

HDF The HDF Group

HDF5

Mike Folk
The HDF Group, Champaign IL
MEDSBIO Workshop at Diamond
August 17, 2007

www.hdfgroup.org

HDF What is "The HDF Group"

- HDF implemented 1988 at NCSA, to manage science data among supercomputers and workstations
- HDF5 implemented in 1998
- Hundreds of different apps of HDF5
- Millions of users of HDF/HDF5
- Spinoff in 2006
 - "The HDF Group," a non profit company
 - Mission: *sustain HDF and serve HDF users*

May 22 2007 NCSA PSP Meeting 2 www.hdfgroup.org

HDF HDF5 is

1. File format for managing any kind of data
2. Software to manage data in the format

- Suited especially to large or complex data
- Designed to be efficient, flexible, extendable
- Runs on every size and type of system
- Open format and software

May 22 2007 NCSA PSP Meeting 3 www.hdfgroup.org

HDF An HDF5 file is a container...

...into which you can put your data objects.



May 22 2007 NCSA PSP Meeting 4 www.hdfgroup.org

HDF Structures to organize objects in files

"Groups"

"Datasets"

May 22 2007 NCSA PSP Meeting 5 www.hdfgroup.org

HDF HDF5 data model

- Groups – structure among objects
- Datasets – data arrays
 - Rich set of datatype options
 - Flexible, efficient I/O
- Attributes, for metadata

Everything else is built essentially from these parts.

May 22 2007 NCSA PSP Meeting 6 www.hdfgroup.org



Storage and I/O options

- Application can select storage and I/O methods most appropriate to type of data, access patterns, file system, etc.
- E.g. structures for efficient access
 - Built-in – indexes, etc.
 - Optional – compression, chunking, etc.

May 22 2007

NCSA PSP Meeting

7

www.hdfgroup.org



HDF5 Software

Tools & Applications

Domain-specific apps (e.g. NeXus)
 Problem-specific apps
 COTS and other tools
 Data management utilities

HDF I/O Library

C, C++, Fortran, Java, other languages
 Filters and I/O drivers for speed, real-time processing, compression, etc.

HDF File

May 22 2007

NCSA PSP Meeting

8

www.hdfgroup.org



The HDF Group

Examples of how HDF5 is used

www.hdfgroup.org



The HDF Group



Understanding Global Warming

--

NASA

Earth Observing System

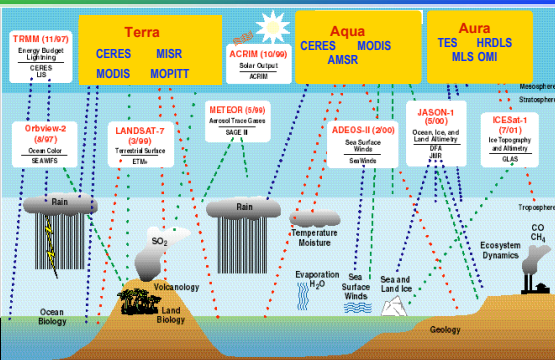
Data & Information System

(EOSDIS)

www.hdfgroup.org

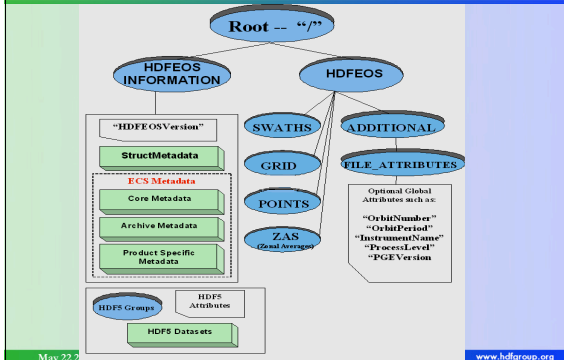


NASA Earth Observing System (EOS)





General HDF-EOS5 File Structure



May 22 2007

www.hdfgroup.org

HDF EOSDIS data requirements

- Volume
 - 3 terabytes/day = 15 petabytes in 14 years
- Complexity
 - dozens of instruments and scores of datasets
- Accessibility
 - search, subset, re-organize
 - software quality and evolution
- Usability
 - 2 million users
 - tools, portability, standardization
 - user support

May 22 2007 NCSA PSP Meeting 13 www.hdfgroup.org

HDF What THG does for EOSDIS

- Maintain/develop HDF formats, I/O libraries
- Help data producers/users/vendors through consulting, documentation, tutorials, workshops
- Tools for viewing, editing and managing data

May 22 2007 NCSA PSP Meeting 14 www.hdfgroup.org

The HDF Group Bioinformatics

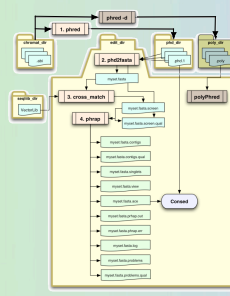
Managing genomic data

caacaagccaaaactcgtacaa
 Cgagatatctcttgaaaaaact
 gtcacaatatgtgacgtacaag
 gttgttcatgaaacttcggta
 Acaatogttgacattgcgacot
 aatacagcccagcaagcagaat



www.hdfgroup.org

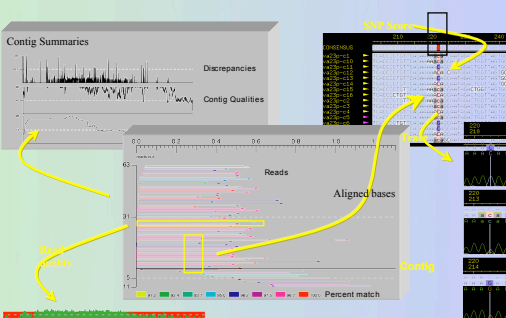
HDF DNA sequencing workflows



- Diverse formats
- Highly redundant data
- Repeated file processing
- Disconnected programs
- Non-scalable storage
- Lack of persistence

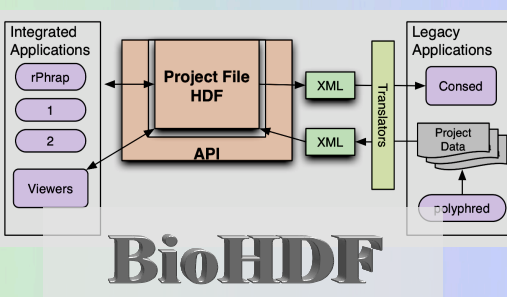
May 22 2007 NCSA PSP Meeting 16 www.hdfgroup.org

HDF Multiple levels and relationships



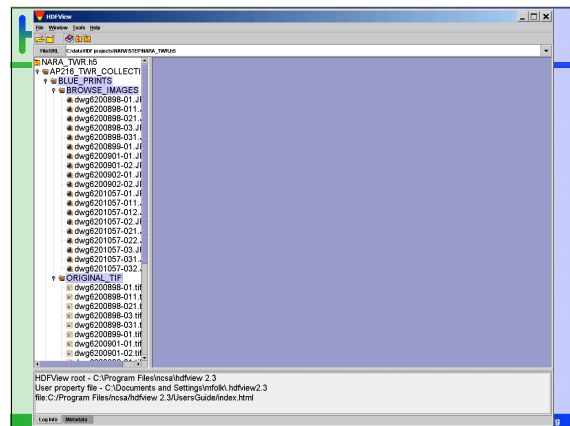
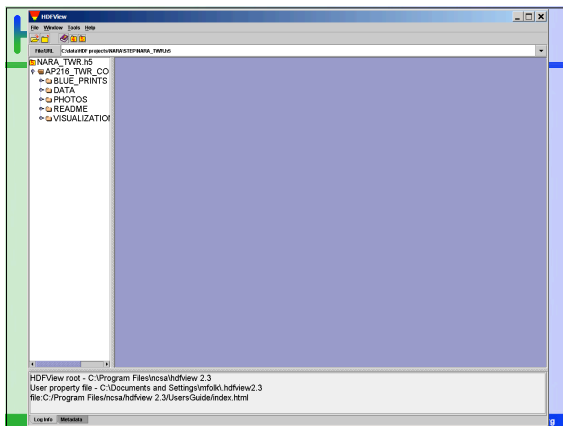
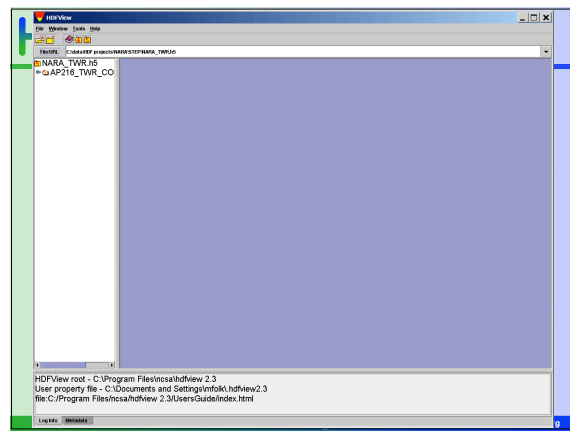
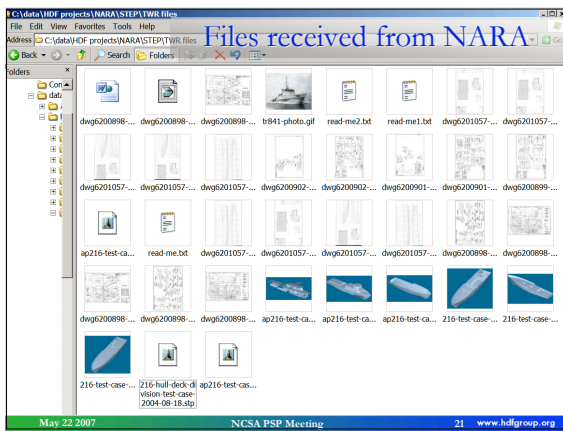
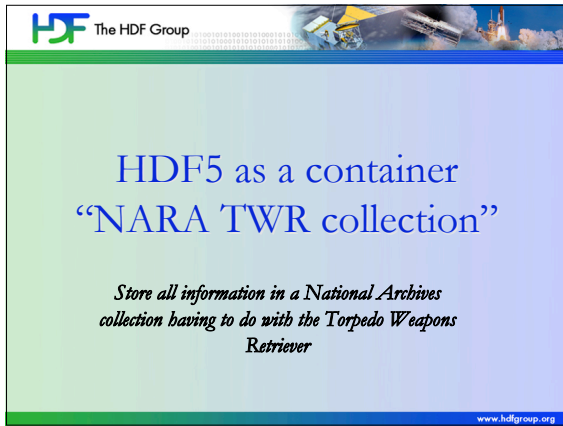
May 22 2007 NCSA PSP Meeting 17 www.hdfgroup.org

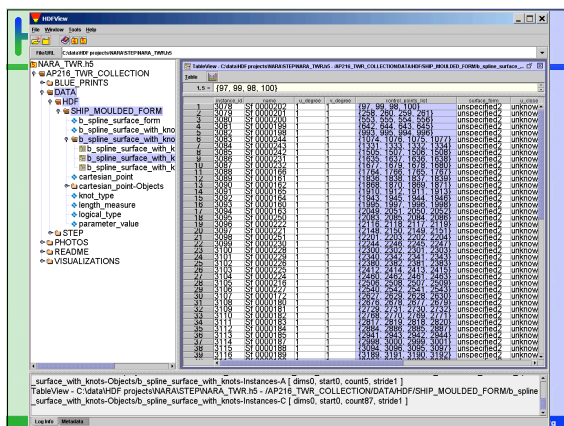
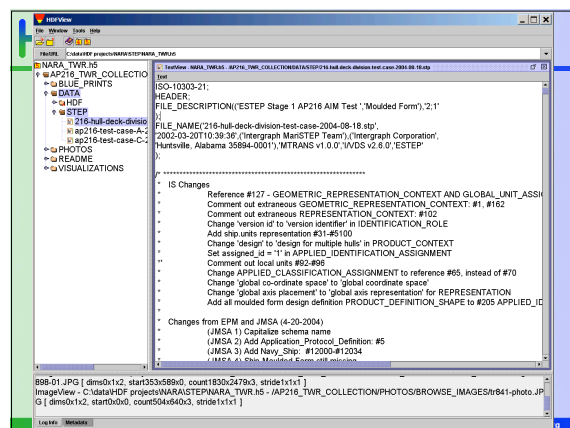
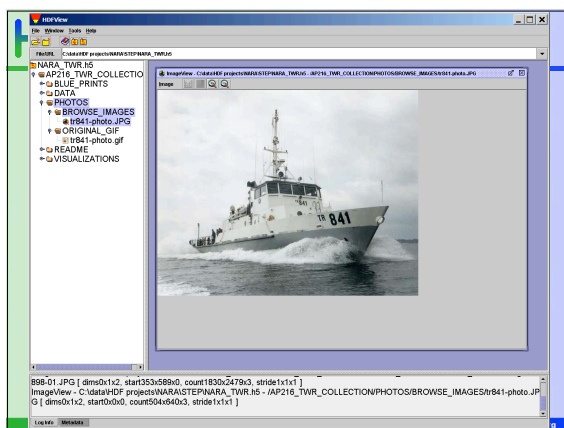
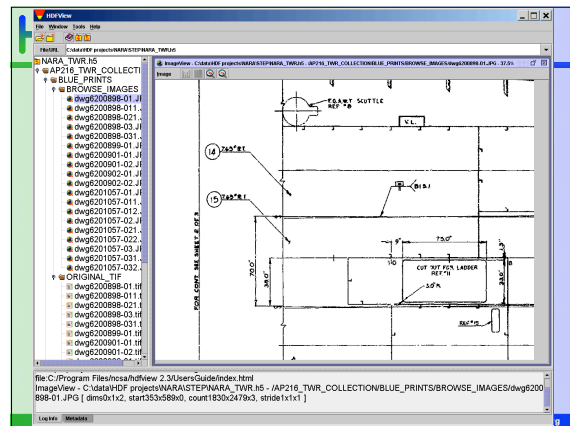
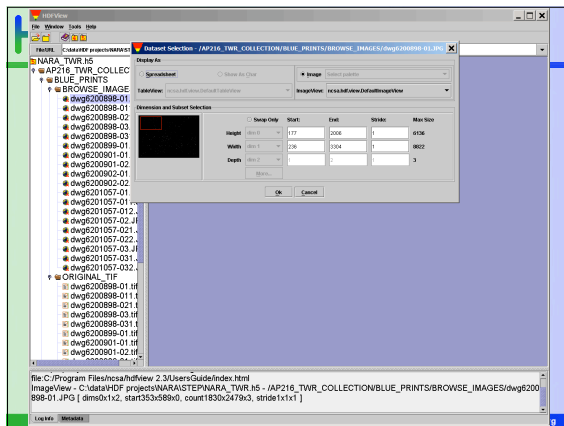
HDF bioinformatics



BioHDF

May 22 2007 NCSA PSP Meeting 18 www.hdfgroup.org





HDF Flight test

May 22 2007 NCSA PSP Meeting 31 www.hdfgroup.org

HDF Flight test data requirements

- Fast data acquisition from 1000s of sources
- Wide variety of data types
- Active archive
- Standard storage for data/software exchange
- Specialized software tools, libraries
- COTS accessibility

May 22 2007 NCSA PSP Meeting 32 www.hdfgroup.org

HDF What we do for Boeing

- Create software for apps ("HDF packet")
- Tune I/O library and format
- Port to specific platforms
- Add features to I/O library and format
- Train/consult to optimize use of HDF5

May 22 2007 NCSA PSP Meeting 33 www.hdfgroup.org

HDF The HDF Group

HDF5 and Databases

May 22 2007 NCSA PSP Meeting 34 www.hdfgroup.org

HDF Hybrid: side-by-side

RDBMS used for SQL queries; points to objects in HDF5 files

Application

Direct access thru HDF5 API for scientific analysis

SQL query Query results

Analysis Analysis results

RDBMS

Indexes

Relational tables

HDF5

May 22 2007 NCSA PSP Meeting 35 www.hdfgroup.org

HDF Hybrid: side-by-side

VISION

"Proud of our past, with a Vision for Tomorrow"

Aberdeen Test Center

May 22 2007 NCSA PSP Meeting 36 www.hdfgroup.org

