# Data and Metadata standardization in the Madrigal Database

Bill Rideout
MIT Haystack Observatory
brideout@mit.edu

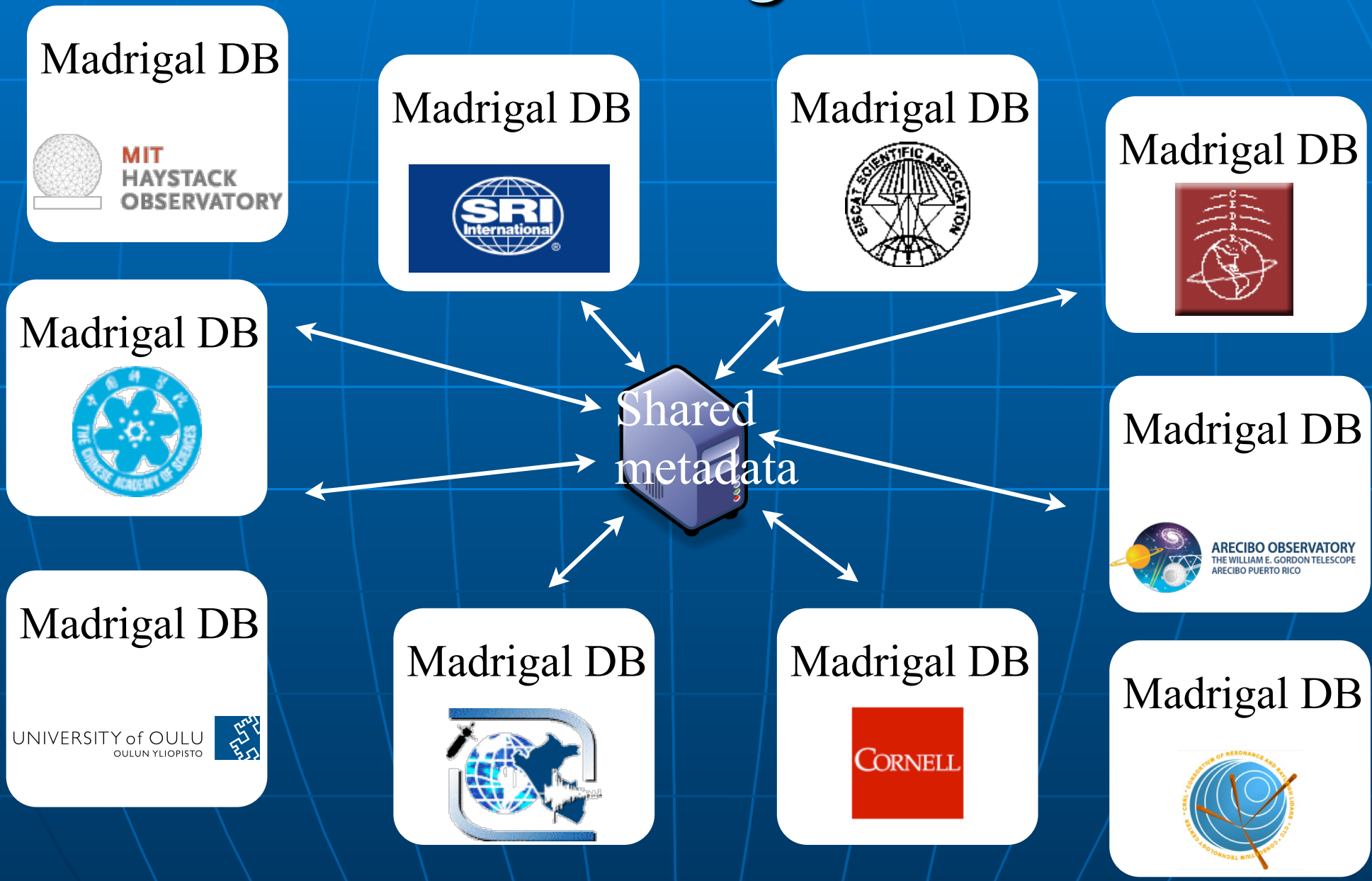
2023 IHDEA Meeting
APL

# Outline

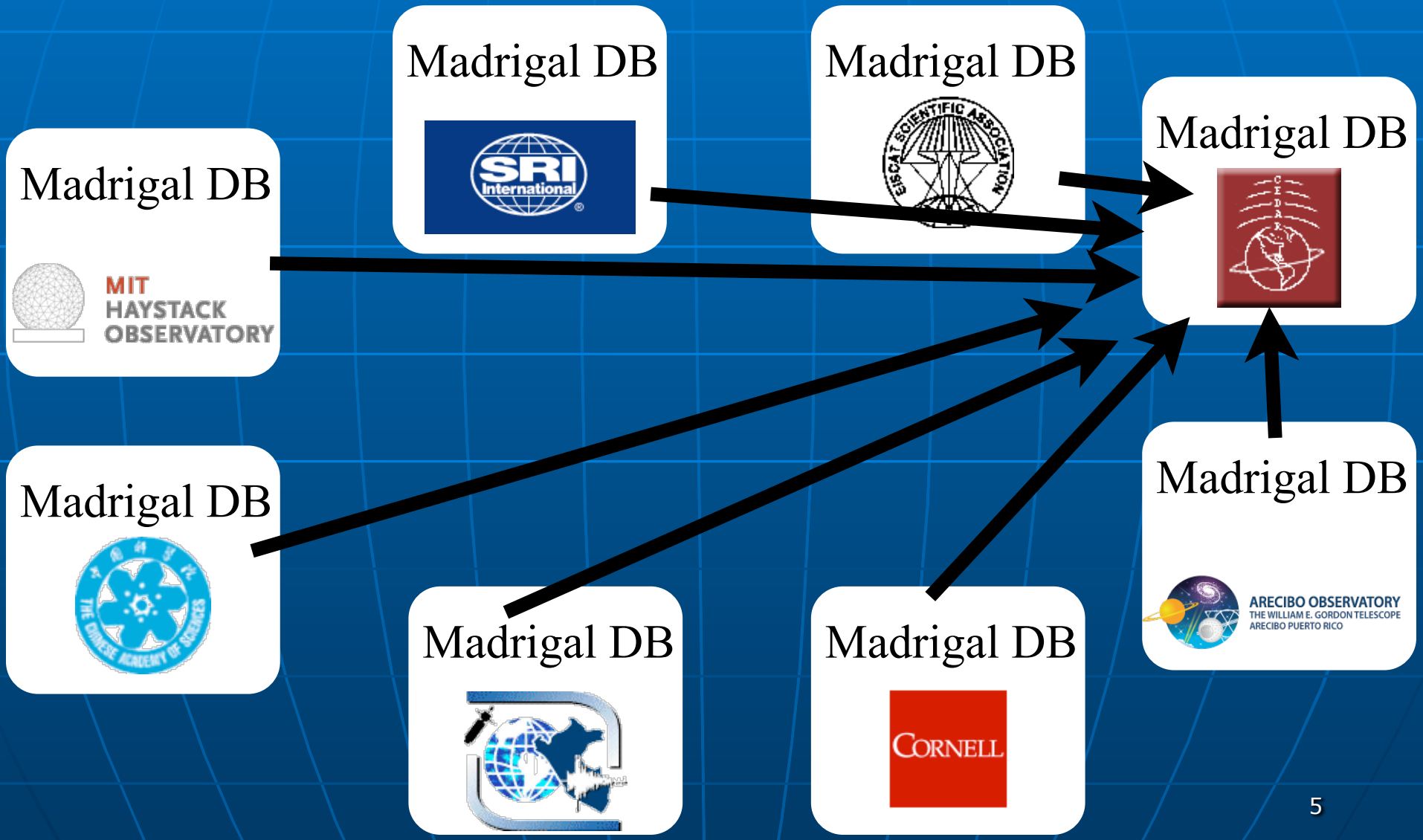
- What is Madrigal?
- Standardization of Madrigal data
- Standardization of Madrigal metadata
- Accessing data and metadata
- Lessons learned
  - Reproducibility of datasets
  - Attribution of datasets

# What is Madrigal?

# Madrigal is a distributed database with a focus on ground-based



# Cedar Madrigal archive imports all data weekly



# The Madrigal database stores data from a wide variety of upper atmosphere research instruments

Incoherent Scatter Radar

TEC via GPS

MF Radar

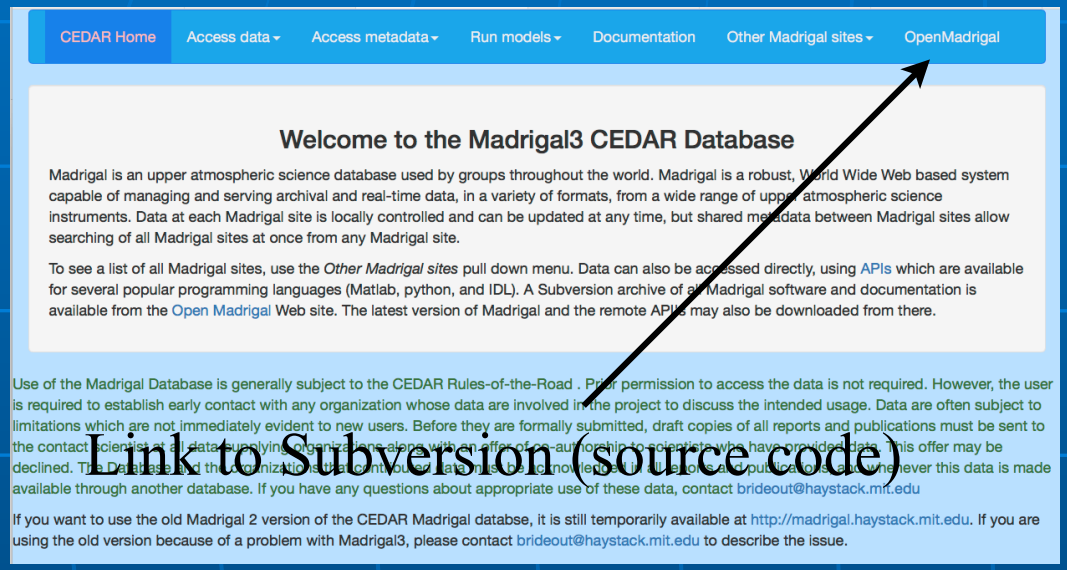
Examples of number of instruments in Madrigal:

- Incoherent scatter radars: 22
- MST radars: 3
- MF radars: 16
- Meteor radars: 11
- FPI: 32
- Michelson Interferometers: 6
- Lidars: 9
- Photometers: 7

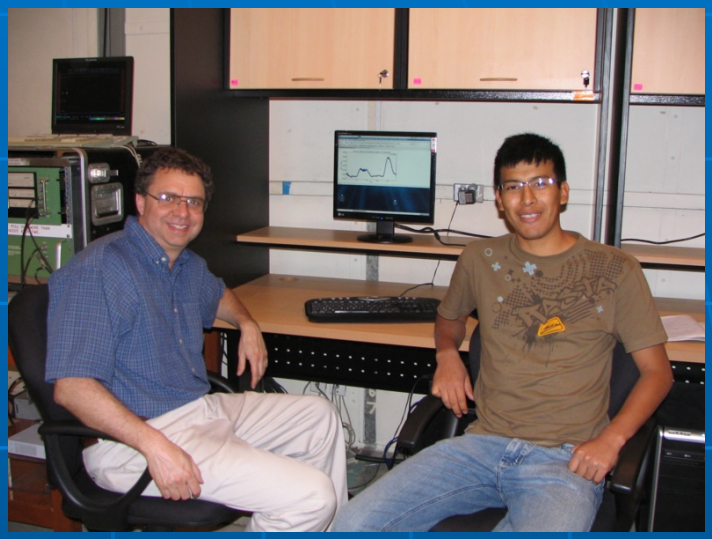
Other examples:

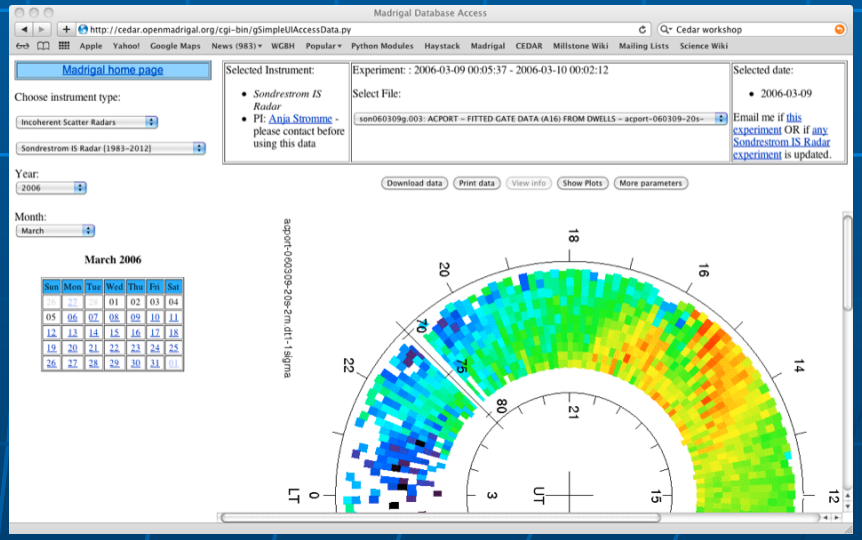
- GPS TEC
- DMSP

# Madrigal is open-source




Link to Subversion (source code)






www.openmadrigal.org

# Standardization of Madrigal data

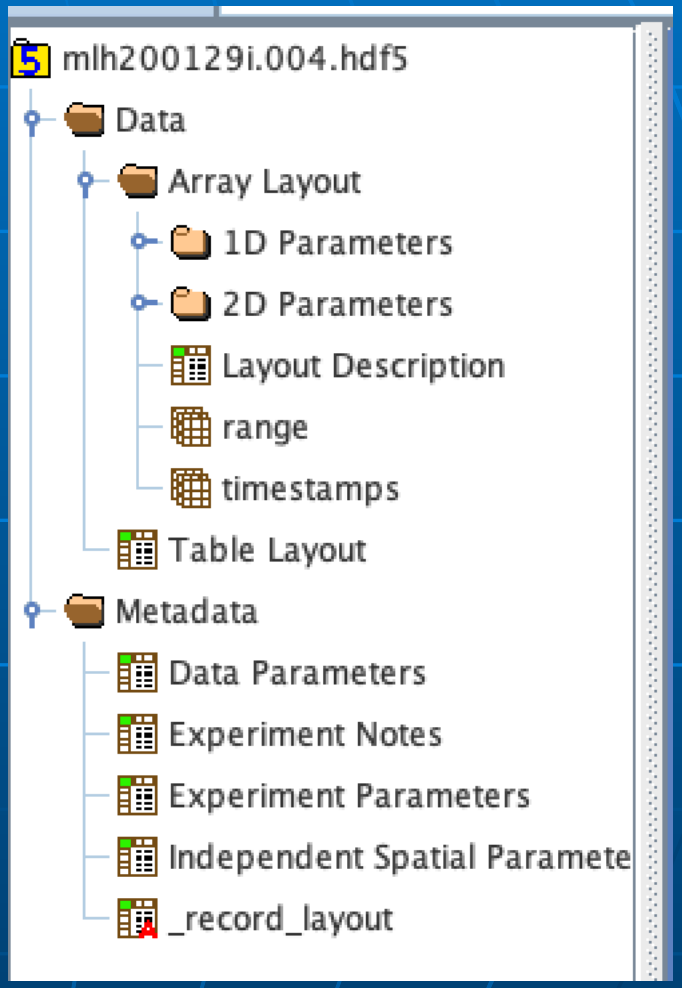
- Madrigal's underlying data format is well-defined Hdf5
  - Required data fields meant that the data should be fully self-defining
  - Standard parameters/units, definitions required in the file
  - Simple Python API for creating files
- Other export formats - Ascii, netCDF4

# Other features of Madrigal HDF5 standard

- Standardized parameters allows the existence of derivation engine
- All parameters have corresponding error parameters
- Independent parameters built into standard

# Cedar file format: Structured Hdf5

- Table Layout always given
- Array Layout default if independent parameters given (other than time)
- Array layout can be turned off if independent parameters non-repeating so that array would be too sparse.



- Standardization of Madrigal metadata

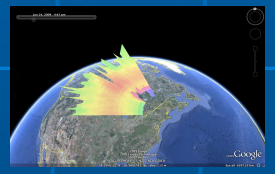
# Madrigal Metadata Model

**Madrigal site**
(typically a facility with scientists and a Madrigal installation)

Madrigal DB

**Instruments**
(ground-based, typically with a set location)

**Experiments**
(typically of limited duration, with a single contact)



Data shared among all Madrigal sites

**Experiment Files**
(represents data from one analysis of the experiment)

**Records**
(measurement over one period of time)

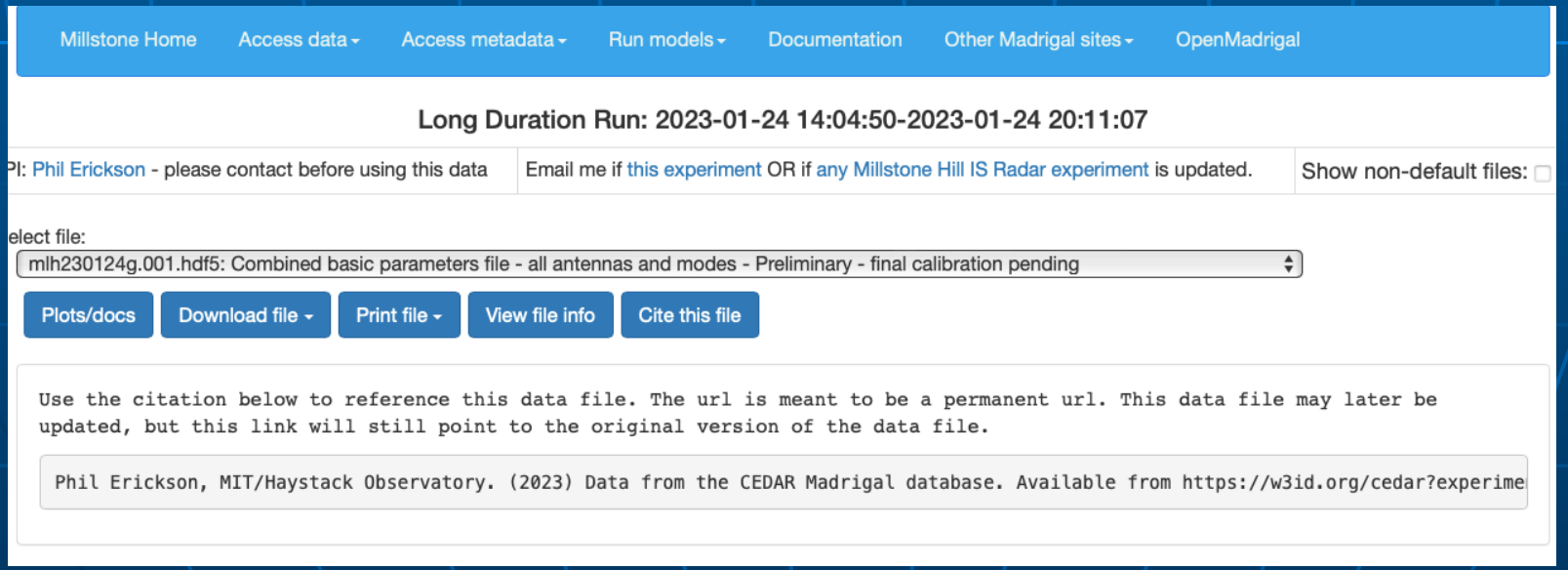
Data unique to one Madrigal site

# Madrigal metadata

- Internally stored as comma-delimited text
- Externally accessible via python API
  - API has been used to make Madrigal export other metadata standards
- Same API can be used to search for and access data

# Data reproducibility

- Madrigal files are never deleted
  - Newer files may make older files by marked as “history”
- Every file has citable url
- Python API allows creation of a citation to multiple files



# Data attribution

- PI's want to know when their data is used vs. free access to data
  - Madrigal has no passwords, but requires a name and email
  - Enforced by cookies in browser, arguments in API
- PI given access to Madrigal usage logs