Ripping Data From Plots

BV

[2015-01-12 Mon 15:21]

1 Overview

Sometimes there is data that is plotted in a paper or a presentation which one would like to have in machine readable form either to incorporate it into some calculation or to plot it in the same plot as some other data. Modern, DOE funded experiments are required to have a Data Management Plan. Part of this plan is supposed to address "making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication". For results not covered by such plans, one must do what one can.

This describes a method to "rip" the data from published plots. The use of the word "rip" is in the same sense as ripping a CD to a WAV file. It is intended to be a lossless procedure. This is not digitization where one plunks down points while zooming in to some rasterized images of the plot. This only has a chance of working if the target plot is in a vector format such as PDF or PS.

This method relies on the curve or other graphical data representation to be stored in the file in a way that its drawing coordinates can be faithfully (and easily!) extracted. It relies on their drawing coordinate system to be related to that of other graphical elements which can be used to provide a scale for the X and Y directions.

2 Inkscape

This method uses Inkscape to assist in ferreting out the needed data. The example uses Fig. 5 from this paper.

First, open the PDF in your favorite reader and save (or print to PDF) just the page with the plot. Open the resulting PDF in inkscape.

```
$ inkscape paper-page6.pdf
```

Click on the target plot and "ungroup" if necessary (Right-click context menu). Keep doing this until you can select the bounding box of just the plot curve.

Open the "XML Editor" (Edit->Xml Editor) and scroll down through the top-level elements it shows. The one co responding to the plot's curve should be highlighted. Open up the element to see the `svg:path` element. It should have a "`d`" attribute with a value like:

```
m 346.422,-188.102 11.773,0 0,7.993 4.969,0.....
```

The full paste is here.

This is the path that draws the curve. Ignoring the "`m`" it is a space-delimited list of pairs of numbers. The first point is the start of the curve in drawing coordinates and each subsequent point represents a relative distance, dX/dY from the prior point. Selecting the "`d`" attribute will show the value in the window. Select the value and paste it into a text file.

Next, repeat this procedure to grab the "`m`" values for major axis tick marks, making sure to record what they represent. Eg:

```
y1=200
m 351.176,-159.043 -4.754,0
y2=800
m 351.176,-71.977 -4.754,0

x1=1
m 363.164,-264.801 0,-1.511
x2=10
m 513.859,-263.285 0,-3.027
```

This is now enough information to shift and scale from drawing coordinates to plotting coordinates and apply this transform to the plot and thus rip the data.