MTH5: A Hierarchical Format for Magnetotelluric Data

J. R. Peacock[1]

[1]U.S. Geological Survey

December 13, 2018

Contents

1 Introduction

The magnetotelluric community is relatively small which has led to various formats for storing the time series data. Some type of ASCII format seems to be the most prevalent because before large data sets that was the easiest method of storage. However, in terms of read/write efficiency, ASCII is the slowest. Various binary formats exist, some proprietary and some open like the Scripps format, though efficient these files lack some critical metadata.

The most widely used format for archiving large data sets is the Hierarchical Data Format (HDF5). The HDF5 Group (https://www.hdfgroup.org/) maintains and updates the format as well as the software needed to read and write. The advantages of HDF5 are the metadata can be stored alongside the data, different components or calibration data or different schedules can be stored as separate folders, and the data is stored to the hard drive making reading and writing very efficient. There is also capability to access a single file from multiple different processors making it versatile for parallel computing. The goal for MTH5 is to develop a format where metadata can easily be stored and searched, as well has have a hierarchical structure where a single station can be stored in one HDF5 file. *Note: this is still in the development stage and any comments are welcome.* jpeacock@usgs.gov.

2 General Structure

The top level of a MTH5 file, the `root` directory, stores attributes important to the location of the data, how the data were collected, the provenance of the data, the software used to write the data, and copyright information on how the data can be used. These metadata are stored as JSON encoded strings. The metadata are split into the following headings **site**, **field_notes**, **copyright**, **provenance**, and **software**. These are described below. Off the `root` directory are the data folders. One for instrument calibrations and one for each different schedule action or sampling rate in which the data was collected. Note that this file structure assumes the data were collected with all the same setup, if the setup was changed a new file should be made.

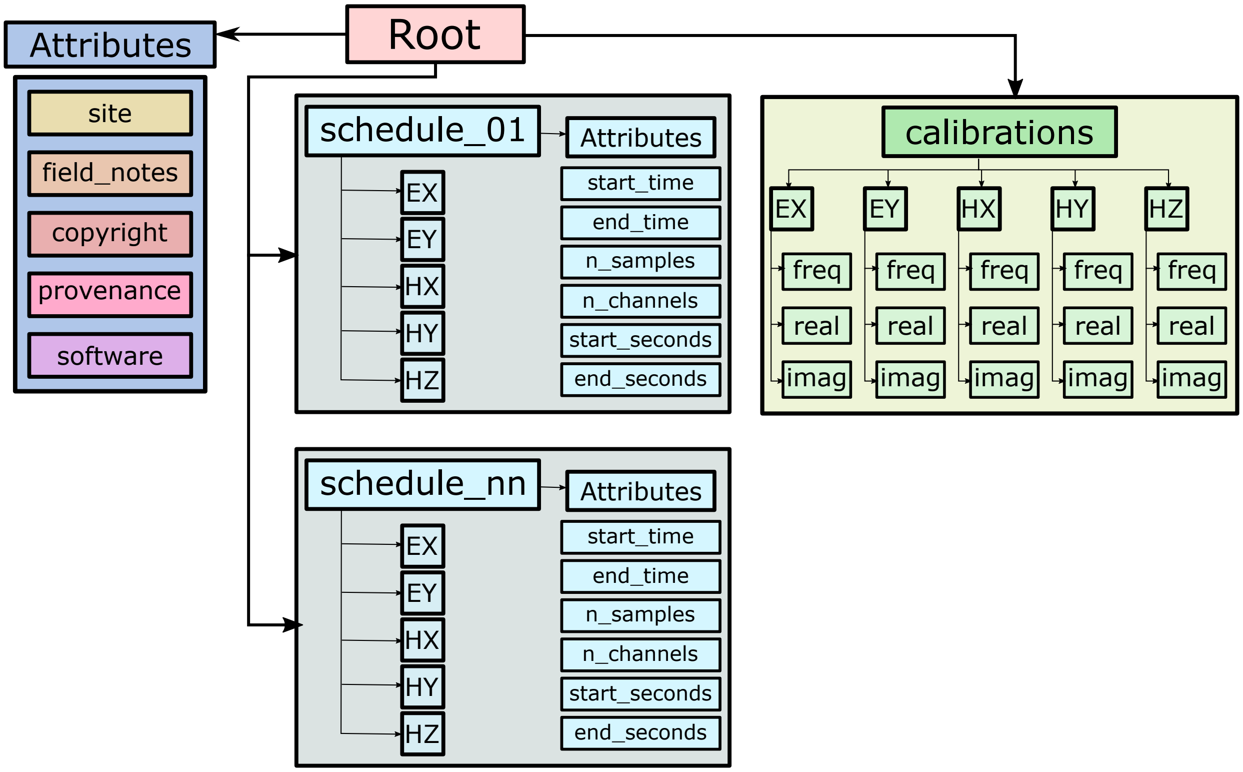

Figure 1: Schematic of how an MTH5 file is setup. See below for description of root attributes **site**, **field_notes**, **copyright**, **provenance**, and **software**.

2.1 Description of site

- **acquired_by**: who acquired the data
  - **email**: email of person responsible for the data
  - **name**: name of person responsible for the data
  - **organization**: organization name for person responsible for the data
  - **organization_url**: organization website for person responsible for the data
- **coordinate_system**: [Geographic North | Geomagnetic North | something else]
- **datum**: Datum that represents the location coordinates, (WGS84)
- **declination**: geomagnetic declination of station location
- **declination_epoch**: epoch from which declination is estimated
- **elev_units**: units of elevation [meters | feet | something else]
- **elevation**: elevation of station in **elev_units**
- **end_date**: date and time of when recording stopped[1].
- **id**: name of the station
- **latitude**: latitude of station[2]
- **longitude**: longitude of station[2]
- **start_date**: date and time of when recording began[1]
- **survey**: survey name and location

2.1.1 Example JSON encoded metadata for site

```
{"acquired_by": {"email": "generic@email.com",
                 "name": "John Doe",
                 "organization": "Company Name",
                 "organization_url": "www.company_name.com"},
"coordinate_system": "Geomagnetic North",
"datum": "WGS84",
"declination": 15.5,
"declination_epoch": 1995,
"elev_units": "meters",
"elevation": 1111.72,
"end_date": "2015-08-17T14:19:38.000000 UTC",
"id": "mshH020",
"latitude": 46.655559999999994,
"longitude": -121.48472,
"start_date": "2015-08-14T13:59:55.000000 UTC",
"survey": "MT survey Washington, USA"}
```

[1] The preferred format is YYYY-MM-DDThh:mm:ss.ms UTC

[2] Preferred format is decimal degrees

2.2 Description of field_notes

- **data_logger**: information about the data logger
  - **id**: ID number of data logger
  - **manufacturer**: Name of manufacturer
  - **type**: type of data logger
- **data_quality**: information about data quality
  - **author**: name of person assessing data quality
  - **comments**: comments on data quality
  - **rating**: [1–5] 1 – poor, 5 – great
  - **warnings_comments**: comments on any warnings flagged
  - **warnings_flags**: number of flags
- **electrode_xx**: information for electric fields EX and EY
  - **azimuth**: heading of dipole relative to **site.coordinate_system**
  - **chn_num**: channel number
  - **contact_resistance**: contact resistance in kOhms
  - **gain**: gain for electric channel
  - **id**: ID number of electrode(s)
  - **length**: dipole length in meters
  - **manufacturer**: electrode maker name
  - **type**: type of electrode
  - **units**: units of electric field data
- **magnetoteter_xx**: information for magnetic fields HX, HY, HZ
  - **azimuth**: heading of magnetotmeter relative to **site.coordinate_system**
  - **chn_num**: channel number
  - **gain**: gain for electric channel
  - **id**: ID number of electrode(s)
  - **manufacturer**: electrode maker name
  - **type**: type of electrode
  - **units**: units of electric field data

2.2.1 Example JSON encoded metadata for field_notes

```
{"data_logger": {"id": "ZEN18",
                 "manufacturer": "Zonge",
                 "type": "32-Bit 5-channel GPS synced"},
"data_quality": {"author": "C. Cagniard",
                 "comments": "testing",
                 "rating": 5,
                 "warnings_comments": "bad data at 2018-06-07T20:10:00.00",
                 "warnings_flag": 1},
"electrode_ex": {"azimuth": 15.5,
                 "chn_num": 4,
                 "contact_resistance": 1,
                 "gain": 1,
                 "id": 1,
                 "length": 100.0,
                 "manufacturer": "Borin",
                 "type": "Fat Cat Ag-AgCl",
                 "units": "mV"},
"electrode_ey": {"azimuth": 105.5,
                 "chn_num": 5,
                 "contact_resistance": 1,
                 "gain": 1,
                 "id": 2,
                 "length": 92.0,
                 "manufacturer": "Borin",
                 "type": "Fat Cat Ag-AgCl",
                 "units": "mV"},
"magnetometer_hx": {"azimuth": 15.5,
                    "chn_num": 1,
                    "gain": 1,
                    "id": 2374,
                    "manufacturer": "Geotell",
                    "type": "Ant 4 Induction Coil",
                    "units": "mV"},
"magnetometer_hy": {"azimuth": 105.5,
                    "chn_num": 2,
                    "gain": 1,
                    "id": 2384,
                    "manufacturer": "Geotell",
                    "type": "Ant 4 Induction Coil",
                    "units": "mV"},
"magnetometer_hz": {"azimuth": 90,
                    "chn_num": 3,
                    "gain": 1,
                    "id": 2514,
                    "manufacturer": "Geotell",
                    "type": "Ant 4 Induction Coil",
                    "units": "mV"}}
```

2.3 Description of copyright

- **additional_info**: any additional information about copyright
- **citation**: Citation that uses this data
  - **author**: citation author
  - **doi**: doi number of citation
  - **journal**: name of publisher
  - **title**: citation title
  - **volume**: citation volume
  - **year**: citation year
- **conditions_of_use**: Conditions of using the data
- **release_status**: status of the data release

2.3.1 Example JSON encoded metadata for copyright

```
{"additional_info": "this is a test",
 "citation": {"author": "Tikhanov",
              "doi": "10.1023/usgs_mt_test",
              "journal": "SI",
              "title": "MT HDF5 test",
              "volume": 1,
              "year": 2018},
 "conditions_of_use": "All data and metadata for this survey are available free of charge
                       and may be copied freely, duplicated and further distributed provided
                       this data set is cited as the reference. While the author(s) strive to
                       provide data and metadata of best possible quality, neither the author(s)
                       of this data set, not IRIS make any claims, promises, or guarantees about
                       the accuracy, completeness, or adequacy of this information, and expressly
                       disclaim liability for errors and omissions in the contents of this file.
                       Guidelines about the quality or limitations of the data and metadata, as
                       obtained from the author(s), are included for informational purposes
                       only.",
 "release_status": "Open to the public"}
```

2.4 Description of provenance

- **creating_application**: software used to create the file
- **creation_time**: creation date and time of file
- **creator**: information about creator
  - **email**: email of person who created file
  - **name**: name of person who created file
  - **organization**: organization of person who created file
  - **organization_url**: organization URL of person who created file
- **submitter**: information about submitter
  - **email**: email of person who submitted file
  - **name**: name of person who submitted file
  - **organization**: organization of person who submitted file
  - **organization_url**: organization URL of person who submitted file

2.4.1 Example JSON encoded metadata for provenance

```
{"creating_application": "MTH5py",
 "creation_time": "2017-11-27T21:54:49.00",
 "creator": {"email": "test@email.com",
             "name": "author",
             "organization": "company name",
             "organization_url": "https://www.company_name.com"},
 "submitter": {"email": "test@email.com",
               "name": "author",
               "organization": "company name",
               "organization_url": "https://www.company.com"}}
```

2.5 Description of software

- **author**: information about the author of the software used to make file
  - **email**: email of software author
  - **name**: name of software author
  - **organization**: organization of software author
  - **organization_url**: organization URL software author
- **name**: software name that made file
- **version**: software version

2.5.1 Example JSON encoded metadata for provenance

```
{"author": {"email": "send@email.com",
            "name": "author",
            "organization": "company name",
            "organization_url": "https://company_name.com"},
 "name": "MTH5py",
 "version": "Beta"}
```

3 Calibrations

Calibrations are put in a dedicated folder where each instrument calibration is put in a sub-folder. Each instrument calibration folder can have metadata attributes. These should contain information about units, date of calibration, and who did the calibration. Typically the magnetometer calibrations are given as a complex value as a function of frequency or period. These data are subsequently put in their own folder.

Attributes should include:

- **instrument_id**: ID number of instrument
- **calibration_date**: date of calibration[1]
- **calibration_person**:
  - **email**: email of person who did calibration
  - **name**: name of person who did calibration
  - **organization**: organization of person who did calibration
  - **organization_url**: organization URL of person who did calibration
- **units**: units of calibration

[1] The preferred format is YYYY-MM-DD

3.0.1 Example JSON encoded metadata for calibration

```
{"calibration_date": "2010-10-01",
 "calibration_person": {"email": "test@email.com",
                        "name": "test",
                        "organization": "house",
                        "organization_url": "www.house.com"},
 "instrument_id": 2284,
 "name": "hx",
 "units": "mV/nT"}
```

4 Schedule Blocks

Commonly MT measurements are recorded at different sampling rates over the course of a measurement to get the full range of data. Or sometimes data recording may be interrupted and restarted. The different recordings are labeled `schedule_xx` where the `xx` represents order of the measurement [0–n], labeled as a 2 character number, with a leading 0 if n is less than 10. The first schedule is `schedule_01` and the nth schedule is `schedule_nn`.

Should the attributes be converted to JSON or left as is?

The `schedule_nn` folder has attributes attached to it. These are:

- **start_time**: date and time data recording started[1]
- **end_time**: date and time data recording ended[1]
- **start_seconds**: seconds from the epoch (1970-01-01) data recording started
- **end_seconds**: seconds from the epoch (1970-01-01) data recording ended
- **sampling_rate**: sampling rate in samples/second
- **n_samples**: number of samples in each channel
- **n_channels**: number of channels recorded

[1] The preferred format is YYYY-MM-DDThh:mm:ss.ms UTC