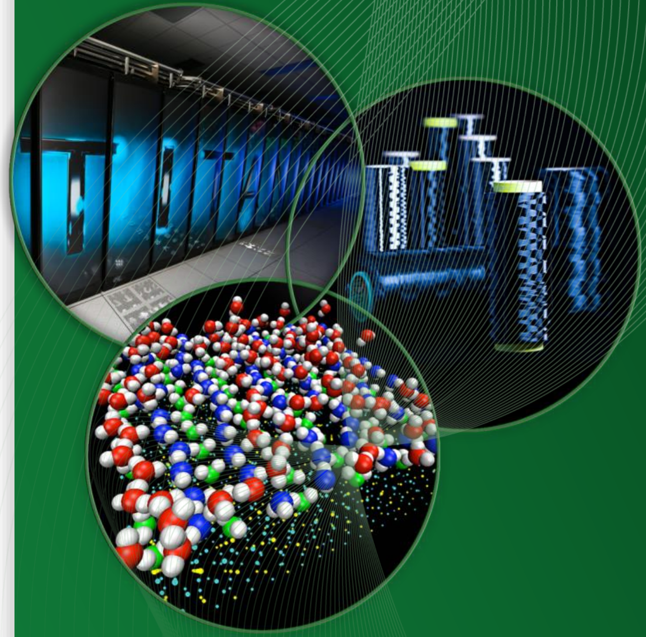
SPM Data Formatting

Stephen Jesse, Suhas Somnath, Chris Smith

01/30/2017



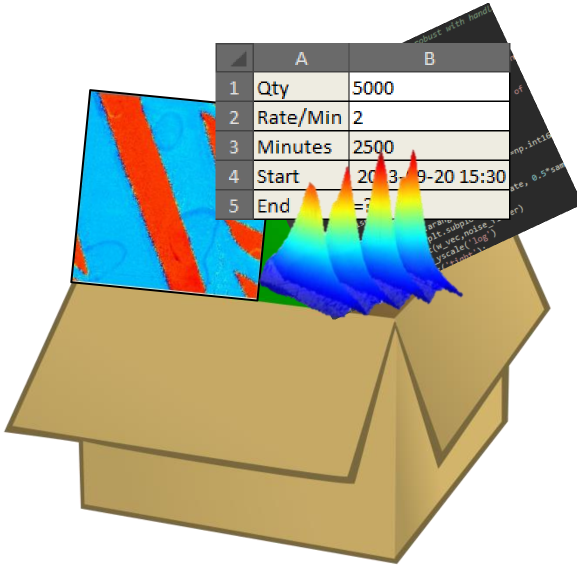
ORNL is managed by UT-Battelle
for the US Department of Energy

INSTITUTE FOR FUNCTIONAL IMAGING OF MATERIALS
OAK RIDGE NATIONAL LABORATORY

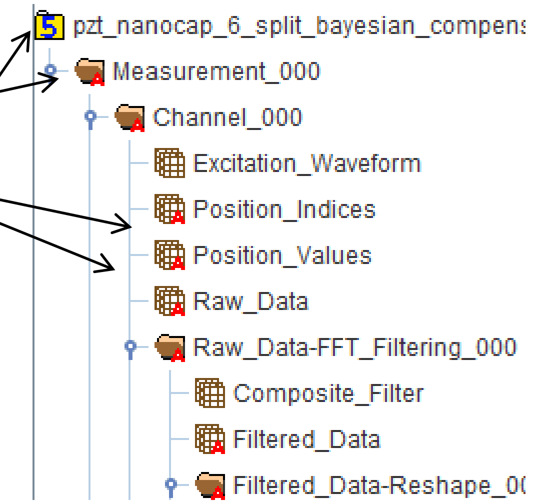
SPM Data Formatting and Analysis Efforts at CNMS

- Data Format:
  - Data stored in HDF5 format
    - Open source format. Hierarchical Data Format v5 (HDF5)
  - Same standard for different microscopes and microscopy methods
  - Works for standard imaging and spectroscopic measurements as well
    - Conventional AFM images would have a single data point at each spatial location. Eg – tapping mode imaging, contact mode imaging
    - Spectral images : One or more parameters systematically varied over a range of values at different spatial positions in a grid or cloud of points. Example – band excitation data

What is an HDF5 File?



- An HDF5 file is a smart container
  - Hierarchical / tree structure
  - Capable of storing:
    - Multidimensional datasets
    - Images
    - text
  - Contents organized like traditional folders and files
  - Important components:
    - **Datagroup** - Analogous to folders in a file system
    - **Dataset** – contains 1 to N dimensional data
      - Integer, floating point, complex numbers etc
    - **Attributes** – Key : value pairs that contain information to describe the data. Eg – units.
    - **References** – Analogous to shortcuts / links



pzt_nanocap_6_split_bayesian_compens
Measurement_000
Channel_000
Excitation_Waveform
Position_Indices
Position_Values
Raw_Data
Raw_Data-FFT_Filtering_000
Composite_Filter
Filtered_Data
Filtered_Data-Reshape_00

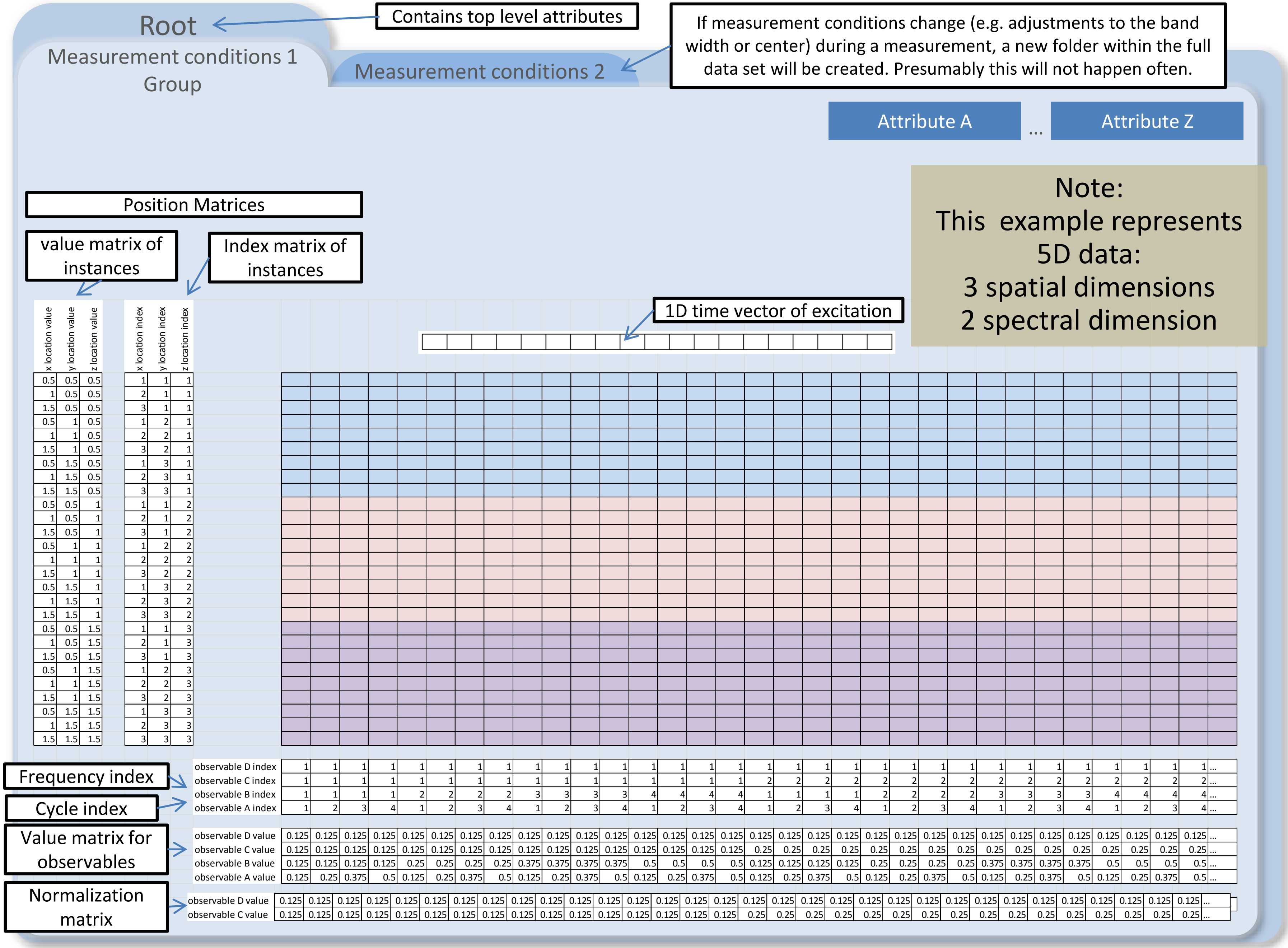
See here for more information on HDF5:
http://extremecomputingtraining.anl.gov/files/2015/03/HDF5-Intro-aug7-130.pdf

Data Formatting – Spectral Images

All data, regardless of dimension, is laid out into a 2D array

- The first dimension for location index
- The second dimension for spectral index
- Keys provide instructions on original data dimensionality
- Allows for irregular position arrays (cloud of points)
- Allows for irregular spectral measurements (set-pulses,…)
- Keeps track measurement sequence
- Format matches what PCA, ICA, etc. expects

HDF5 File Format for BE: Ver. 4



Root

Contains top level attributes

Measurement conditions 1 Group

Measurement conditions 2

If measurement conditions change (e.g. adjustments to the band width or center) during a measurement, a new folder within the full data set will be created. Presumably this will not happen often.

Attribute A ... Attribute Z

Note:
This example represents 5D data:
3 spatial dimensions
2 spectral dimension

Position Matrices

value matrix of instances

Index matrix of instances

1D time vector of excitation

x location value	y location value	z location value
0.5	0.5	0.5
1	0.5	0.5
1.5	0.5	0.5
0.5	1	0.5
1	1	0.5
1.5	1	0.5
0.5	1.5	0.5
1	1.5	0.5
1.5	1.5	0.5
0.5	0.5	1
1	0.5	1
1.5	0.5	1
0.5	1	1
1	1	1
1.5	1	1
0.5	1.5	1
1	1.5	1
1.5	1.5	1
0.5	0.5	1.5
1	0.5	1.5
1.5	0.5	1.5
0.5	1	1.5
1	1	1.5
1.5	1	1.5
0.5	1.5	1.5
1	1.5	1.5
1.5	1.5	1.5

x location index	y location index	z location index
1	1	1
2	1	1
3	1	1
1	2	1
2	2	1
3	2	1
1	3	1
2	3	1
3	3	1
1	1	2
2	1	2
3	1	2
1	2	2
2	2	2
3	2	2
1	3	2
2	3	2
3	3	2
1	1	3
2	1	3
3	1	3
1	2	3
2	2	3
3	2	3
1	3	3
2	3	3
3	3	3

Frequency index

Cycle index

observable D index	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1 ...
observable C index	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2 ...
observable B index	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4 ...
observable A index	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4 ...

Value matrix for observables

observable D value	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125 ...
observable C value	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25 ...
observable B value	0.125	0.125	0.125	0.125	0.25	0.25	0.25	0.25	0.375	0.375	0.375	0.375	0.5	0.5	0.5	0.5	0.125	0.125	0.125	0.125	0.25	0.25	0.25	0.25	0.375	0.375	0.375	0.375	0.5	0.5	0.5	0.5 ...
observable A value	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5	0.125	0.25	0.375	0.5 ...

Normalization matrix

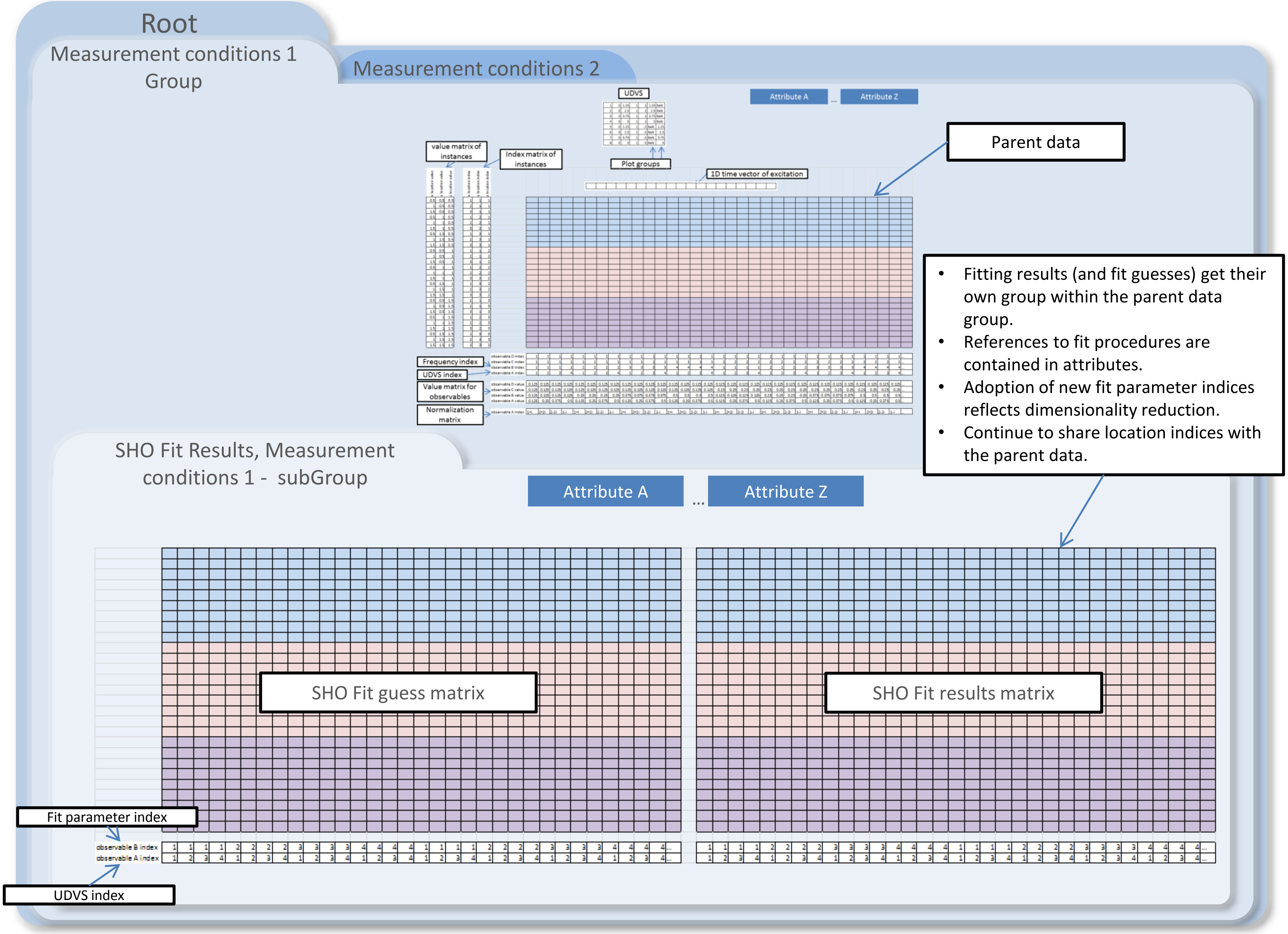
observable D value	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125 ...
observable C value	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25 ...

Data Formatting – Spectral Images

All data, regardless of dimension, is laid out into a 2D array

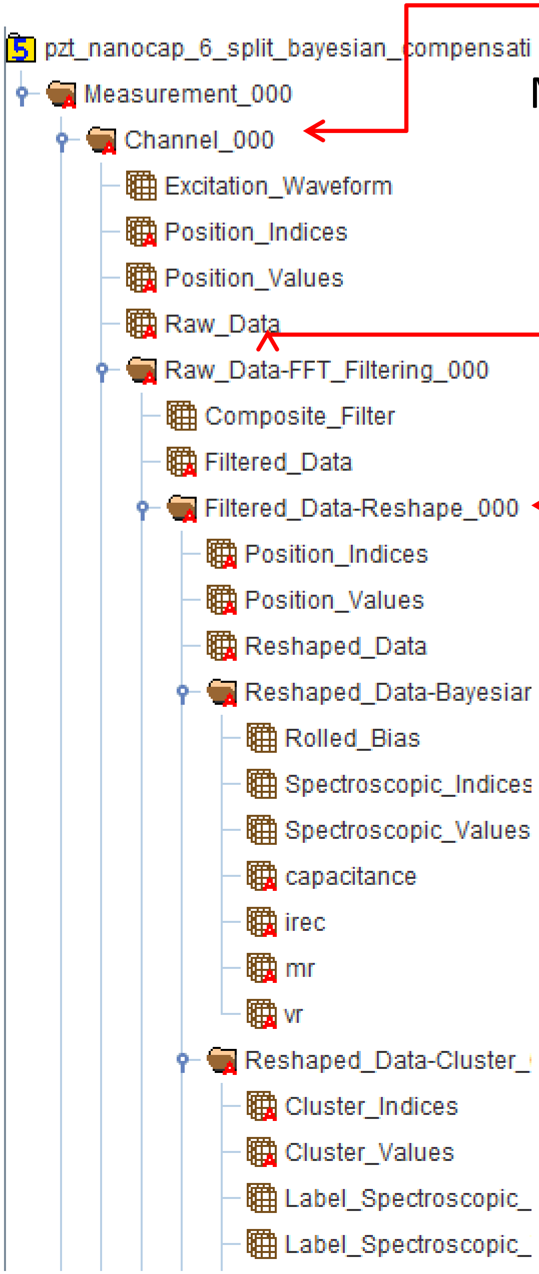
- The first dimension for location index
- The second dimension for spectral index
- Keys provide instructions on original data dimensionality
- Allows for irregular position arrays (cloud of points)
- Allows for irregular spectral measurements (set-pulses,…)
- Keeps track measurement sequence
- Format matches what PCA, ICA, etc. expects
- Subsequent analysis, such as fitting, is contained in ‘subfolders’
- of the parent data within the HDF5 structure

HDF5 File Format for BE: Ver. 4



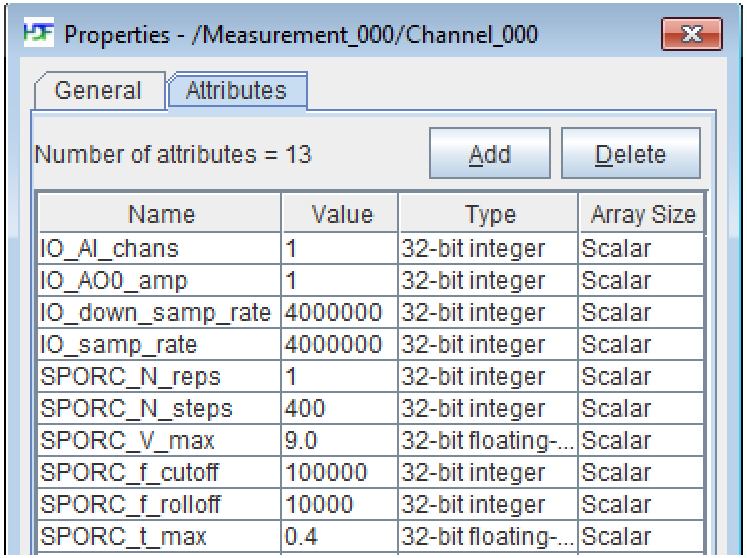
- Fitting results (and fit guesses) get their own group within the parent data group.
- References to fit procedures are contained in attributes.
- Adoption of new fit parameter indices reflects dimensionality reduction.
- Continue to share location indices with the parent data.

Example Dataset



- pzt_nanocap_6_split_bayesian_compensati
  - Measurement_000
    - Channel_000
      - Excitation_Waveform
      - Position_Indices
      - Position_Values
      - Raw_Data
      - Raw_Data-FFT_Filtering_000
        - Composite_Filter
        - Filtered_Data
        - Filtered_Data-Reshape_000
          - Position_Indices
          - Position_Values
          - Reshaped_Data
          - Reshaped_Data-Bayesiar
            - Rolled_Bias
            - Spectroscopic_Indices
            - Spectroscopic_Values
            - capacitance
            - irec
            - mr
            - vr
          - Reshaped_Data-Cluster_
            - Cluster_Indices
            - Cluster_Values
            - Label_Spectroscopic_
            - Label_Spectroscopic_

Metadata stored as attributes

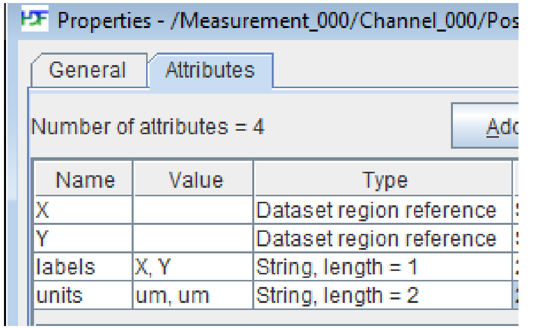


Properties - /Measurement_000/Channel_000

Number of attributes = 13

Name	Value	Type	Array Size
IO_AI_chans	1	32-bit integer	Scalar
IO_AO0_amp	1	32-bit integer	Scalar
IO_down_samp_rate	4000000	32-bit integer	Scalar
IO_samp_rate	4000000	32-bit integer	Scalar
SPORC_N_reps	1	32-bit integer	Scalar
SPORC_N_steps	400	32-bit integer	Scalar
SPORC_V_max	9.0	32-bit floating-...	Scalar
SPORC_f_cutoff	100000	32-bit integer	Scalar
SPORC_f_rolloff	10000	32-bit integer	Scalar
SPORC_t_max	0.4	32-bit floating-...	Scalar

Data columns / rows can be accessed by name instead of indices



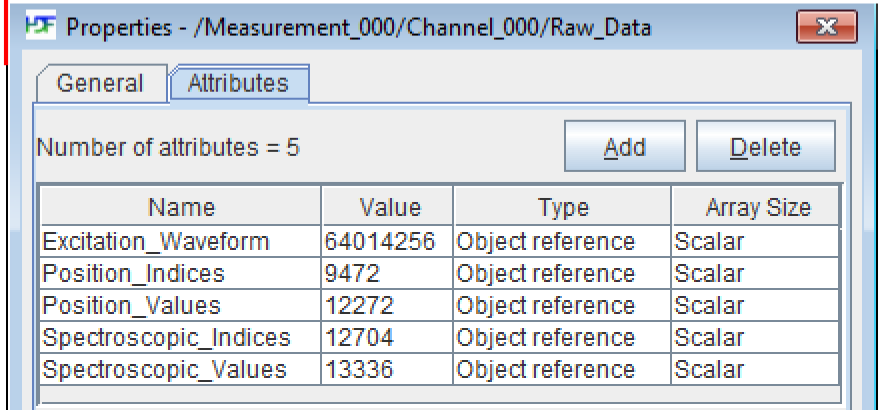
Properties - /Measurement_000/Channel_000/Pos

Number of attributes = 4

Name	Value	Type
X		Dataset region reference
Y		Dataset region reference
labels	X, Y	String, length = 1
units	um, um	String, length = 2

Nomenclature of datagroups provides simple way to understand sequence of steps applied to process data

Datasets link to relevant ancillary data



Properties - /Measurement_000/Channel_000/Raw_Data

Number of attributes = 5

Name	Value	Type	Array Size
Excitation_Waveform	64014256	Object reference	Scalar
Position_Indices	9472	Object reference	Scalar
Position_Values	12272	Object reference	Scalar
Spectroscopic_Indices	12704	Object reference	Scalar
Spectroscopic_Values	13336	Object reference	Scalar

Generic SPM File Rules – Mandatory Datasets

- <Root>
  - **Measurement_000** (new measurement group each time parameters change)
    - **Channel_000** (one for each physical channel like deflection, lateral..)
      - **Raw_Data** (positions x time or spectroscopic values)
        - type : uint8, float32, complex64 etc.
        - Required attributes:
          - References to all ancillary datasets below
          - **Quantity** - Single string that explains the data – eg – Current or Voltage
          - **units** – list of a single string for units like nA, V, F, etc.
      - **Position_Indices** (positions x spatial dimensions)
        - type : uint32
        - Required attributes:
          - **labels** - list of strings for the column names
          - Region references based on column names
      - **Position_Values** (positions x spatial dimensions)
        - type : float32
        - Required attributes:
          - **labels** - list of strings for the column names
          - **units** – list of strings for units like nm / um
          - Region references based on column names
      - **Spectroscopic_Indices** (spectroscopic parameter x spectroscopic indices)
        - type : uint32
        - Required attributes:
          - **labels** - list of strings for the row names
          - Region references based on row names
      - **Spectroscopic_Values** (spectroscopic parameter x spectroscopic values)
        - type : (at least) float32 or complex64
        - Required attributes:
          - **labels** - list of strings for the row names
          - **units** – list of strings for units like mV / rad / sec
          - Region references based on row names
  - **Measurement_001…**

Legend:

- **Dataset**
- **Datagroup**
- **Attribute**

Additional datasets, data groups, and attributes can be added as necessary depending on the measurement

Generic SPM File Rules – Attributes for Root

- <Root>:
  - comments = ’10X amplifier used’
  - data_tool = ‘be_analyzer’
  - **data_type = ‘BELine’ ← mandatory – used for reading data**
  - experiment_date = 2015_10_15-14_55_05
  - experiment_unix_time = 1.35654765E+9
  - microscope = ‘Asylum Research Cypher’
  - instrument = ‘Cypher West CNMS’
  - project_id = ‘CNMS_2015B_X0252’
  - project_name = ‘HfO2 investigation’
  - sample_description = ‘8 nm HfO2 with 300um2 capacitors’
  - sample_Name = ‘HFO2’
  - translate_date = 2015_10_15-14_55_05
  - translator = ‘ODF’
  - user_name = ‘John Doe’
  - xcams_id = ‘jdoe’

Incorporating units:

attribute_name_[unit] = Value
read_voltage_[V] = 3.9

Time stamp:

YYYY_MM_DD-HH_mm_ss
24 hour format for hours

Nomenclature for Processing Tools

Analysis tools include function fitting, multivariate analysis functions etc. while processing tools include signal / image filtering, flattening functions etc.

General Rule

- DatasetName
- DatasetName-ToolName_00x
  - time_stamp
  - machine_id
  - tool_name
  - algorithm
  - Other relevant attributes
  - ToolResult0
    - ~~Reference to DatasetName~~
    - Reference to mapping matrices (position / spectroscopic) for unpacking and/or plotting
    - labels
    - units
  - ToolResult1
    - .....

Current methodology facilitates:

- Same tool (with different parameters) to be applied to same dataset (different suffixes)
- Tracing of all processing applied to any given dataset (using paths)

Example -> Chain of analysis tools (SVD and kMeans)

Legend:

- Dataset
- Datagroup
- Attribute

- Raw_Data
- Raw_Data-SVD_000
  - <SVD Attributes>
  - S
    - <Relevant references>
  - U
    - <Relevant references>
  - V
    - <Relevant references>
  - U-Cluster_000
    - Type = ‘KMeans’
    - Labels
      - <Relevant references>
    - Mean_Response
      - <Relevant references>
- Raw_Data-SVD_001
  - S
  - U…

Example Rules for Processing Tool – Singular Value Decomposition (SVD)

- Raw_Data
- Raw_Data-SVD_000
  - time_stamp
  - machine_id
  - tool_name
  - algorithm
  - S
  - Component_Indices
  - U
    - Reference to Position_Values from attribute of Raw_Data
    - Reference to Position_Indices from attribute of Raw_Data
    - Reference to Component_Indices named as ‘Spectroscopic_Indices’
    - Reference to S named as ‘Spectroscopic_Values’
    - labels
    - units
  - V
    - Reference to Component_Indices named as ‘Position_Indices’
    - Reference to S named as ‘Position_Values’
    - Reference to Spectroscopic_Values from attribute of Raw_Data
    - Reference to Spectroscopic_Indices from attribute of Raw_Data
    - labels
    - units

Do NOT store references to source dataset – Should the user want to only export a certain analysis / processing result (group), all the references within the group will also be copied over.

Legend:
- Dataset
- Datagroup
- Attribute