Data Management and Open Access
Creating Data Files for Published Figures

Josh Stillerman, Martin Greenwald, Mark London, Jason Thomas
February, 2016

Publishing Data for Figures

- The DOE requirement is not specific about exactly which data and metadata must be included with published figures.
  - We are interpreting the requirement to be:
    - The actual values plotted in the figure
    - Metadata about those values
      - Name, Description, Units
    - Metadata about how the data are displayed in the figure
      - Labels, Display Parameters
  - They are are also not dictating how the data should be stored.
    - File Format / Data Organization …

PSFC Standardized Format

- Choosing a standard file format has several advantages:
  - Easier access for readers of the publication
  - Easier verification for librarians, curators, and sponsors
  - Slower obsolescence, and easier conversion as standards evolve
  - Standard general purpose tools for browsing and viewing contents.
- We have chosen HDF5
  - https://www.hdfgroup.org/HDF5/

PSFC Standardized Schema

- Using a standard file format is good, but not good enough.
  - If all of the data files for figures in PSFC publications were for example MS Excel
    - This would not dictate the organization of labels, rows and columns in those spreadsheets.
    - In order to interpret one of them a user would have to open the file interactively and attempt to understand the organization.
  - The same is true for HDF5, so
    - We have defined a standard HDF5 file organization to represent the data in published figures.
    - Easy access for all consumers (since they are all the same in structure)
    - Easy to creation from the programing languages in use at the PSFC.
      - IDL
      - PYTHON
      - MATLAB
      - This list can be expanded as needed.

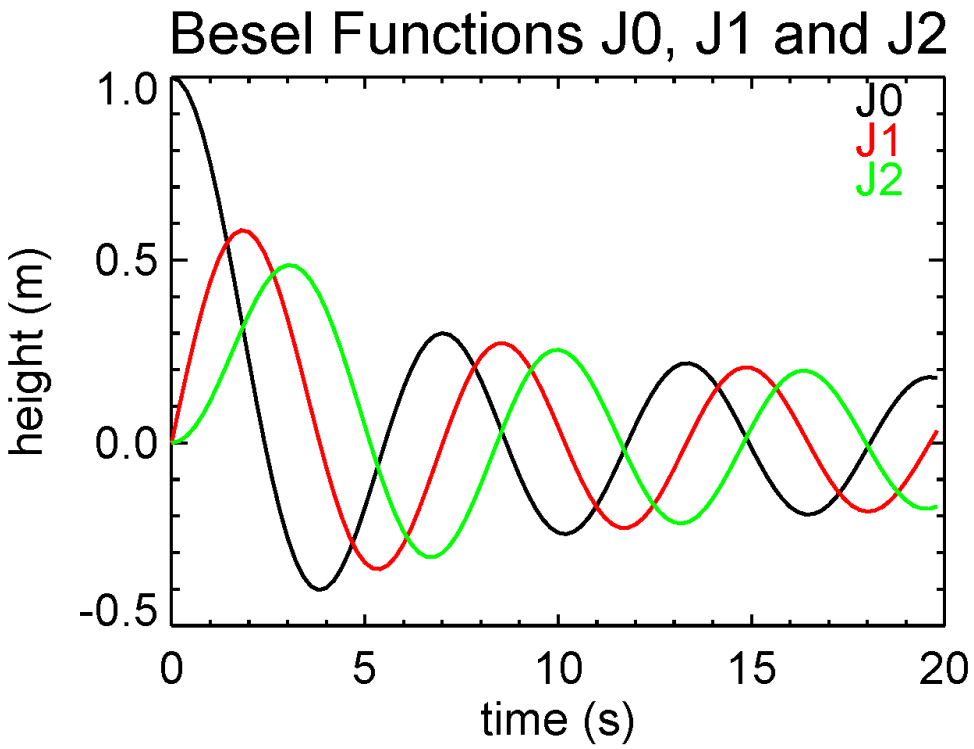
PSFC Standardized Schema (2)

- One file per figure - the library system will name the file based on the publication's ID
  - Root level attributes: author, username, date, description, caption…
  - One Group per 'trace' displayed.
    - Group level attributes for this trace:
- One Group per set of data displayed
  - Group level attributes: name, legend string, plot-information
  - x_data – values for the X axis
    - Units, label
  - Y_data – values for the Y axis
    - Units, label
  - Z_data – values for the Z axis
    - Units, label

Creating data files

- The time to create (or update) the data files is when the figures are being created
  - At that time, all of the data is available in some programming language.
  - It is much more likely the file will match the figure, if it is created at that time.
- APIs are set up to mimic the plotting APIs.
- Files can be created and consumed in any programming language interchangeably
- Example in IDL
- Example in Python
- Other languages to follow

IDL - The figure

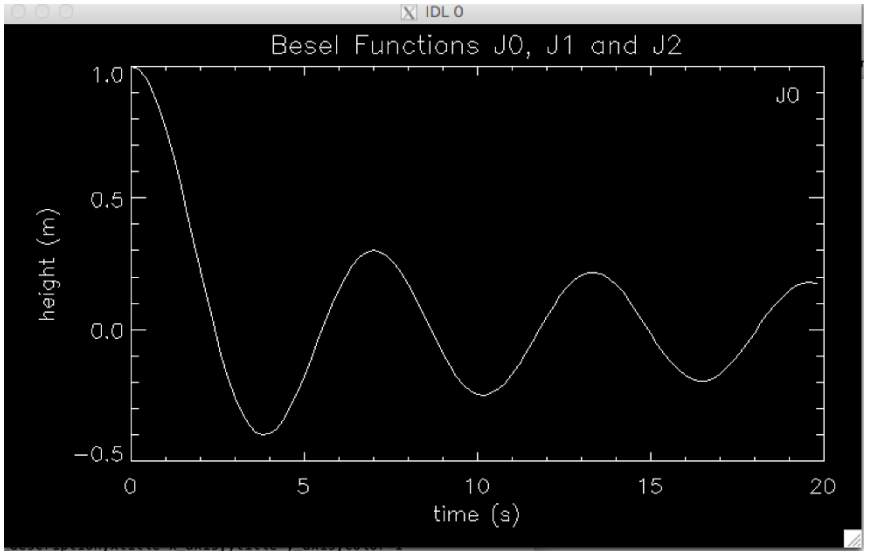


IDL

```
file = 'Fig_1'
fig_description = 'Besel Functions J0, J1 and J2'
fig_source = 'Phys. Plasmas 17, 1234 2010'
comment = 'This is the way the ball bounces'
user_fullname = 'John Doe'
date = systime(0)

;set up a simple color table (just for plotting)
r = [000,255,255,000,000]
g = [000,255,000,000,255]
b = [000,255,000,255,000]
tvlct,r,g,b

;start a new hdf5 file
hdf5_new, file=file, fig_description=fig_description, fig_source=fig_source, $
        comment=comment, user_fullname=user_fullname,date=date
```

IDL(2)

```
x_units = 's'
x_axis = 'time (s)'
x_name = 'measured with a stopwatch'
x_type = 'float'

y_units = 'm'
y_axis = 'height (m)'
y_name = 'measured with a ruler'
y_type = 'float'

legend = 'J0'
```

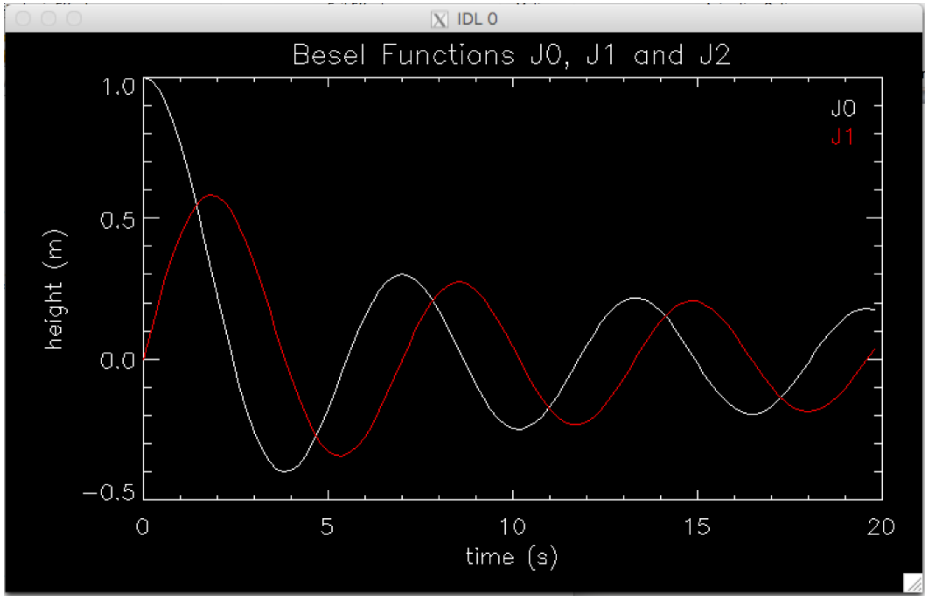



```
;compute and plot the first curve (you'll do this to create the plot file)
x = indgen(100)/5.
y0 = beselj(x,0)
plot,x,y0,charsize=1.8,title=fig_description,xtitle=x_axis,ytitle=y_axis,color=1
xyouts,/norm,.9,.85,legend,size=1.8
```

```
hdf5_add, x,y0,file=file,group_name=group_name,$
        x_units=x_units,x_axis=x_axis, x_name=x_name,x_type=x_type,$
        y_units=y_units, y_axis=y_axis,y_name=y_name,y_type=y_type,$
        legend=legend, plot_graphics=plot_graphics
```

IDL(3)

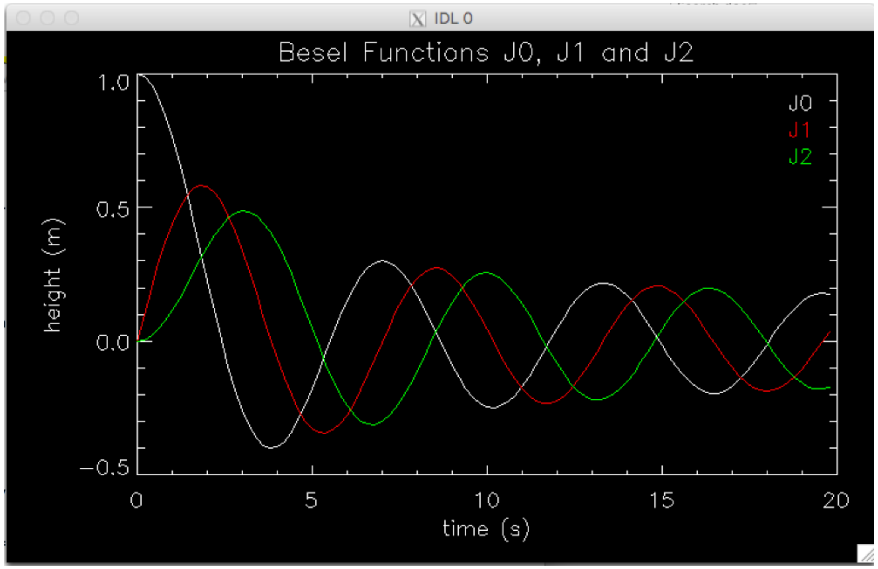
```
legend = 'J1'

y1 = beselj(x,1)
oplot,x,y1,color=2
xyouts,/norm,.9,.8,legend,size=1.8,color=2

group_name = legend
plot_graphics = 'red line'
```




```
hdf5_add, x,y1,file=file,group_name=group_name,$
        x_units=x_units,x_axis=x_axis, x_name=x_name,x_type=x_type,$
        y_units=y_units, y_axis=y_axis,y_name=y_name,y_type=y_type,$
        legend=legend, plot_graphics=plot_graphics
```

IDL(4)

```
legend = 'J2'

;compute and plot the third curve
y2 = beselj(x,2)
oplot,x,y2,color=4
xyouts,/norm,.9,.75,legend,size=1.8,color=4

group_name = legend
plot_graphics = 'green line'
```

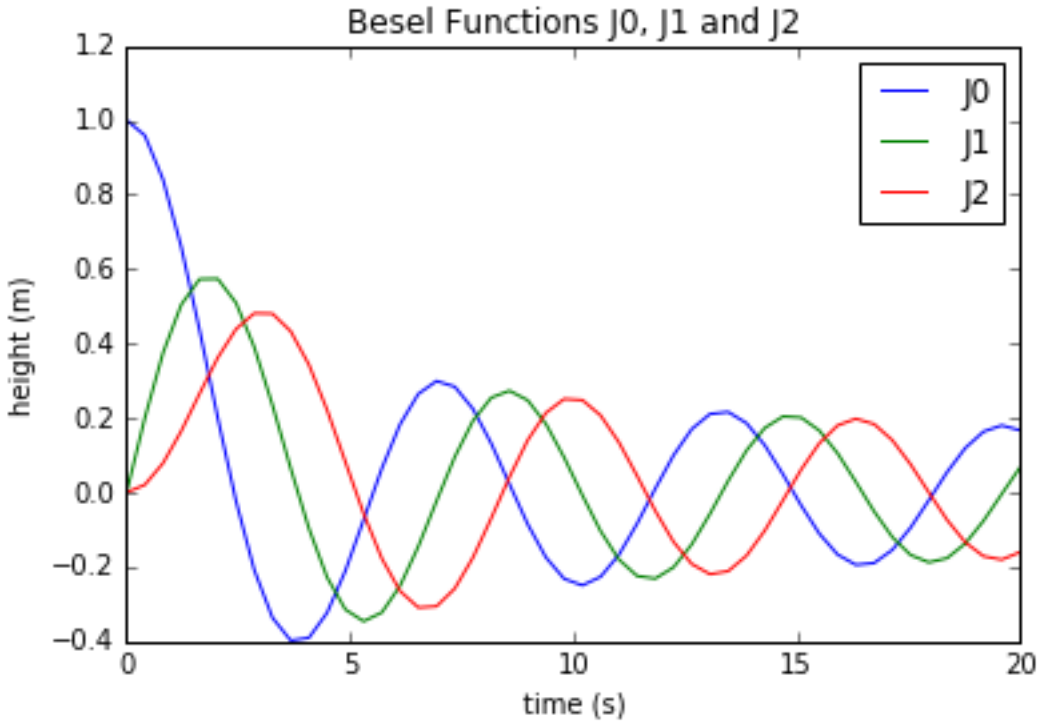



```
;add data group for this trace to file
hdf5_add, x,y2,file=file,group_name=group_name,$
        x_units=x_units,x_axis=x_axis, x_name=x_name,x_type=x_type,$
        y_units=y_units, y_axis=y_axis,y_name=y_name,y_type=y_type,$
        legend=legend, plot_graphics=plot_graphics
```

The Result

```
<HDF5 file "Fig_1.hdf5" (mode r, 12.4k)> (File) /
    root (Group) /root
        ('user_fullname', 'John Doe')
        ('user_id', 'g')
        ('date', 'Thu Feb  4 13:52:10 2016')
        ('fig_description', 'Besel Functions J0, J1 and J2')
        ('fig_source', 'Phys. Plasmas 17, 1234 2010')
        ('n_groups', 3)
      J0 (Group) /root/J0
          ('group1 plotting information', 'black line')
          ('legend', 'J0')
        x_values (Dataset) /root/J0/x_values     len = (100,)
            ('units', 's')
            ('axis label', 'time (s)')
            ('data type', 'float')
            ('nx', 100)
        y_values (Dataset) /root/J0/y_values     len = (100,)
            ('units', 'm')
            ('axis label', 'height (m)')
            ('data type', 'float')
            ('ny', 100)
      J1 (Group) /root/J1
          ('group1 plotting information', 'red line')
          ('legend', 'J1')
        x_values (Dataset) /root/J1/x_values     len = (100,)
```

Python - The figure

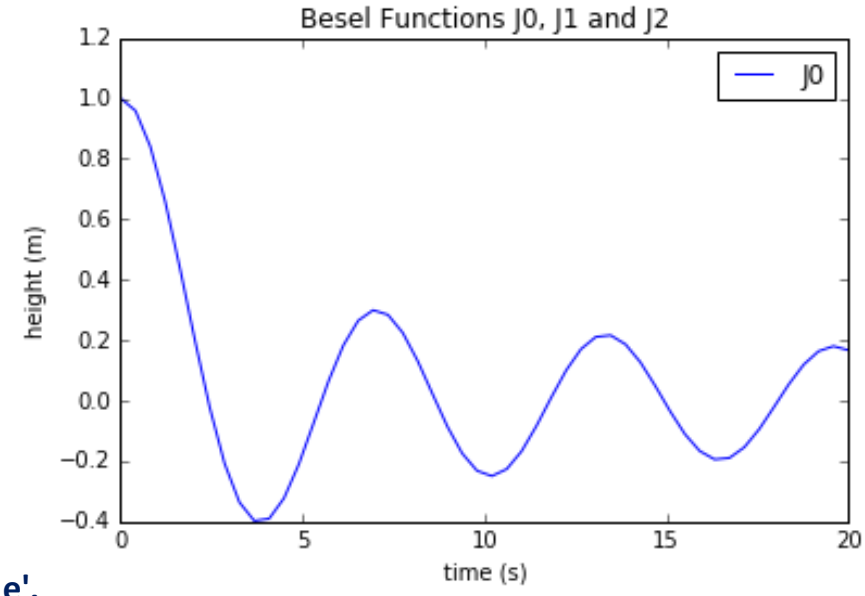


Python

```
from scipy.special import jv
rom h5_data import h5_data

file_name = 'Fig_4'
fig_description = 'Besel Functions J0, J1 and J2'
fig_source = 'Phys. Plasmas 17, 1234 2010
'comment = 'This is the way the ball bounces
'user_fullname = 'John Doe'

#Create the datafile, with file level metadata
hdf_file = h5_data("%s.hdf5"%(file_name,),
            fig_description = fig_description,
            fig_source = fig_source,
            comment = comment,
            user_fullname = user_fullname)
```

Python(2)

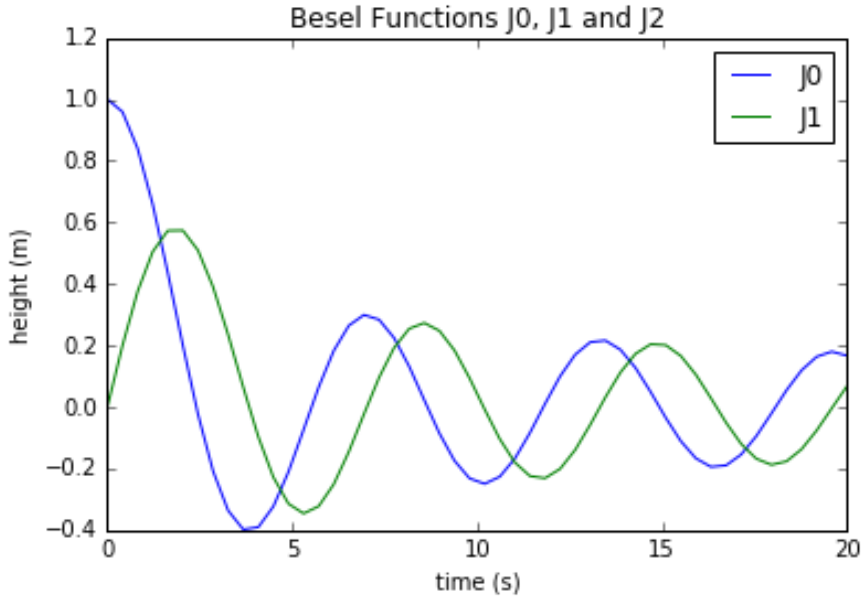
```
#Draw the first curve
x = linspace(0,  20)
y0 = jv(0,x)
plot(x,y0,  '-b', label='J0')
x_units='s'
x_label='time (s)'
y0_units='m'
y0_label='height  (m)'
```




```
#Add the first curve to the file
hdf_file.add_dataset('J0', x, y0,
                legend=None, plot_info='Blue Line',
                 x_units=x_units, x_label=x_label,x_datatype='float',
                 y_units=y0_units, y_label=y0_label, y_datatype='float' )
```

Python(3)

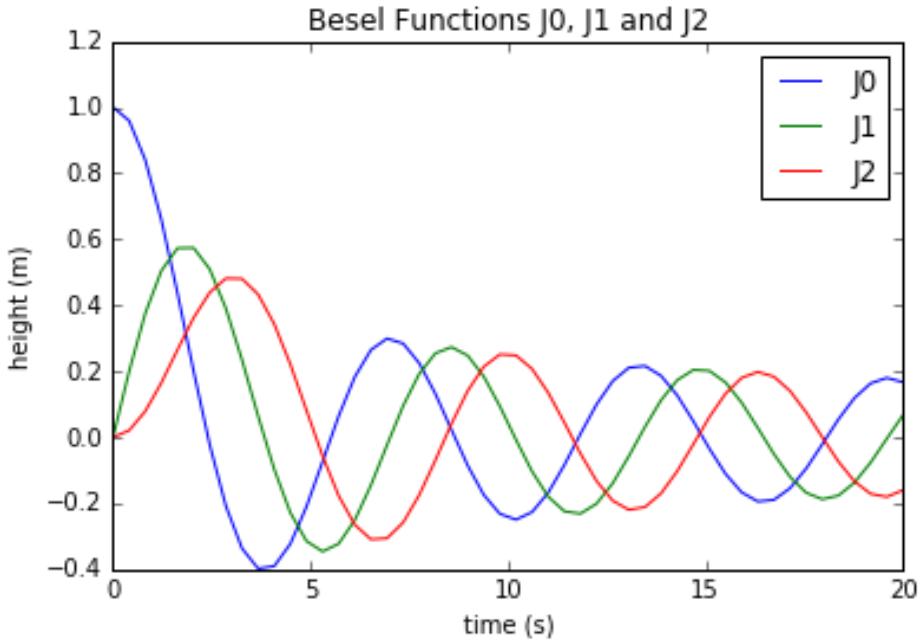
 # Draw the second curve
y1 = jv(1,x)
plot(x, y1, '-g', label='J1')
y1_units='m'
y1_label='height (m)'



#Add the second curve to the file
hdf_file.add_dataset('J1', x, y1,
legend=None, plot_info='Green Line',
x_units=x_units, x_label=x_label,x_datatype='float',
y_units=y1_units, y_label=y1_label, y_datatype='float')

Python(4)

```
# Draw the third curve
y2 = jv(2,x)
plot(x, y2, '-r', label='J2')
y2_units='m'
y2_label='height (m)'
title(fig_description)
xlabel(x_label)
ylabel(y0_label)

# Add a legend
legend(loc='upper right')
```




```
#add the third curve to the file
hdf_file.add_dataset('J2', x, y2,
            legend=None, plot_info='Red Line',
             x_units=x_units, x_label=x_label,x_datatype='float',
             y_units=y2_units, y_label=y2_label, y_datatype='float' )
```

The Result

```
<HDF5 file "Fig_1.hdf5" (mode r, 12.4k)> (File) /
    root (Group) /root
        ('user_fullname', 'John Doe')
        ('user_id', 'g')
        ('date', 'Thu Feb  4 13:52:10 2016')
        ('fig_description', 'Besel Functions J0, J1 and J2')
        ('fig_source', 'Phys. Plasmas 17, 1234 2010')
        ('n_groups', 3)
      J0 (Group) /root/J0
          ('group1 plotting information', 'black line')
          ('legend', 'J0')
        x_values (Dataset) /root/J0/x_values     len = (100,)
            ('units', 's')
            ('axis label', 'time (s)')
            ('data type', 'float')
            ('nx', 100)
        y_values (Dataset) /root/J0/y_values     len = (100,)
            ('units', 'm')
            ('axis label', 'height (m)')
            ('data type', 'float')
            ('ny', 100)
      J1 (Group) /root/J1
          ('group1 plotting information', 'red line')
          ('legend', 'J1')
        x_values (Dataset) /root/J1/x_values     len = (100,)
```

END