The HDF Group

By Elena Pourmal, Director & Dax Rodriguez, Director

The HDF Group

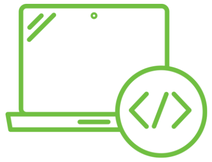
Agenda

- **Overview of The HDF Group and the HDF5 Library**
- **The HDF Dataverse: Standards Building**
- **Building Sustainability as a Not-for-Profit**

Who is the HDF Group?

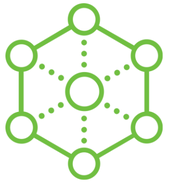
“De-facto standard for scientific computing” and integrated into every major analytics + visualization tool

HDF Group has developed open source solutions for over 30 years



Small not-for-profit company focus on Performance Computing and Scientific Data

Headquarters in Champaign, IL



Our flagship platform – HDF5

Thousands use + build on HDF5 every day (~1000+ projects on Github)

What We Do?

Products

- HDF5 Community Edition
- HDF5 Enterprise Support Edition (Future)
- HDF Cloud Platform: HDF5 object storage service (Beta)

Consulting & Support Services

- Create semi-custom and custom data platforms for scientific communities, e.g. IoT, Deep Learning, etc.
- Add features to HDF5
- Performance analysis of HPC applications
- Embedded with federal agencies and engineering teams
- Training

Metadata Services

- Facilitate creation of new standards
- Data conversion and compliance
- Vendor-independent reference implementations
- Metadata for variables, data quality, and lineage
- Integration of standard metadata (e.g. ISO, SensorML) with data in HDF5 files.

What Sets HDF Apart?

The HDF Group

- Open source: vendor independent
- Large dedicated community: we are here to stay
- Certified: used in government, healthcare, and finance

Our Industries

The HDF Group

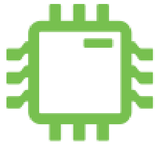
Financial Services

Oil and Gas

Aerospace

Automotive

Medical & Biotech



Silicon Manufacturing

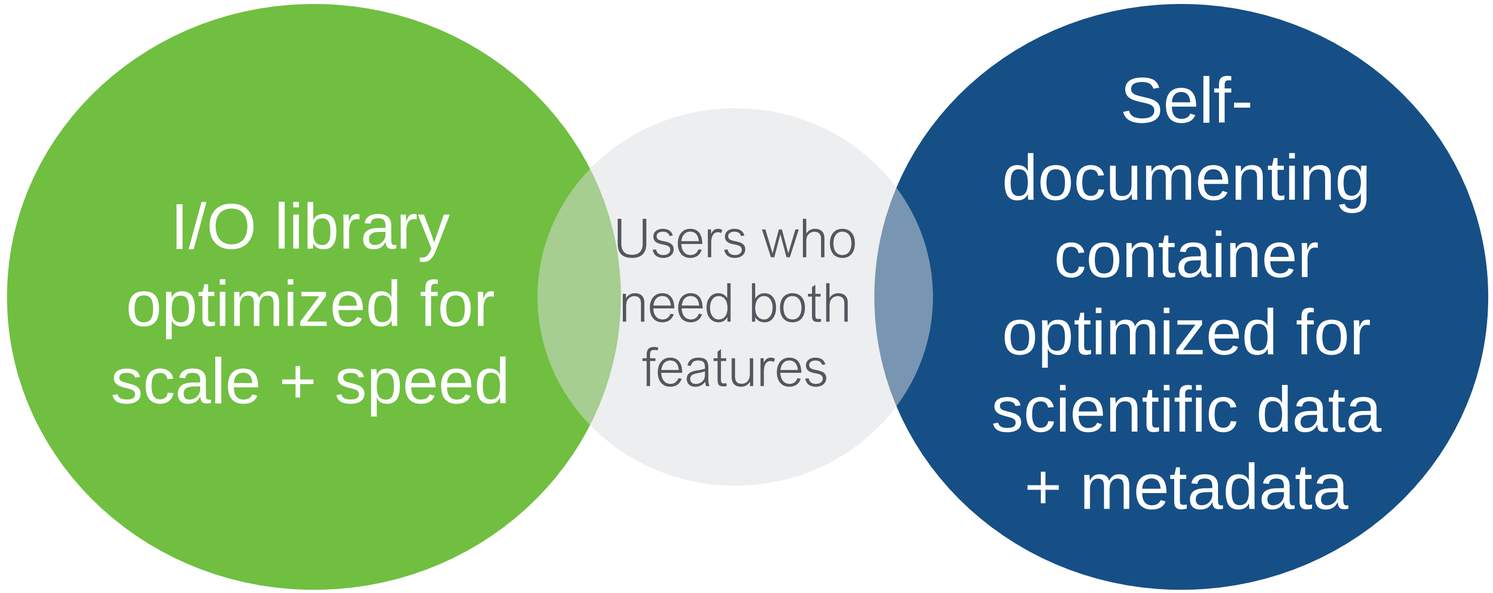
Electronics Instrument

Government

Defense & National Security

Academic Research

Why HDF Technologies?

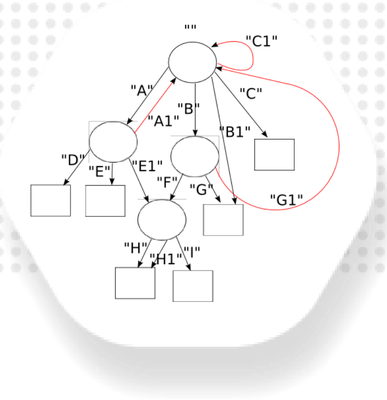


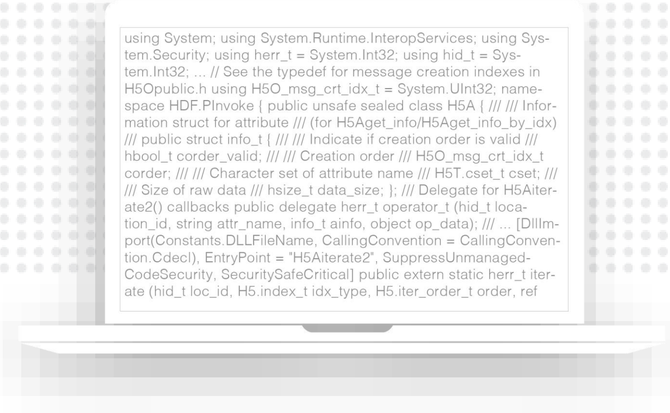
High performance, infrastructure built for big data storage and processing

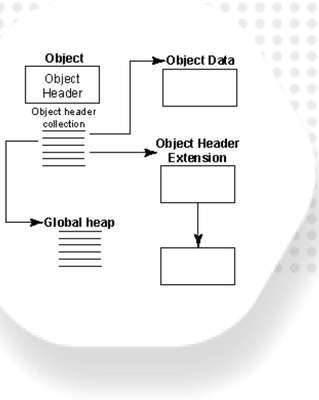
Platform to build community-specific and domain-specific data types and conventions

The HDF5 Platform

Marriage of data model + I/O software + binary container







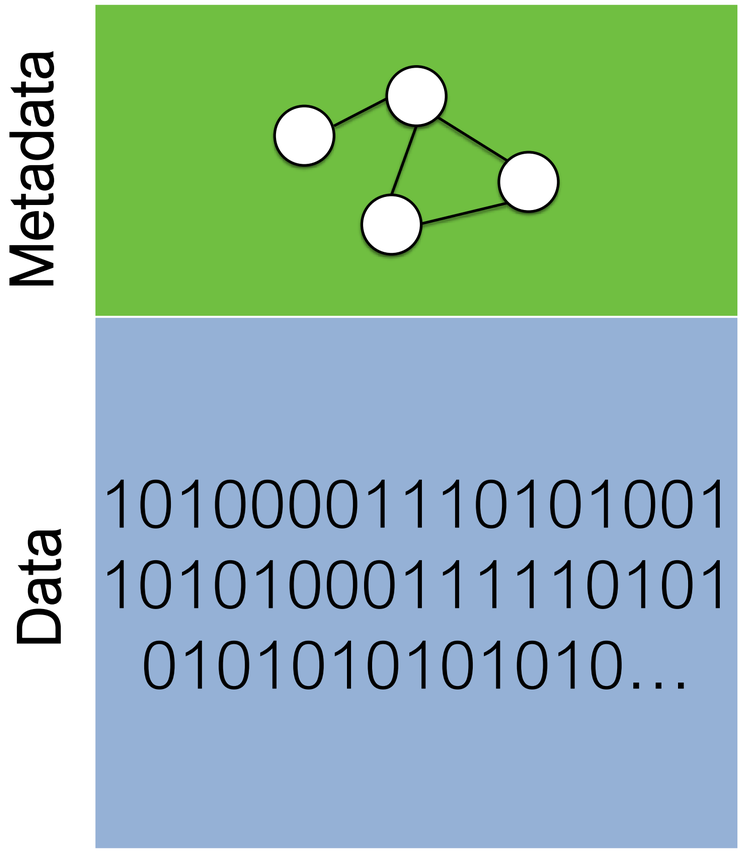
HDF5 **abstract data model**

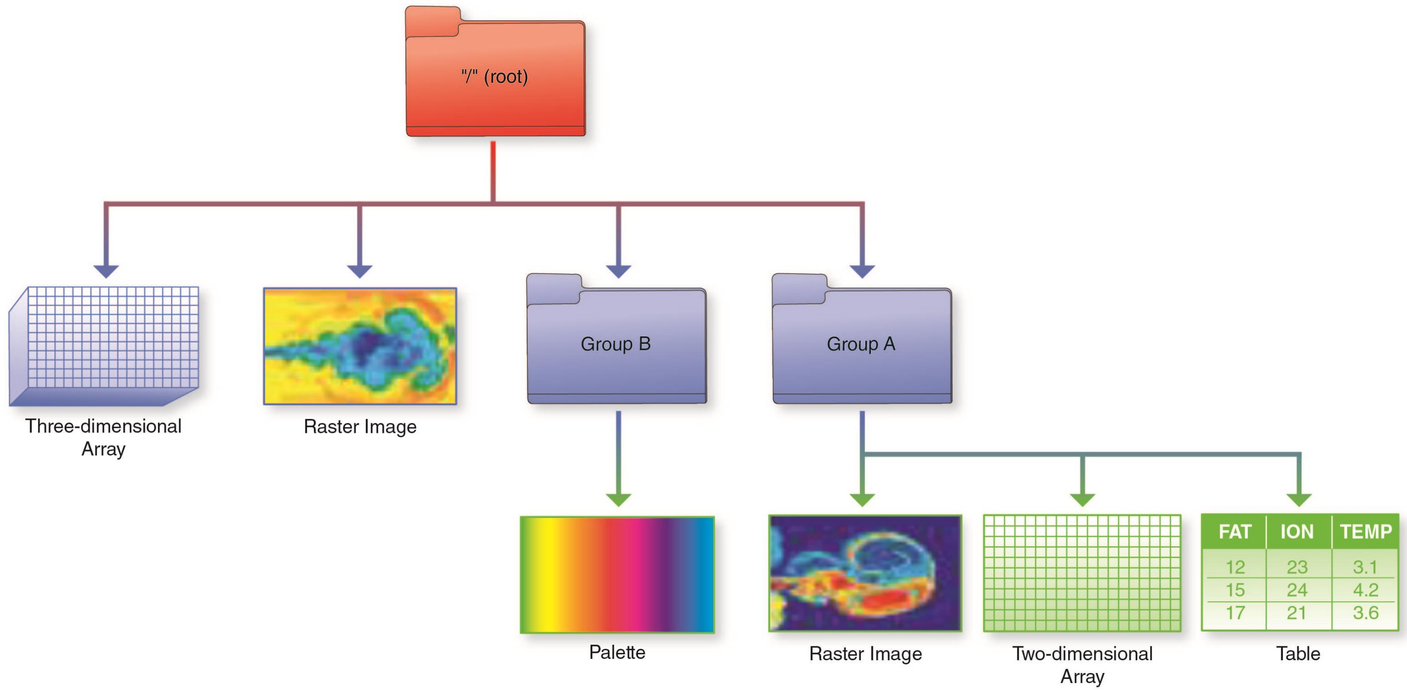
HDF5 **library**

HDF5 **file format**

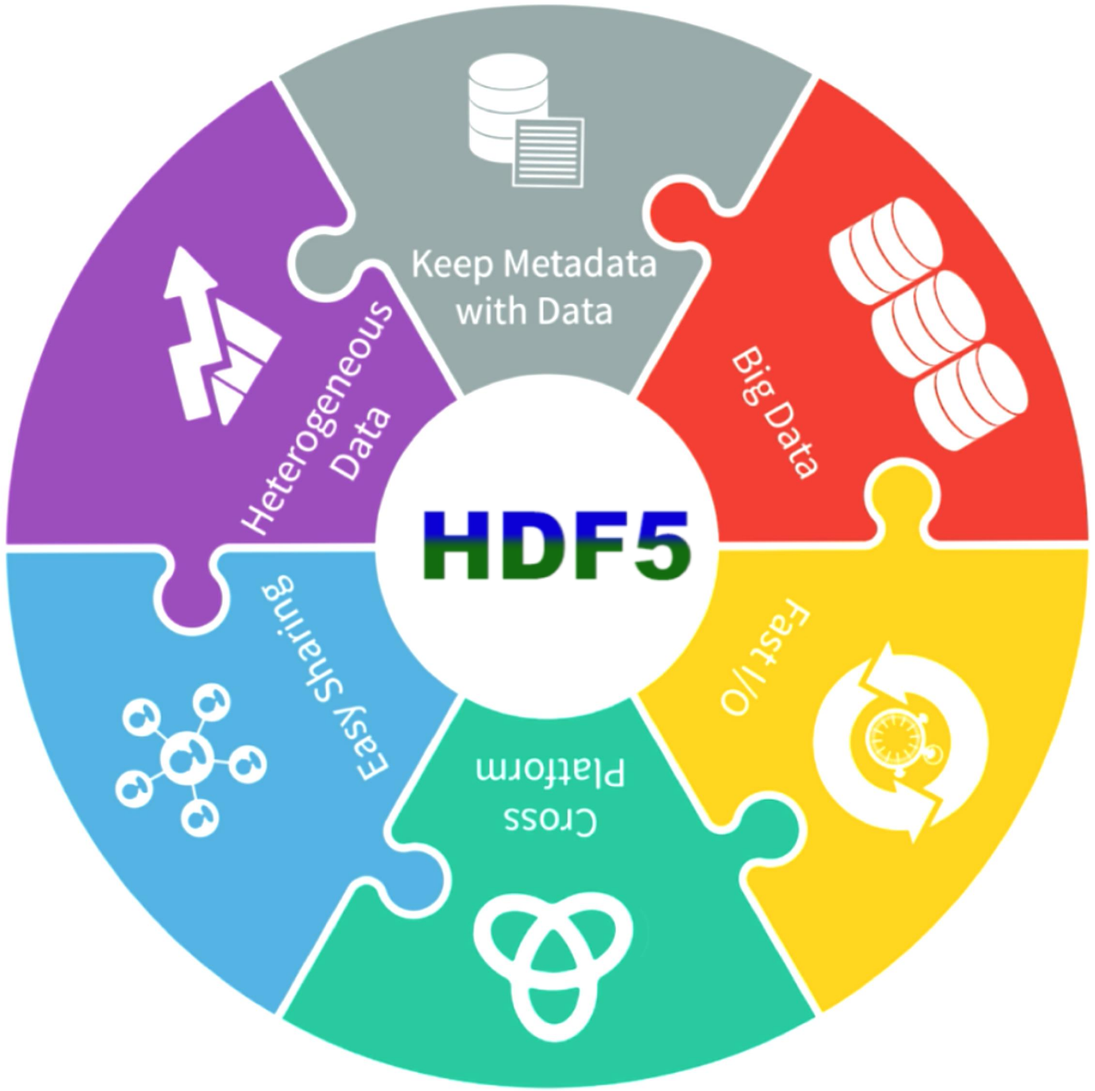
C library with APIs for <u>every</u> programming language: python, C, C++, Java, Fortran, etc.

HDF5 Container





Why is this Concept so Different + Useful?



- **Native support for multidimensional data**
- **Data and metadata in one place => streamlines data lifecycle & pipelines**
- Portable, no vendor lock-in
- Maintains logical view while adapting to storage context
- In-memory, over-the-wire, on-disk, parallel FS, object store
- Pluggable filter pipeline for compression, checksum, encryption, etc.
- High-performance I/O
- Large ecosystem (1000+ Github projects)

Agenda

- Overview of The HDF Group and the HDF5 Library
- The HDF Dataverse: Standards Building
- Building Sustainability as a Not-for-Profit

We don’t make standards…

… We help communities turn standards into software

The HDF5 Dataverse

Building HDF5 Standards

- Communities build domain specific data types, objects and conventions on top of HDF5:
  - It is a high-performance infrastructure built for big data storage, processing, archiving, mining, and exchange. ***Scientists can focus on doing science and don't need to think about I/O and storage.***
  - Domain specific data types can be easily represented by HDF5 primitives
  - HDF5 is well supported and evolves!

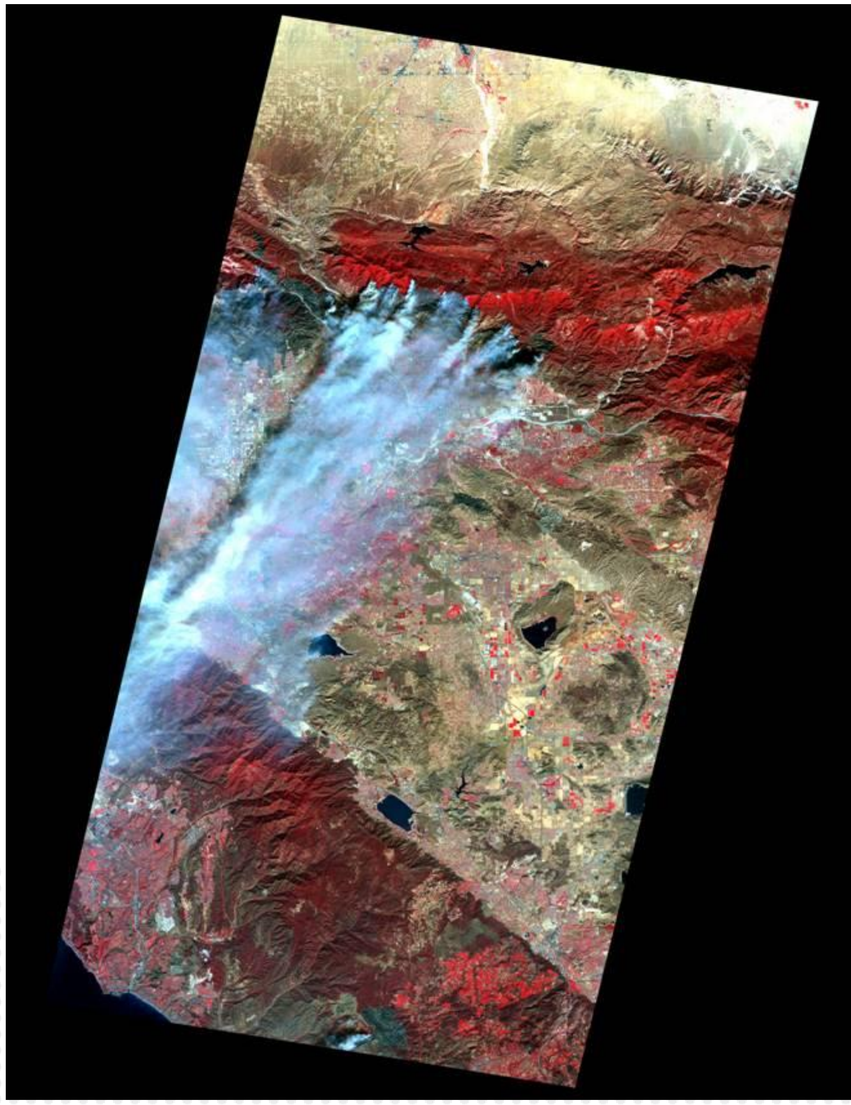


Image of California fires (HDF-EOS Data product from LP DAAC)

Industry Formats Built on HDF5

The HDF Group

Industry	Format	Notes
Research & Science	netCDF-4	NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data
Bio-Tech & Pharma	ADF	The Allotrope Data Format (ADF) is a federation of standards that features the ability to store datasets of nearly unlimited size and complexity in a single file, organized as a single or multiple n-dimensional arrays to record the measurements of experiments
Oil & Gas	RESQML	RESQML™ is an industry initiative to provide open, non-proprietary data exchange standards for reservoir characterization, earth and reservoir models.
Entertainment	Alembic	Alembic is an open computer graphics interchange framework used as data representation scheme for storing computer graphics scenes (Lucas Films)

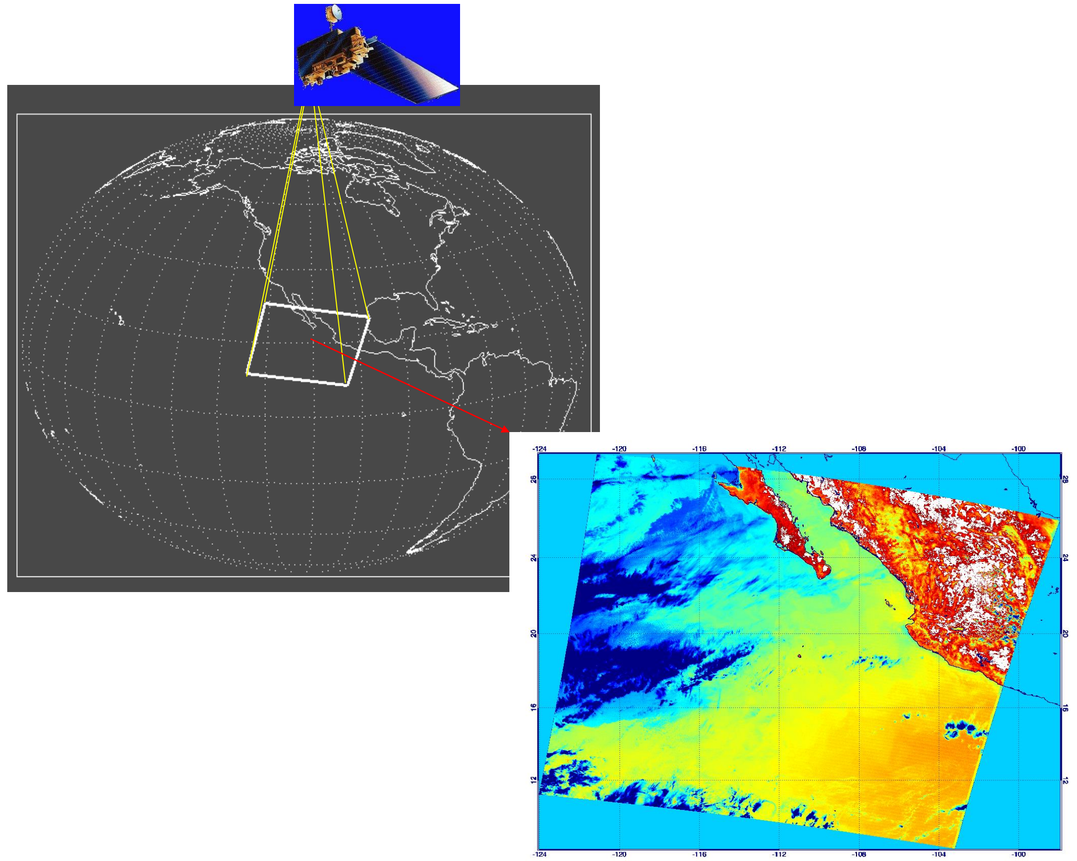
Insights on Standards Building

The HDF Group

The

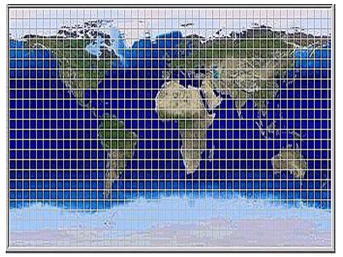
NASA

Case Study



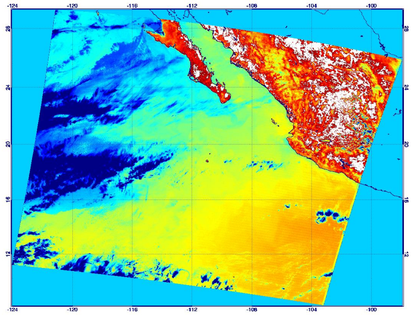
Earth Science Data Structures

- **GRID - Data which is organized by regular geographic spacing, specified by projection parameters.**



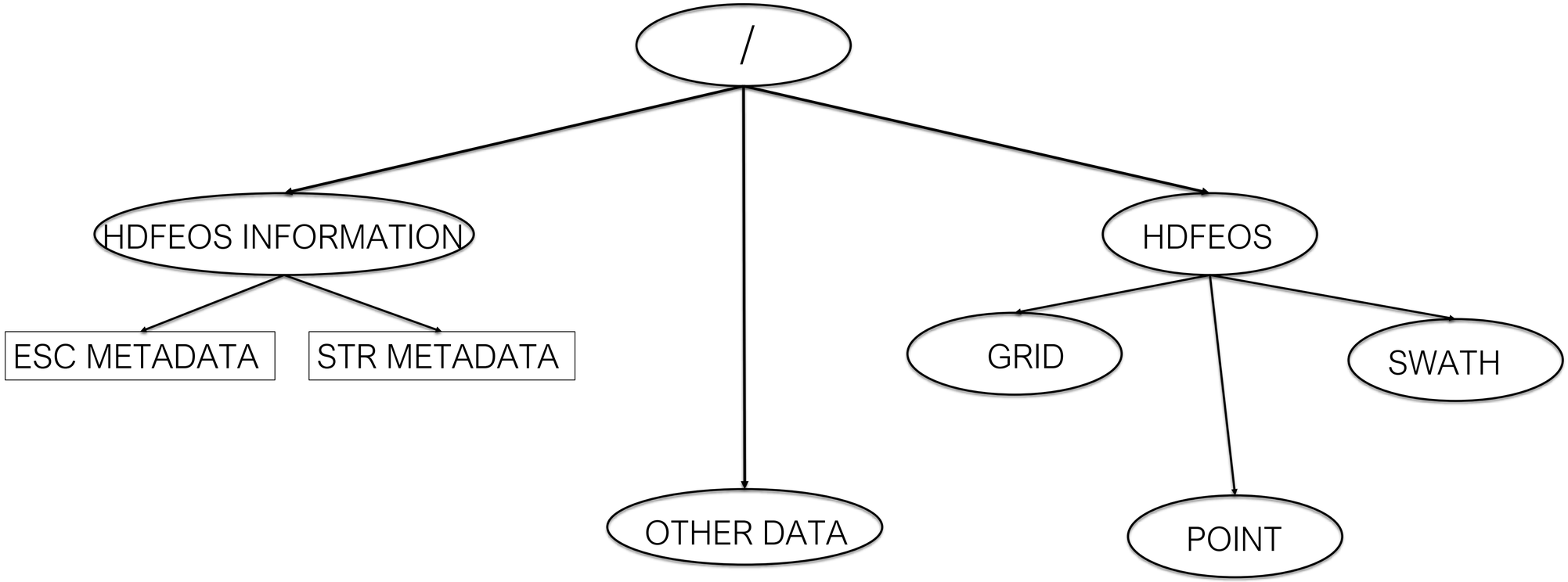
- **Structure**
  - Any number of 2-D to 8-D data arrays per structure, one per data type (e.g. temperature)
  - Geolocation information contained in projection formula, coupled by structural metadata.
  - Any number of Grid structures per file allowed.

- **SWATH - Data which is organized by time, or other track parameter. Spacing can be irregular.**



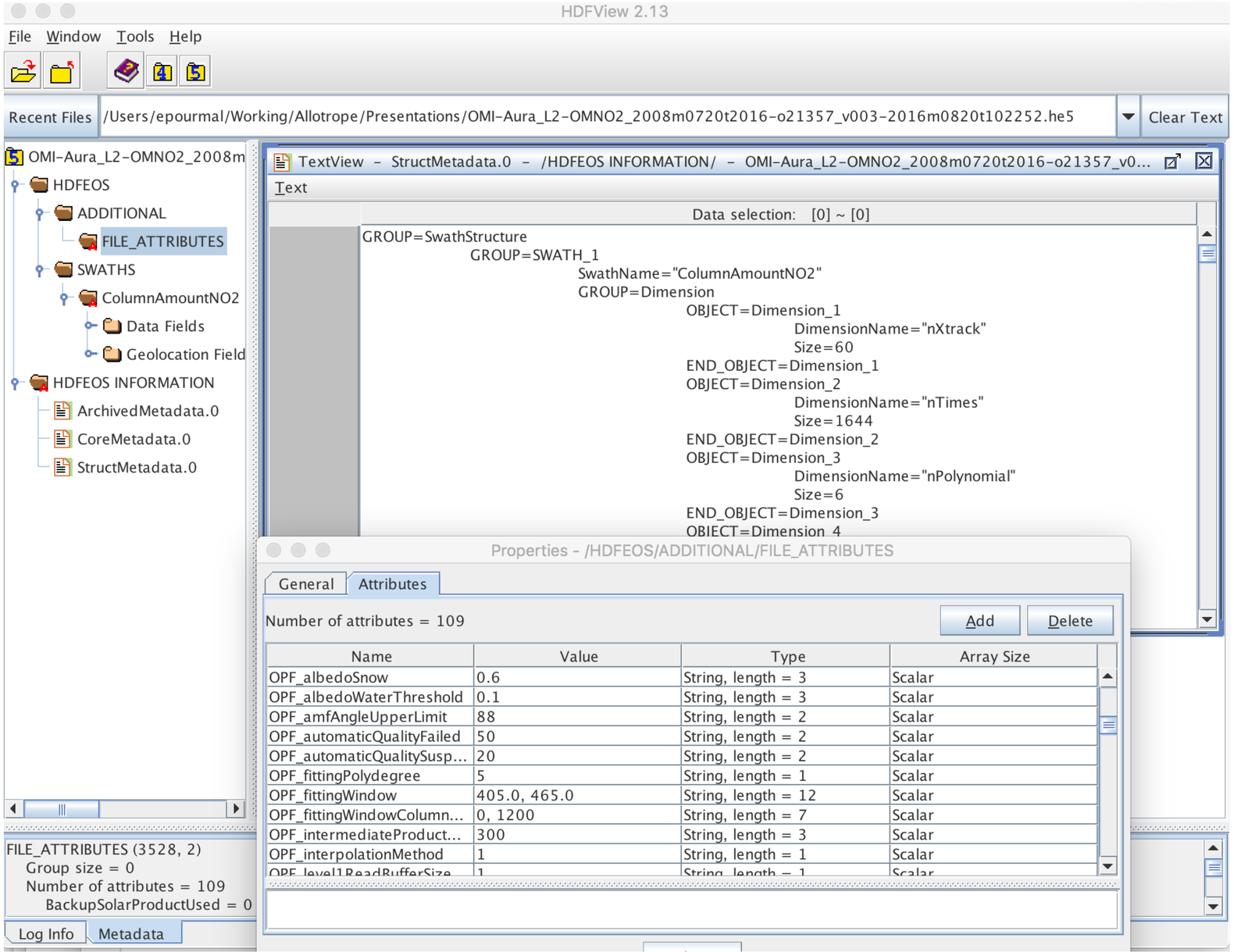
- **Structure**
  - Geolocation information stored explicitly in Geolocation Field (2-D array)
  - Data stored in 2-D or 3-D arrays
  - Time stored in 1-D or 2-D array
  - Geolocation/science data connected by structural metadata

HDF-EOS File Structure



HDF-EOS5 library creates necessary structures

Aura HDF-EOS5 File



AURA products standards based on HDF-EOS5

- https://cdn.earthdata.nasa.gov/conduit/upload/518/ESDS-RFC-018v1.pdf

- **GOAL: Help the end user to develop one universal reader to read the primary data within the Aura teams' data files.**
  - Items which did not affect the reading of the data were not standardize (e.g., compression)
  - Examples of the items standardization was done on:
    - Names of the fields
    - Names and ordering of dimensions for each field
    - Datatype and sizes of each field (e.g., 32-bit integer, no endianess)
    - Attributes for each field and their types and definitions
    - Units for each field
    - Coordinate system

Agenda

- Overview of The HDF Group and the HDF5 Library
- The HDF Dataverse: Standards Building
- Building Sustainability as an Organization

Organizational Challenges

Acceptance

- Obtaining wide acceptance and usage
  - ✓ Provide education and training

Financial Support

- Even if the standard is accepted, we must financially support the standard and organization
  - ✓ Membership and license fees

Building Community

- It is dangerous for one individual or entity to assume all financial burden or set the direction
- We must have an active, diverse community for healthy evolution
  - ✓ Get Involved – Technical Support, Workshops, Webinars, and Committees

How Can HDF Help?

Form a mutually beneficial partnership between our organizations!

Acceptance

Leverage our experience and expertise implementing HDF5 standards & solutions

- Provide education and formal training of advanced HDF5 features and best practices

Financial Support

Leverage existing HDF Infrastructure tech and data interoperability tools through the Allotrope Foundation

- HDF Cloud
- HDF Connectors (ODBC, JDBC, Spark)
- New (.Net and Win32 Wrapper)

Community

The HDF Group will proactively participate in committees, conferences, and webinars. We will also be an additional, expert resource for technical support

THANK YOU!

Questions & Comments?