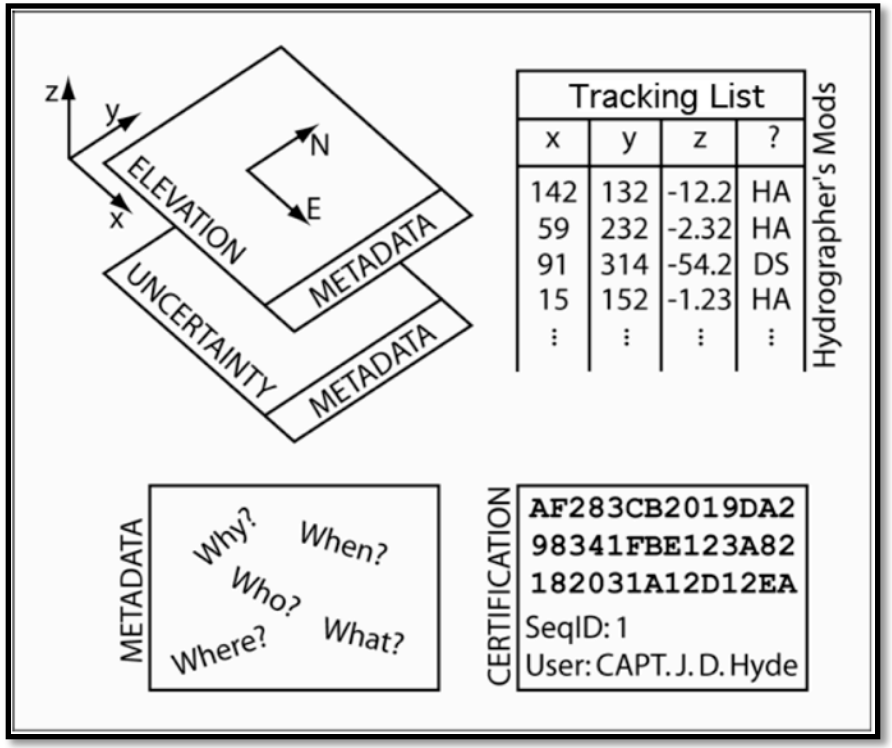
Eddy-covariance Data

Open Navigation Surface (BAG – IHO)



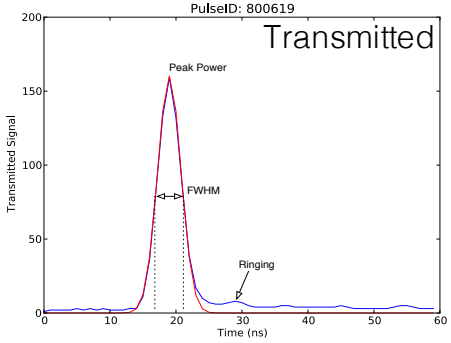
Open Navigation Surface

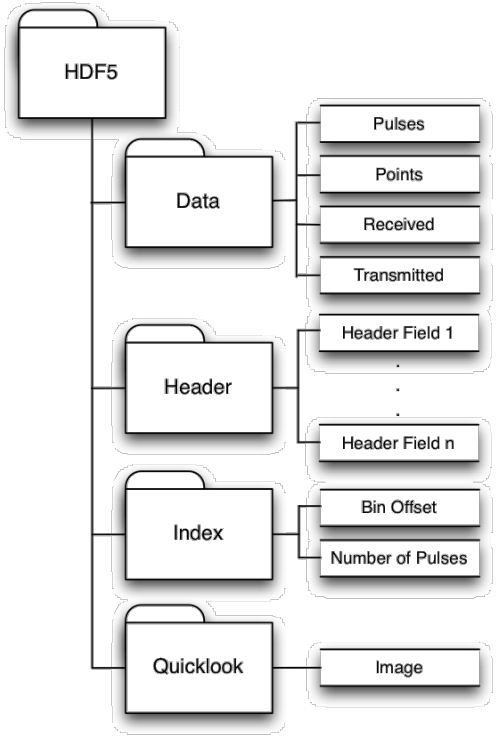
http://www.opennavsurf.org

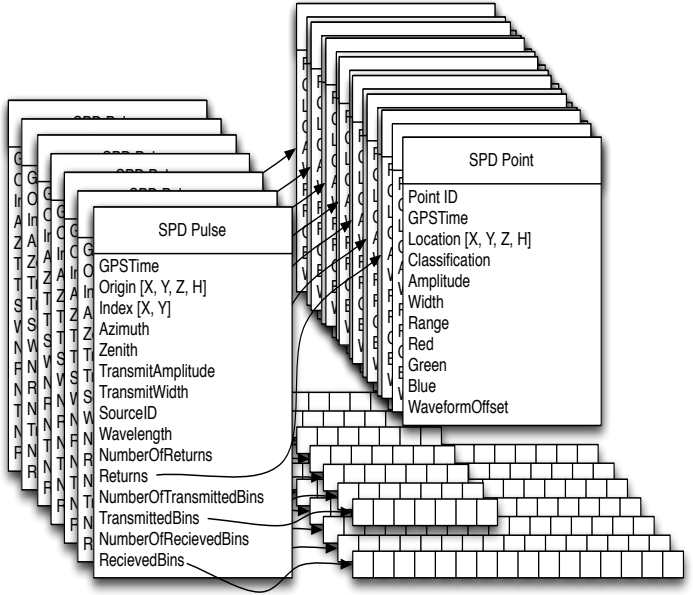


https://www.iho.int/mtg_docs/com_wg/HSSC/HSSC3/Draft_S-102_Ed1.0.0.pdf

Sorted Pulse Format (Lidar etc.)



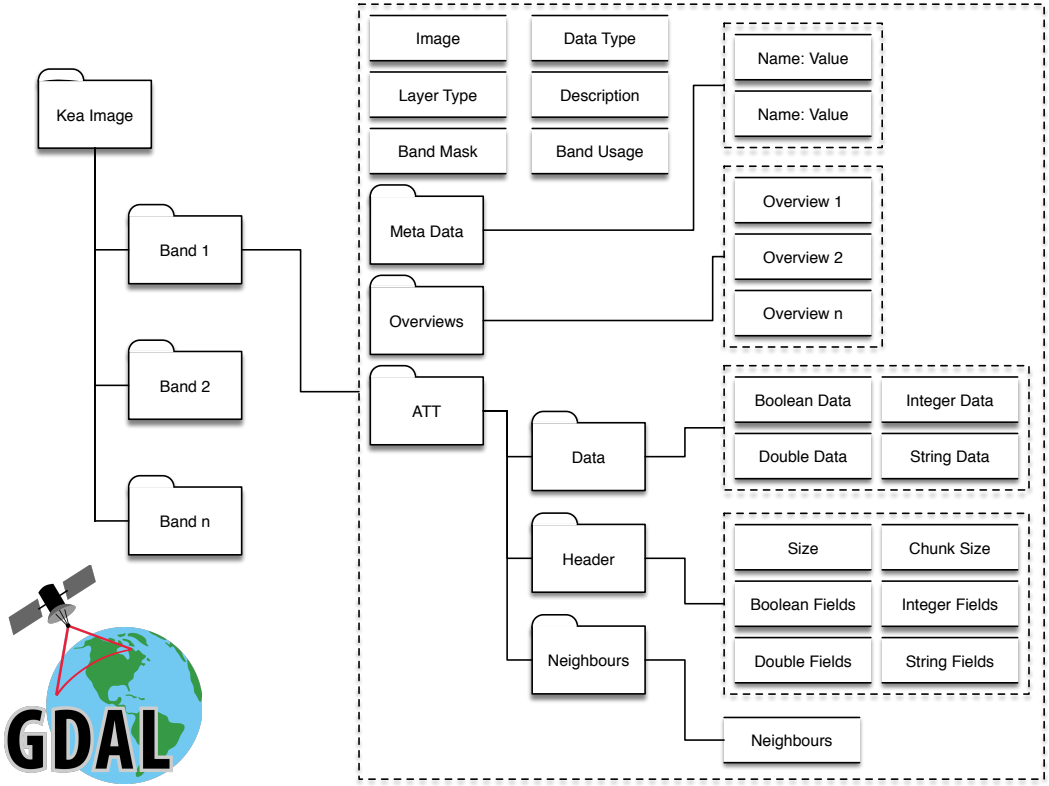


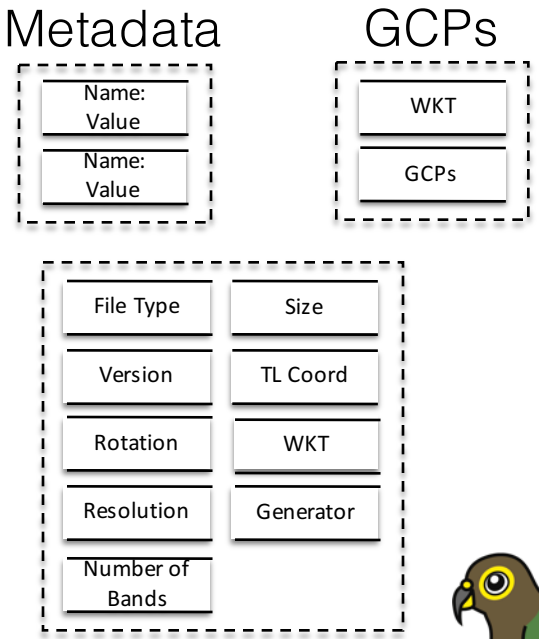


[SPD Documentation]

Kea – Land Cover (Images and Pixel Attributes)

Problem: How to have large attribute tables of data alongside raster data

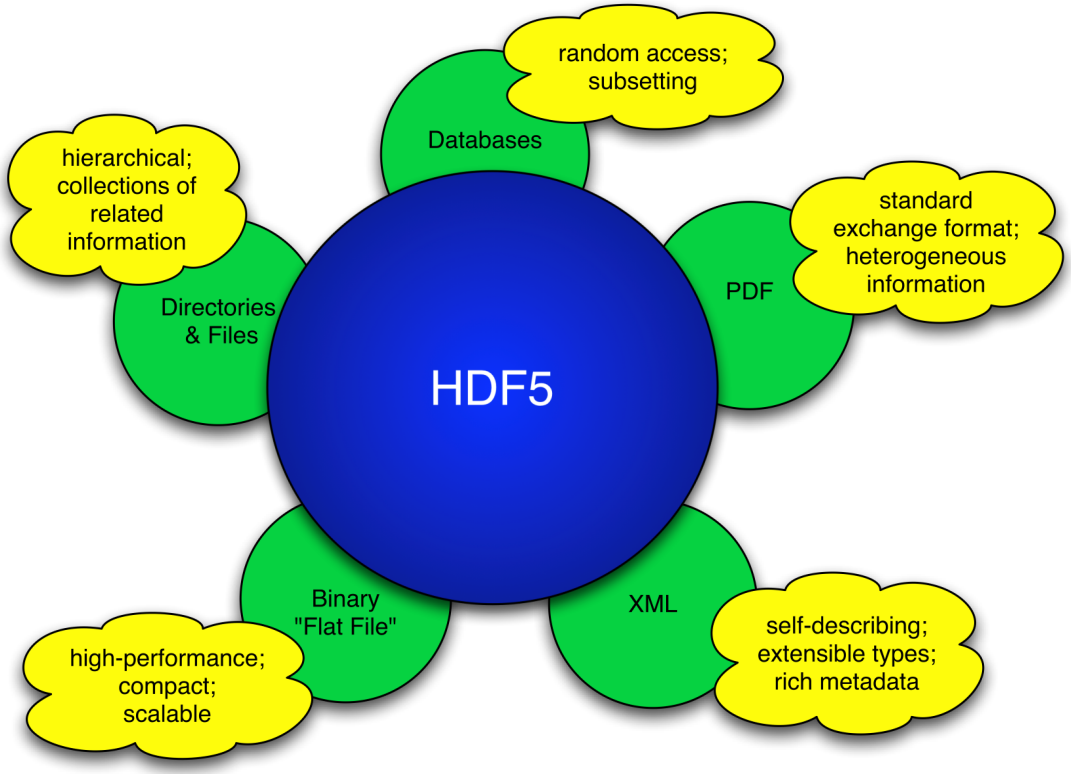




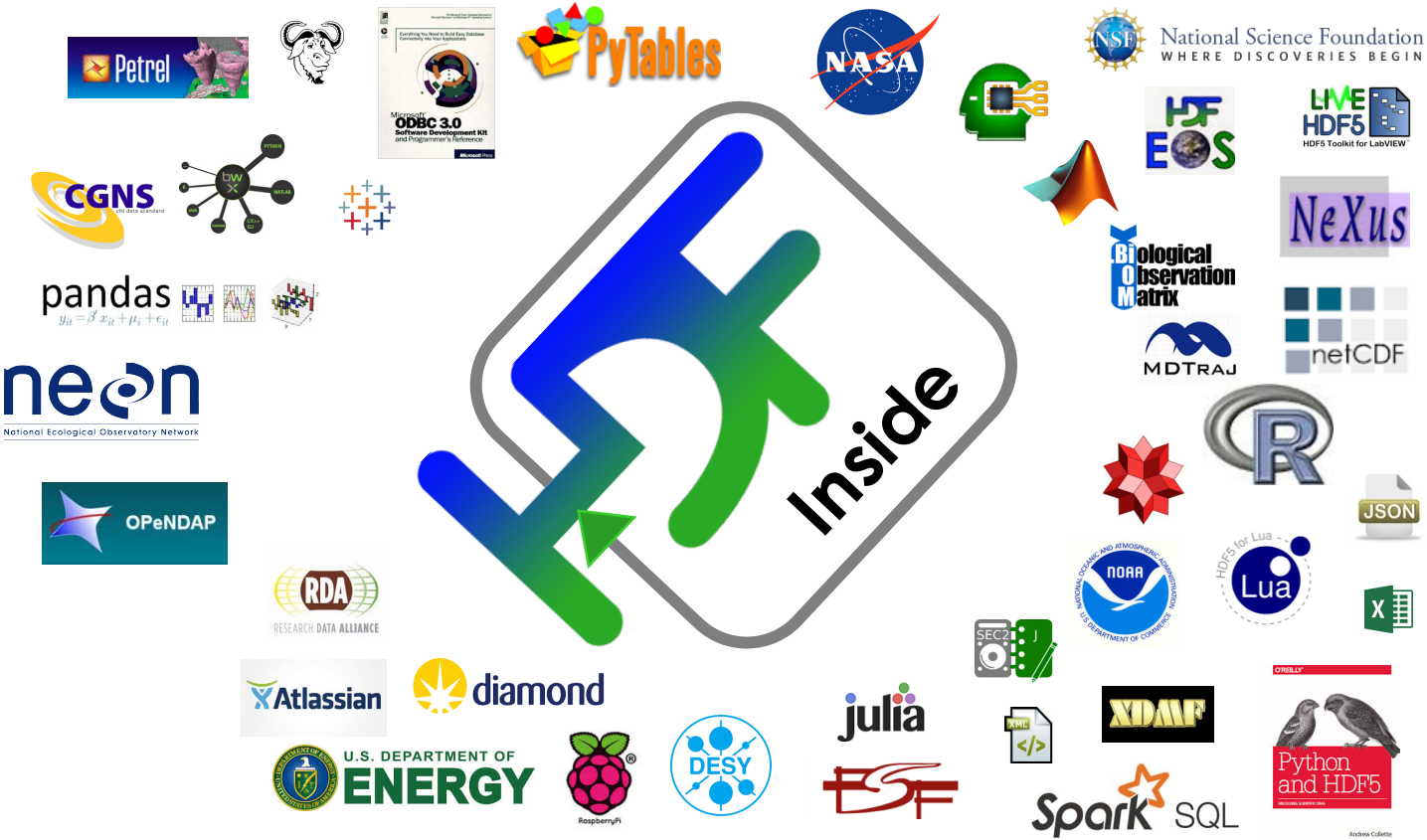


http://kealib.org

HDF5 is like …


HDF5
Databases
random access;
subsetting
Directories
& Files
hierarchical;
collections of
related
information
PDF
standard
exchange format;
heterogeneous
information
Binary
"Flat File"
high-performance;
compact;
scalable
XML
self-describing;
extensible types;
rich metadata

HDF5 Community



Source: http://www.yourarticlelibrary.com/ecosystems/ecosystem-useful-notes-on-our-ecosystem-with-diagram/9878/

Justification (ASDF)

- **Efficiency –** Data storage is cheap but operations on the data are quickly becoming the limiting factor. More efficient and performant data processing and analysis tools are badly needed.
- **Data Organization** – Different types of data are needed for a variety of tasks. This results in ad hoc data organizations and formats that are hard to maintain, integrate, reproduce, and exchange.
- **Data Exchange** – We need to exchange complex data sets.
- **Reproducibility** – Oftentimes not existent and should be strongly encouraged.
- **Mining, Visualization and Analysis of Data** – As data volumnes growand data become more complex, new techniques to query and visualize large datasets are needed.
- **Understanding and Trusting Data –** Detailed documentation standardized for specific domain data types is required for data understanding, trust and reuse.

Useful resources and tutorials for new users

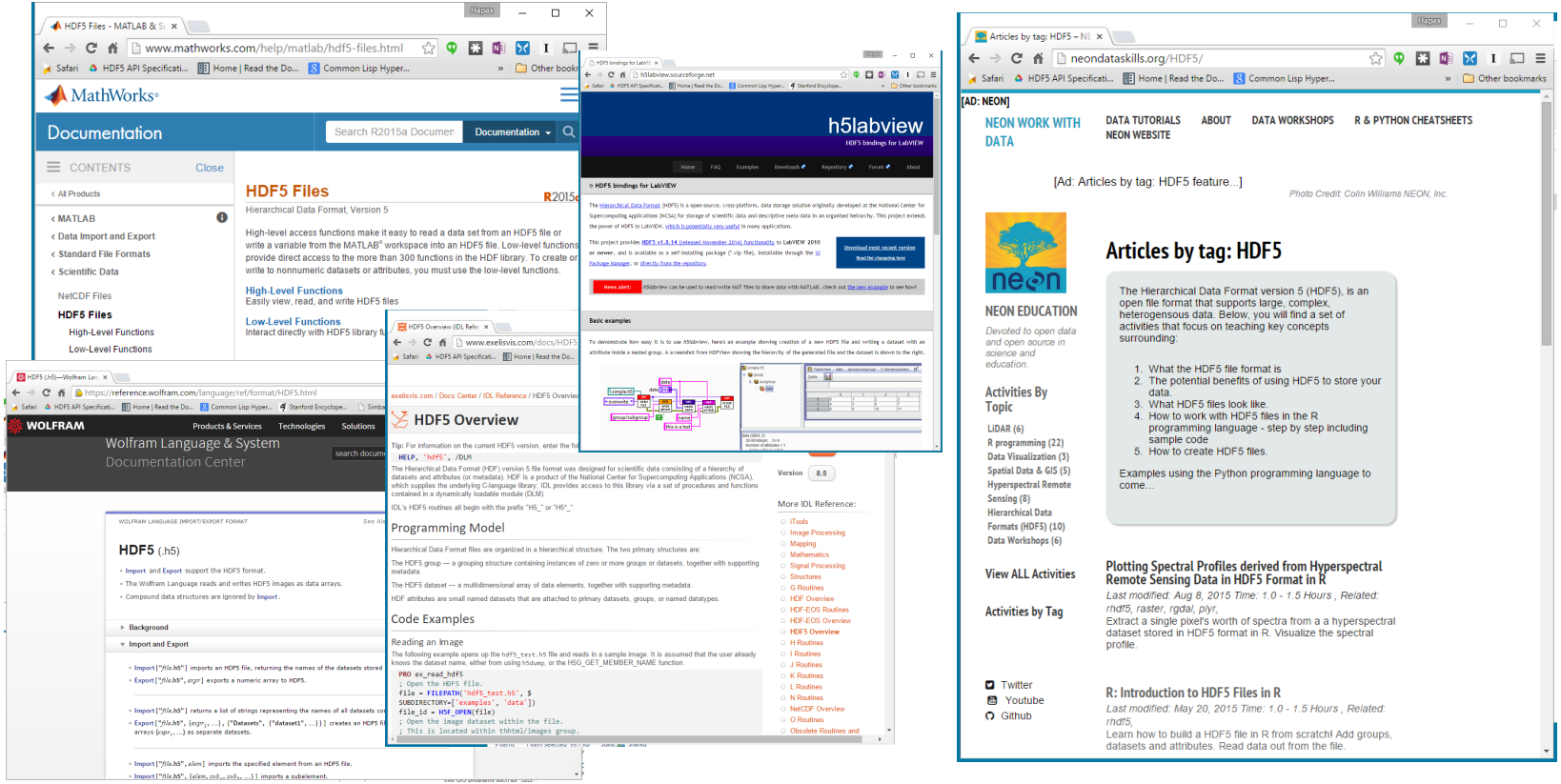
HDF Group Tutorials https://www.hdfgroup.org/HDF5/Tutor/

NEON Tutorials http://neondataskills.org/HDF5/

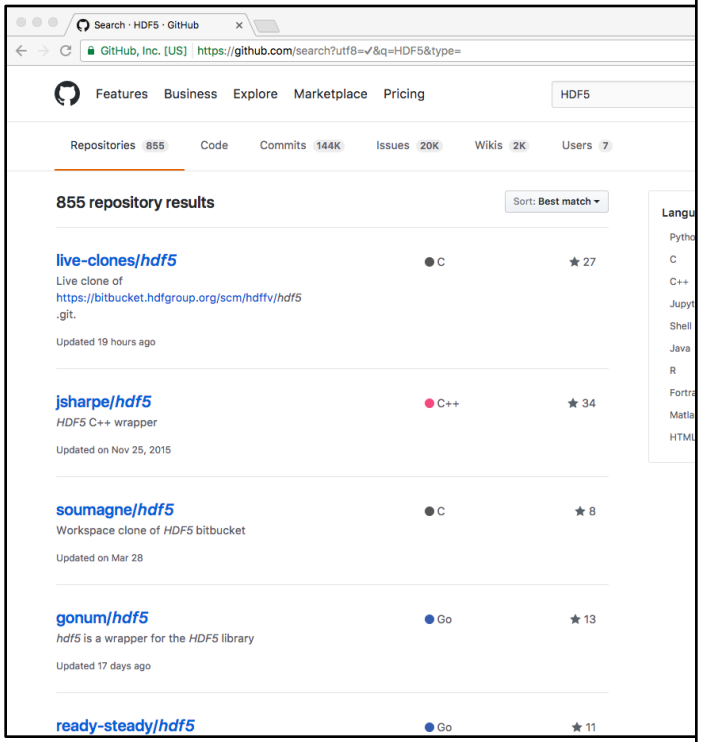
HDF5 is for lovers tutorial https://vimeo.com/63260833

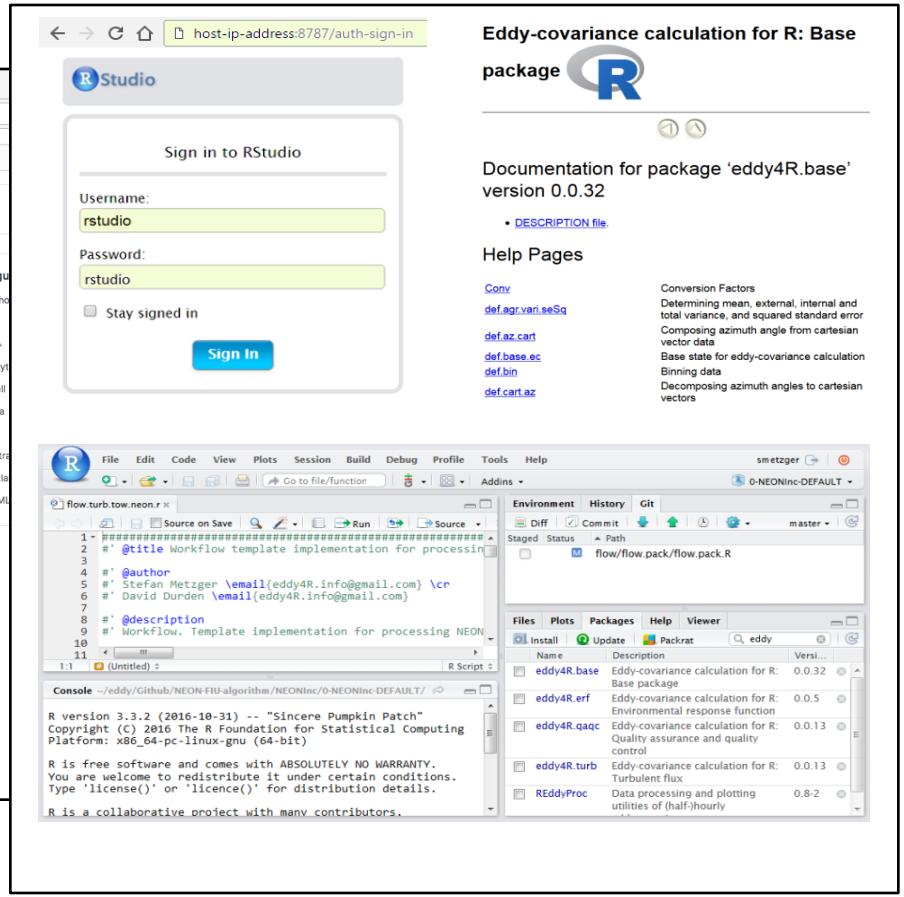
HDF5 is for lovers GitHub https://github.com/scopatz/hdf5-is-for-lovers

Maybe a Third Party's Got Your Back?



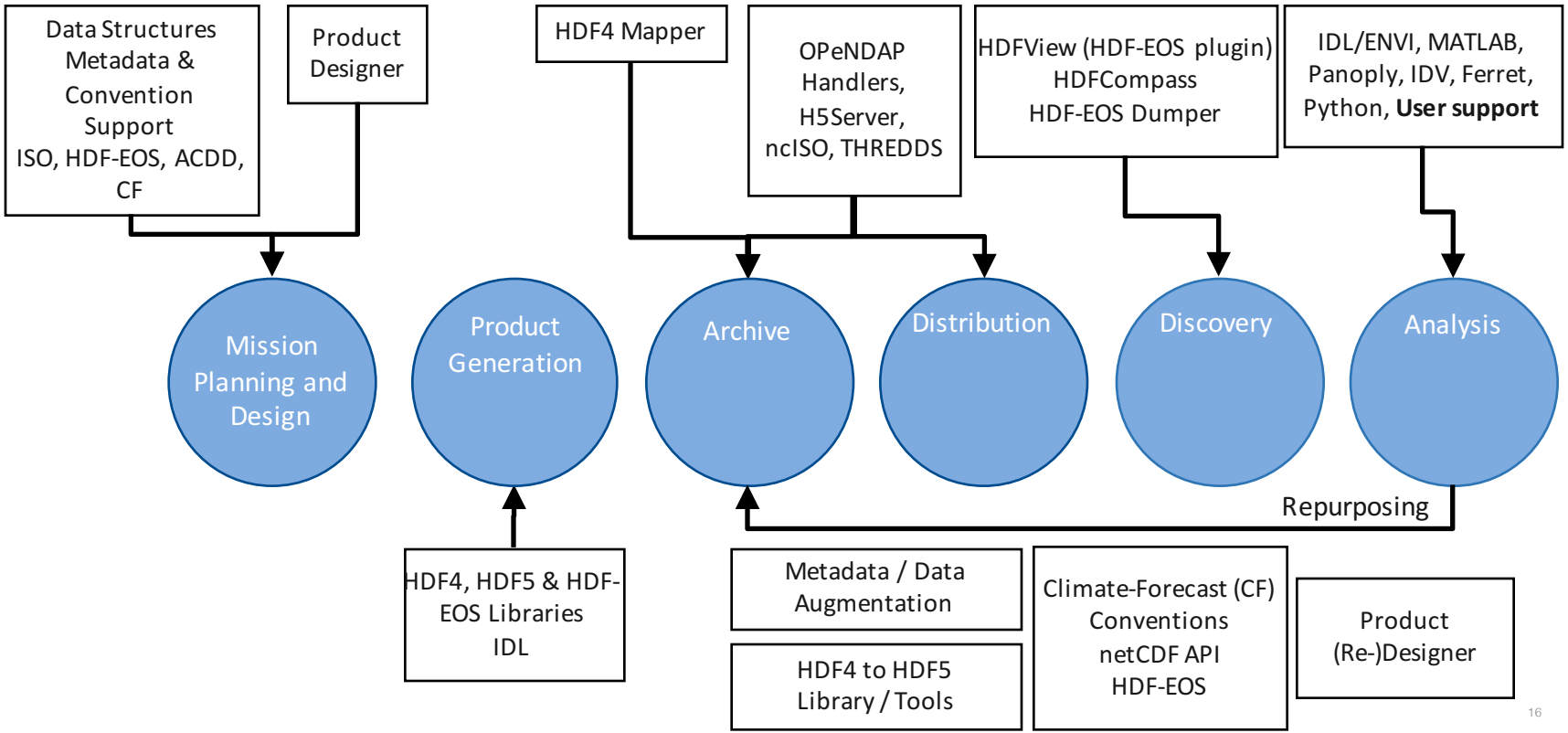
Community Tools





HDF Tools

Supporting the entire data life cycle

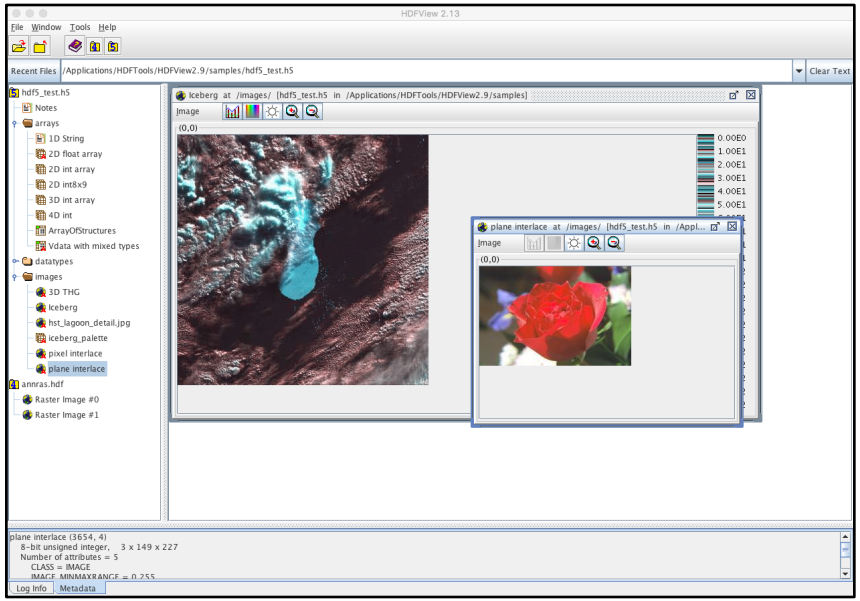


HDFView

HDFView (Java)

Tool for viewing and editing local HDF files

Modify content of dataset AND add, delete or modify attributes



https://www.hdfgroup.org/products/java/hdfview/index.html

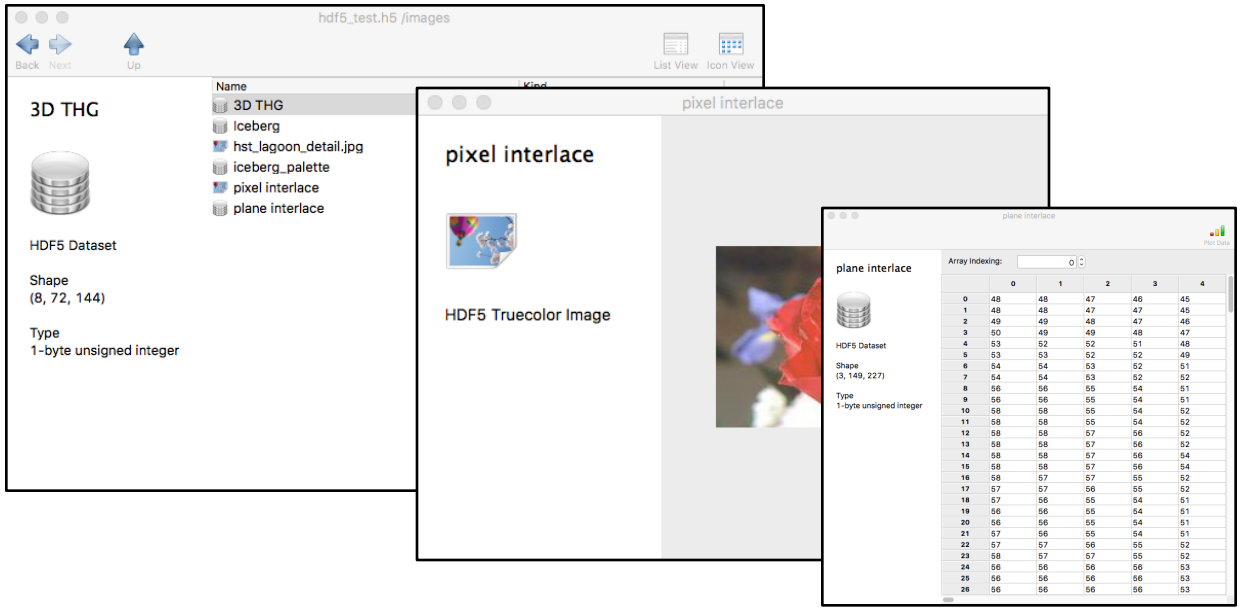
HDF Compass

HDF Compass (Python) https://www.hdfgroup.org/projects/compass/index.html

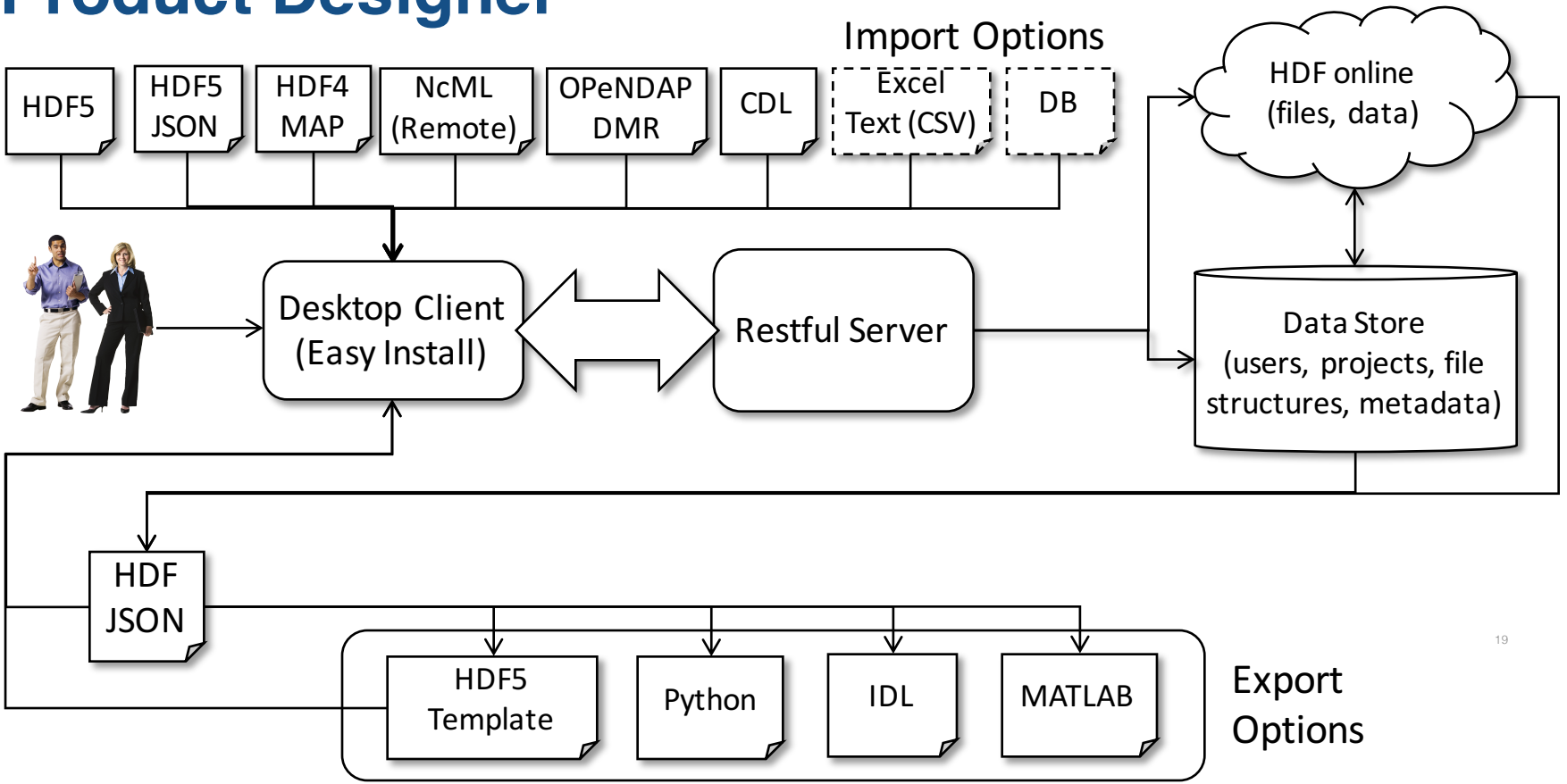
View datasets, attributes and groups, local and remote (No edit capability)

Simple image, line and contour plots

Cross-platform



Product Designer



Product Designer Hub – V0



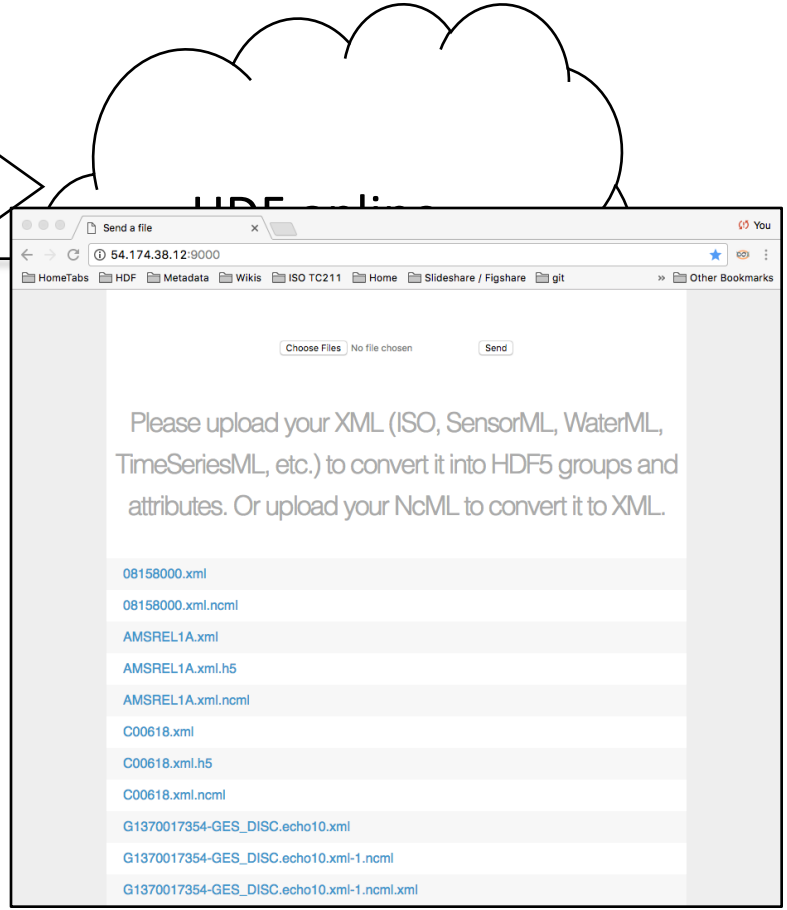
Browser

HDF5
Template

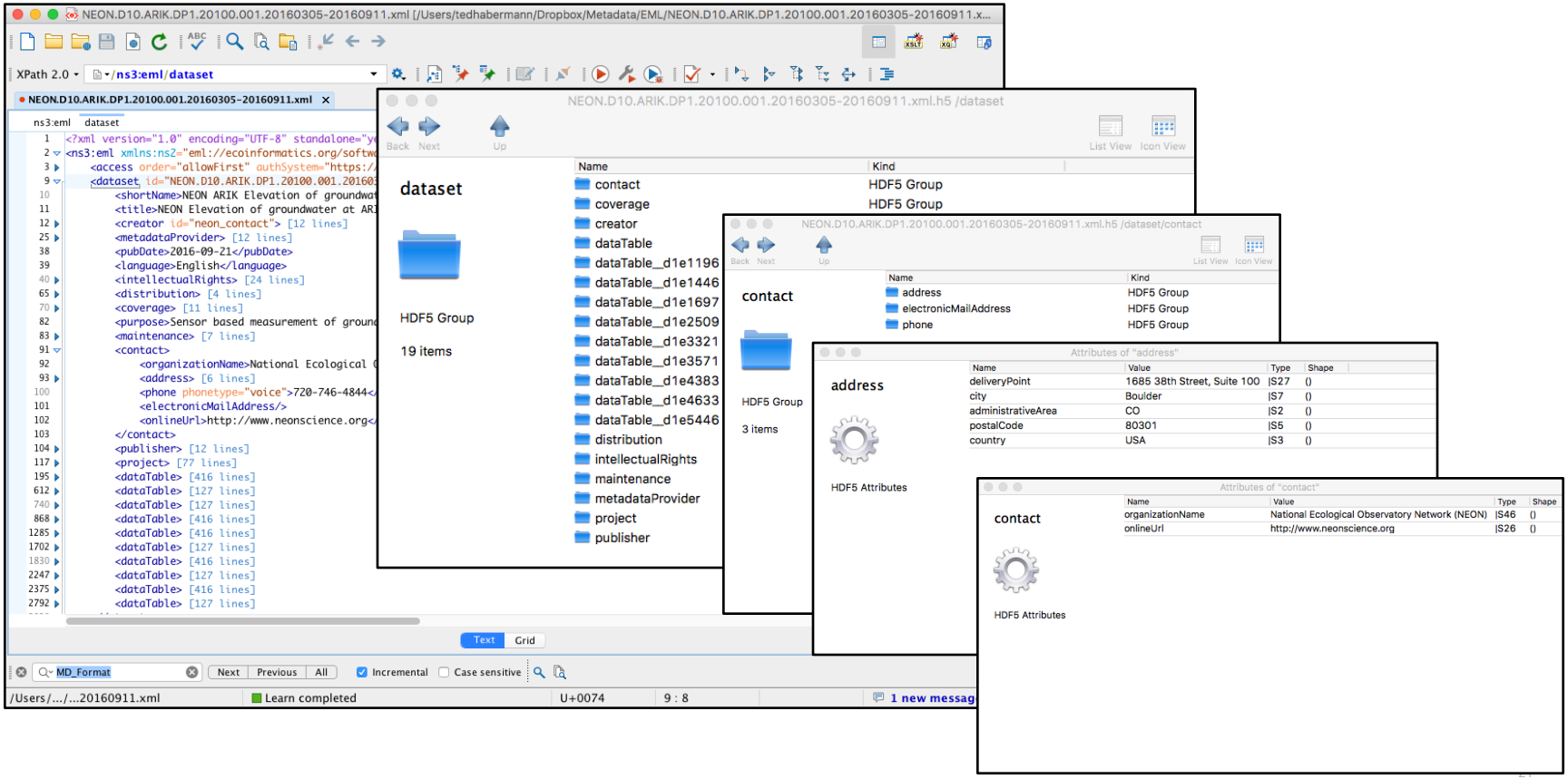
Local
Disk

XML
ISO / NcML / WaterML / SensorML / TSML

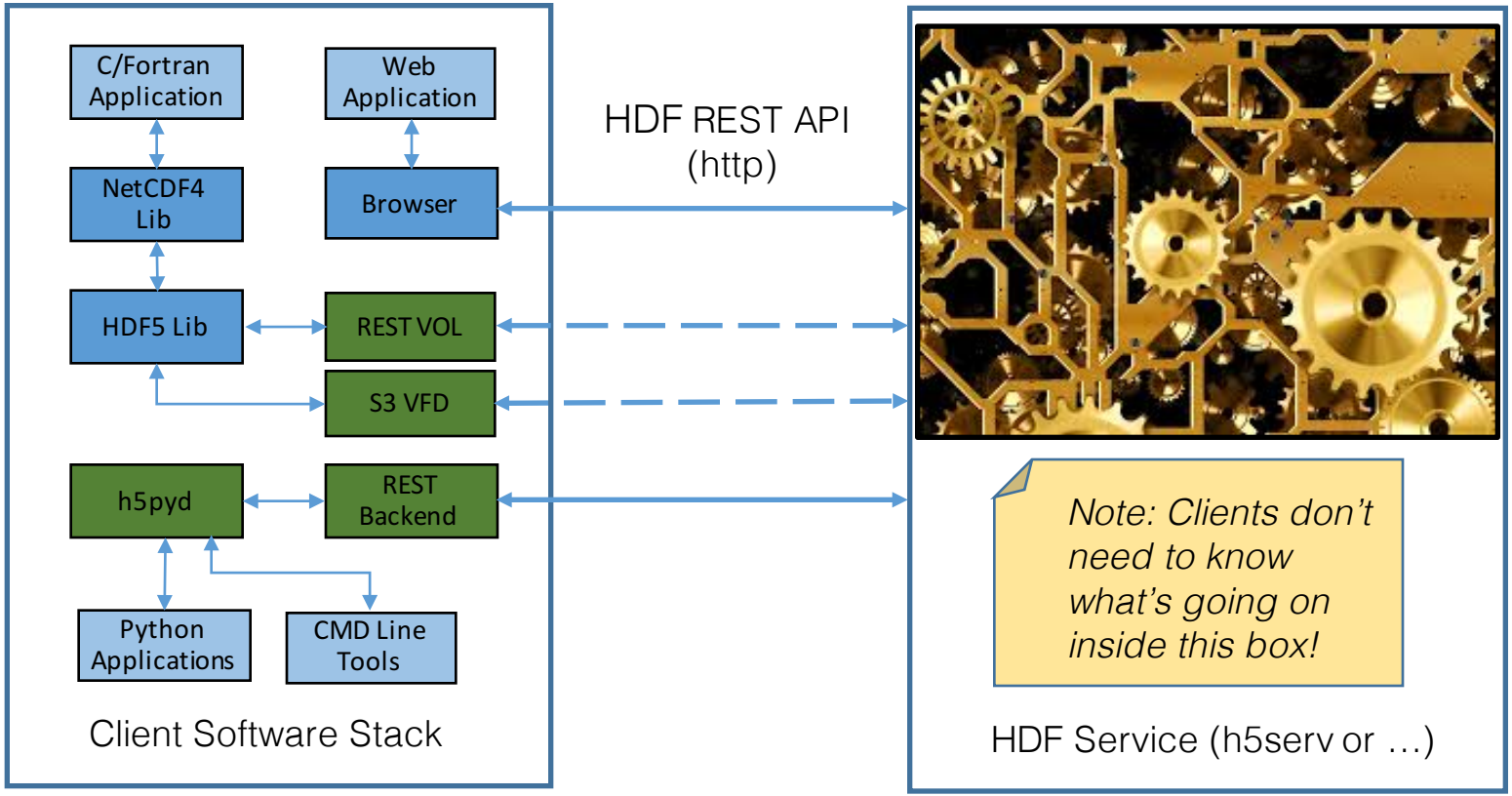
XML
EML



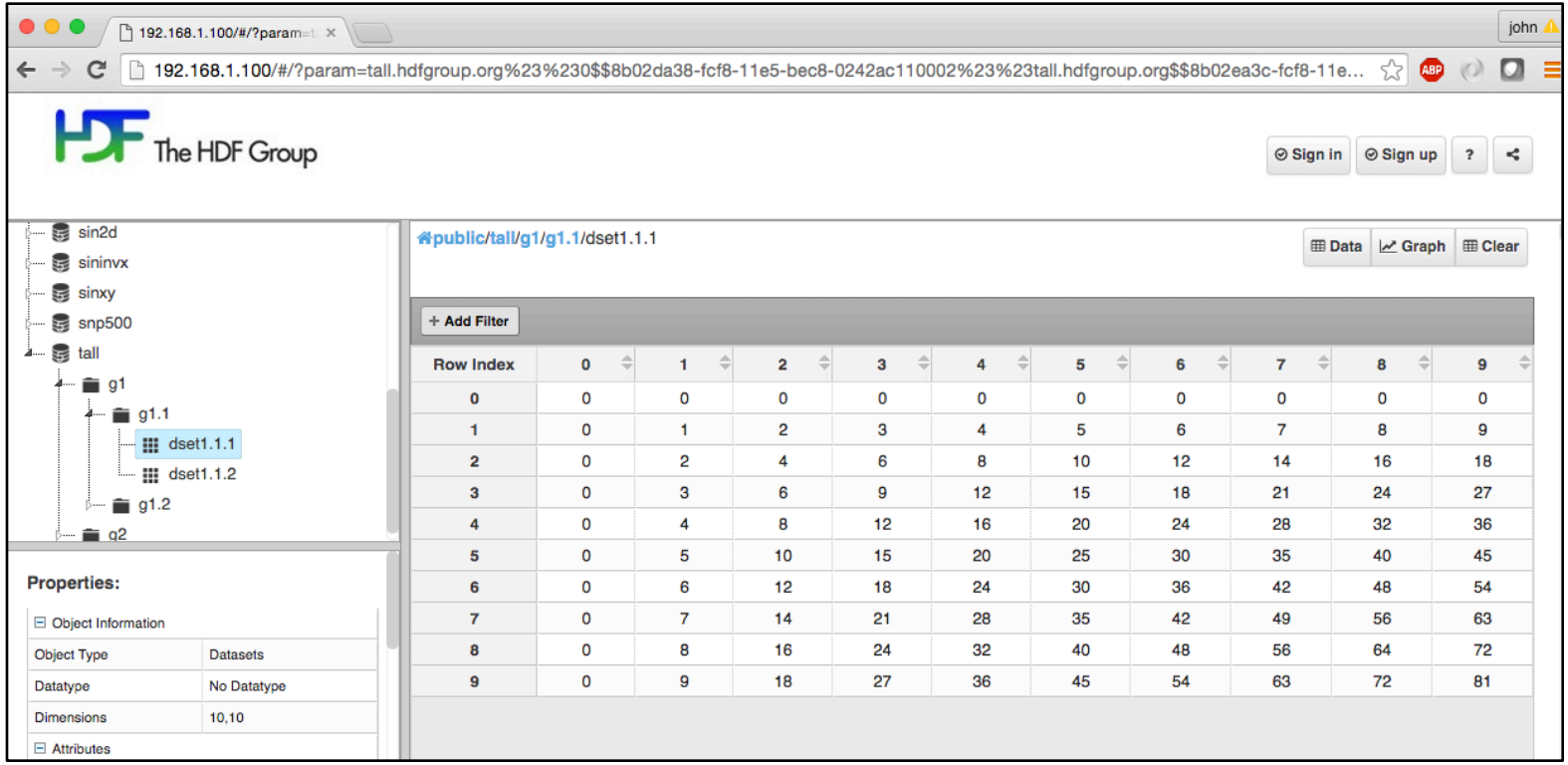
Community Metadata in HDF



Client/Server Architecture

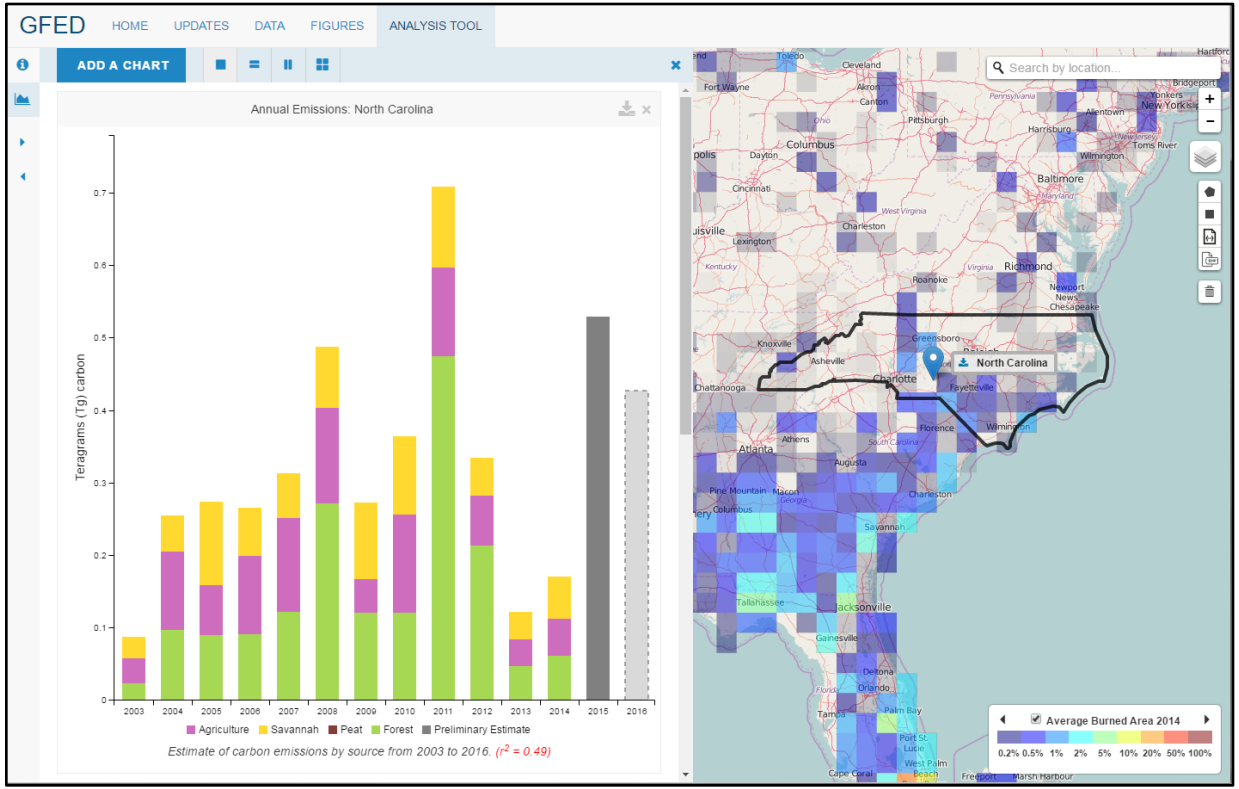


Web UI – Display HDF Content in a browser

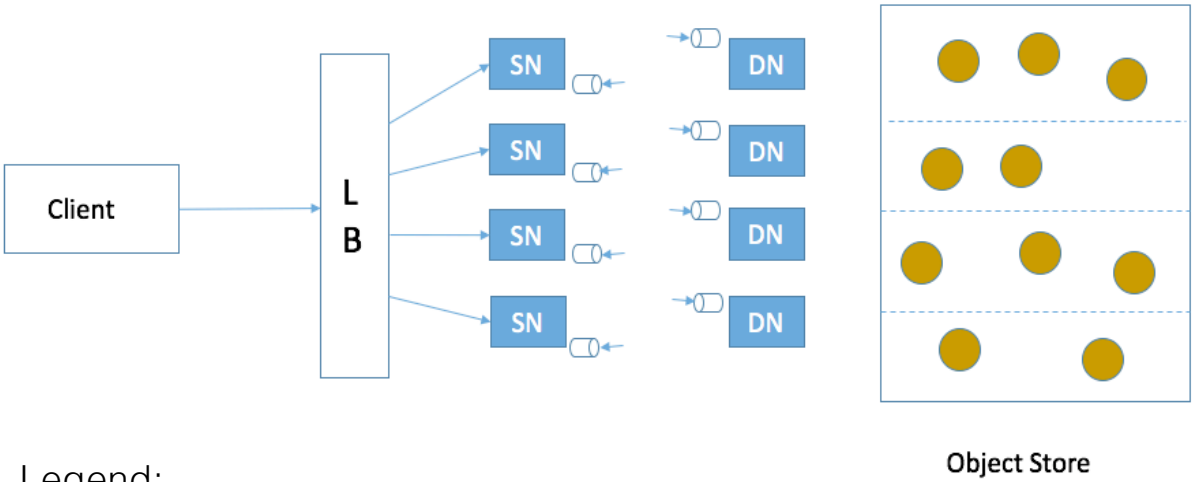


Row Index	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9
2	0	2	4	6	8	10	12	14	16	18
3	0	3	6	9	12	15	18	21	24	27
4	0	4	8	12	16	20	24	28	32	36
5	0	5	10	15	20	25	30	35	40	45
6	0	6	12	18	24	30	36	42	48	54
7	0	7	14	21	28	35	42	49	56	63
8	0	8	16	24	32	40	48	56	64	72
9	0	9	18	27	36	45	54	63	72	81

Global Fire Emissions Database



Architecture for HSDS



Legend:

- Client: Any user of the service
- LB: Load balancer – distributes requests to Service nodes
- SN: Service Nodes – processes requests from clients (with help from Data Nodes)
- DN: Data Nodes – responsible for partition of Object Store
- Object Store: Base storage service (e.g. AWS S3)

To Find out More:



- H5serv: https://github.com/HDFGroup/h5serv
- Documentation: http://h5serv.readthedocs.io/
- H5pyd: https://github.com/HDFGroup/h5pyd
- RESTful HDF5 White Paper: https://www.hdfgroup.org/pubs/papers/RESTful_HDF5.pdf
- OpenNex: https://nex.nasa.gov/nex/static/htdocs/site/extra/opennex/
- Blog articles:
  - https://hdfgroup.org/wp/2015/04/hdf5-for-the-web-hdf-server/
  - https://hdfgroup.org/wp/2015/12/serve-protect-web-security-hdf5/

Thank You!

thabermann@hdfgroup.org

Questions?