Integrating Professional Design Software into a Science Visualization Workflow

M. Jentoft-Nilsen

Wyle / NASA GSFC, NASA GSFC Code 613.1, Greenbelt, MD 20771, USA – marit.jentoft-nilsen@nasa.gov

Abstract – Creating science visualizations that appeal to a general audience has recently become both easier and more challenging. The wide availability of Earth mapping software such as Google Map/Google Earth has increased familiarity with mapping and the appearance of Earth from space–on the other hand such professional products have increased people's expectations in terms of visualization quality and interactivity. The standard output of many scientific processing packages now appears crude and dated. NASA's Earth Science Outreach group at Goddard Space Flight Center has integrated professional graphics and animation software into our workflow to create visualizations for display on physical globes such as Digital Earth and Science on a Sphere, virtual globes such as Google Earth and regular flat display screens. Using such software allows the visualizer more interactive control over over the design aspects of a visualization such as color, line width and anti-aliasing and labeling.

Keywords: Visualization

1. INTRODUCTION

Scientific visualization of remote sensing data for public outreach has many of the same requirements as visualization for science. Both sets of viewers are no doubt attracted to beautiful and complex, detailed images, yet there are important differences in how these two sets of viewers will interact with a visualization. For most types of scientific data, a visualization, whether a chart, graph, image or animation, is the only way to "observe" phenomena in the data; so the typical scientist/expert has acquired a visual language made up of the the conventions of science imagery in their particular field. This language, acquired through exposure to visualizations in previous literature, textbooks, etc., may include elements that are perceptually non-ideal, or elements that facilitate technical exploration of the data. However long experience with those non-ideal elements makes interpretation seem intuitive to the expert. For example, the use of the rainbow palette for pseudo-color images may seem natural to science experts and can facilitate reading metrical values off the image, but it makes seeing the overall shape of the data more difficult (Ware, 1988). The non-expert audience may very well find such visualizations beautiful and exciting; but how much they will understand, and how long they will interact with the visualization? The message received may be that the science is pretty but also impenetrable and obscure.

Like the papers they are placed in, science figures are a specialized form of communication that require training and a large amount of acquired knowledge to understand. And just as you wouldn't give a non-expert a copy of the latest paper published in scientific journal and expect them to understand its main points easily or quickly, you can't give them a standard science visual and expect them to extract meaning. A visualization for a non-expert will generally need to be more single-purpose, and be designed to reveal or focus attention on some specific feature in the data – or to tell some story about it. Nearly every feature in the science oriented visualization needs to be adapted for the non-expert. Their eyes are not trained to know where to look in a science visual, so its colors, saturation, labels, and lines widths all have to be modified to give obvious perceptual clues as to what are the important parts of the figure. None of these changes are impossible to implement using a science-oriented software tool; science tools are simply optimized for a different purpose, and they are much less efficient at creating polished visuals. Design oriented tools such as Gimp, Photoshop, Inkscape, Illustrator, Lightworks, Motion, Blender, Maya, and others are designed to be interactively used by visual artists and provide a much more flexible way to create high quality visualizations. However there are issues to be aware of in moving data from a scientific data format into professional design software.

Although NASA data is generally freely available to anyone who cares to download it, it is also archived and distributed in scientific data formats that require significant processing to create a visualization. For a scientist who possibly is already working with similar data in similar formats, who is familiar with general conventions for storing data, or who has more or less ready access to colleagues with expertise in the data format, the technical challenges of working with a novel data format are usually manageable, but that is not the case for the general public. Even if in the ideal case that public will interact with the data using a virtual globe or other tool, images for use in those tools still need to be created from data files. So, to provide true access to NASA data to non-expert audiences, it is necessary to create the specific visualizations from the data for them, and when creating these visualizations, to avoid the visual conventions of scientists, instead following perceptual design guidelines. We describe one common visualization workflow and point out issues arising in working with scientific data.

2. 2D RASTER IMAGERY

A typical scientific visualization task involves taking a single band of data from from a sensor or a single derived parameter and creating a pseudo-color image to highlight some interesting aspect of the scene. In this section we describe an example workflow using Adobe Photoshop to create an Advanced Microwave Scanning Radiometer - EOS **(**AMSR-E) sea ice concentration (Cavalieri, 2004) image.

AMSR-E data is distributed by the National Snow and Ice Data Center (NSIDC) in Hierarchical Data Format (HDF) format. This format is well supported by many scientific software tools and scripting languages, but not by graphics editing programs such as GIMP or Adobe Photoshop. So one of the main issues involved in preparing scientific data for import into non-scientific software are preserving as much of the original data range as possible in an intermediate file format that can be used by the more general purpose graphics software. Science data is often distributed as 1, 2 or 4-byte integers or as floating point numbers, while most graphics programs only support operations on 1 or 2-byte integers. Support for 4-byte integers or floats is less common and often such images are treated as High Dynamic Range (HDR) images, and scaled to 1 or 2-byte integers using a local scaling operator (i.e. pixels with the same input data count will be mapped to different output values depending on surrounding pixel values). Since we are usually trying to preserve the data's scaling, it is better to do the necessary scaling in the science software.

In the case of AMSR-E sea ice concentration, the data counts are 1-byte integers and so can be written directly to an image file format without any scaling, in this case a tiff file since we use the GeoTIFF format to retain geolocation information. A typical AMSR-E image is shown in Figure 1.

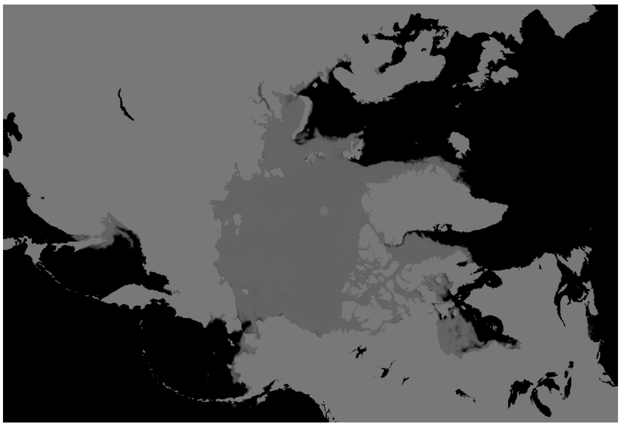

Figure 1. AMSR-E northern hemisphere sea ice concentration from December 19, 2010. Valid data, missing data and land are all represented in the 1-byte integer.

The data count information of the original data is maintained by creating a matching colorbar, which has a range of values matching the valid data range of the the sea ice data; in this case since the valid value range from 0 to 100 (representing 0 to 100 percent sea ice coverage), the colorbar is a grayscale image with a linear ramp from 0 to 100. Figure 2 shows a colorbar matching the AMSR-E data. Since the labels will be recreated in the graphics program it is not necessary place them correctly for this initial colorbar version.

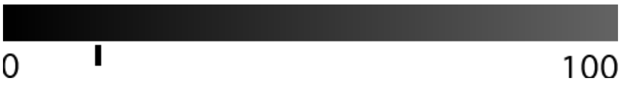

Figure 2. A grayscale color bar for AMSR-E sea ice concentration. Placeholder labels show the range of data, while ticks mark interesting levels. The tick here indicates the 15 percent level, a lower cutoff commonly used to identify sea ice extent.

Creating a colorbar is important because graphics editing programs that manage colorspaces apply a transformation to image file values when converting from one color space to another; and switching from grayscale to rgb within such a program almost alway modifies the values in an image, even before any colorizing steps have been taken. Some graphics programs also dither during a conversion in an effort to minimize banding, which further modifies the original data values. Converting a grayscale image to RGB (by placing the gray band in each of the 3 rgb channels) in scientific software can avoid this transformation in data values. One further caveat with Adobe Photoshop is that it represents 2-byte integers with 15 bits not 16 bits – it does this by rounding up all the odd integers to the next highest even integer; this emphasizes the point that all calculations that require the data counts to be actual data counts should take place before switching to a graphics editing program.

There are two categories of non-sea ice data values in AMSR-E data: invalid data points and land data points. These two values need to be treated differently. Invalid data is retained as a mask in a separate image file, but since the AMSR-E sea ice data is much lower resolution that the supporting ancillary vector and texture data that will be used in the visualization, the land values adjacent to valid sea ice values are filled with a neighborhood average. We do this calculation in the same script that extracts the data from the HDF file. Figure 3 shows the original data along with the filled data, both in the original grayscale, and after color and overlay map have been applied. The non-filled data shows distracting dark pixels around the coastlines where the original data set had land pixels.

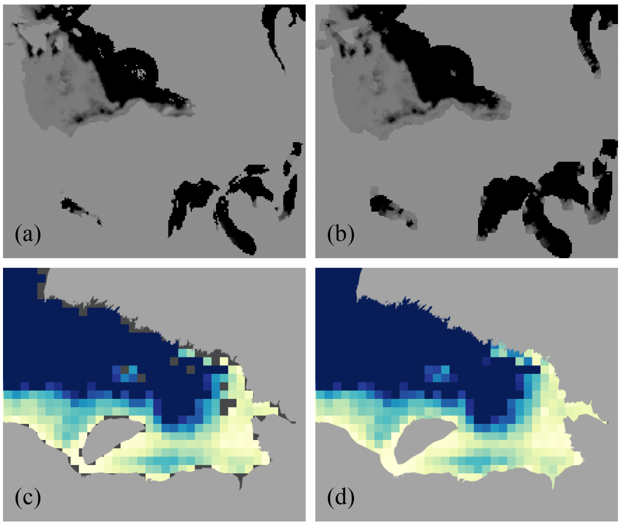

Figure 3. A region around Hudson Bay showing original compared to filled data; (a) the original data, (b) valid data pixels extended into land areas. A further zoom into James Bay of the pseudo-colored image with high resolution map overlay; (c) the original data shows gaps along the coastline where land pixels have no color, while (d) the filled version extends the data to the coastline.

Another type of metadata that should be preserved is geolocation information. In this case it is stored in GeoTIFF tags in the image file itself. This information is required in order to make matching ancillary images such as coastal outlines or background images. If the visualization is being created for display on a real or virtual globe, it is reprojected into a rectangular projection before import into the graphics editing program. If the reprojection step is done after the colorization step, it may introduce color errors because converting from one color space into another is not guaranteed to be a linear mapping and so is not guaranteed to preserve addition (e.g. a common map that transforms the anomaly data range [-1,0,1] to blue-white-red rgb color space [{0,0,1},{0,0,0},{1,0,0}] does not preserve addition, so the addition operation -1 + 1 = 0 in the data space does not map into {0,0,1}+{1,0,0} = {0,0,0} in the color space.) An additional issue to consider for reprojection, is whether the data is of a type that requires flux preservation.

Once we have created the three image files: the data file, the missing data mask and the valid data mask file, we can import them into our graphics editing program, in this case Adobe Photoshop. It is a straightforward process to combine the separate pieces into one layered image. The data image and colorbar are arranged in one layer. The valid data mask is used as an alpha channel on the data, so that invalid data pixels are transparent and valid pixels are opaque. A color lookup table can be applied with a Photoshop layer called a gradient map, which assigns an RGB value to each grayscale data value. Figure 4 shows the pseudocolor image and colorbar after this step.

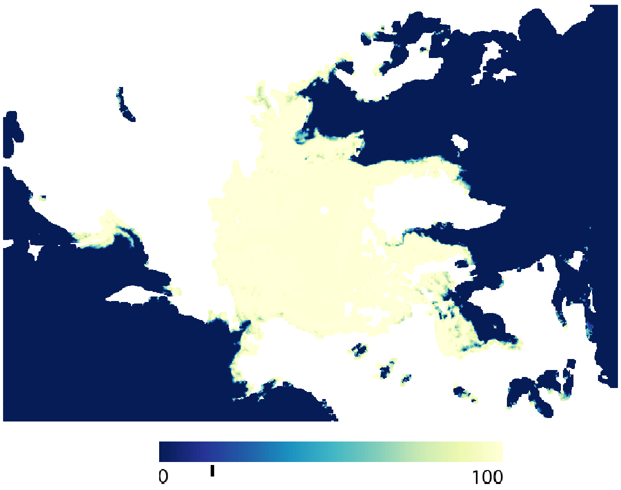

Figure 4. Pseudocolor version of the AMSR-E sea ice image with a reference colorbar.

3. VECTOR DATA

Creating a visualization often requires combining a raster data with vector data overlays – either of other science data or of ancillary data in the form of map outlines, coastlines, administrative boundaries, annotations, etc. For map-based visualizations, the original vector layers are created in a tool that has map projection capabilities, but the vector output of such programs is not often organized in a way that makes editing in a vector graphics editor easy. For example, vectors are not labeled or organized by feature type, they are not arranged to make group operations possible, the clipping algorithms used may produce polygons that do not allow for easy modification, there are limits to the number of points per vector path, etc. If the overlay and dataset being combined include different features then the overlay will have to edited to match the data. For example, the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (GSHHS) (Wessel, 1996) is a global vector coastline often combined with data. As Figure 5 shows, some GSHHS lakes have valid AMSR-E sea ice data while nearby ones are classified as land. It may be possible to fine-tune the vector plotting to leave out smaller lakes based on size, but that may also remove features such as Manitoulin Island in Lake Huron, which is in the AMSR-E land mask.

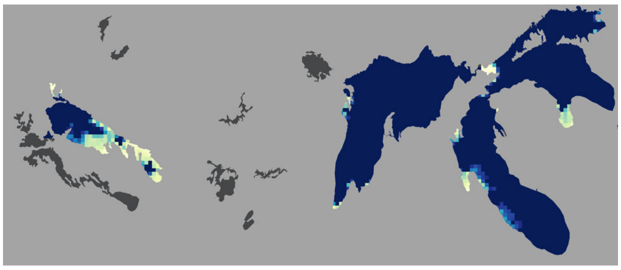

Figure 5. Closeup of AMSR-E data and GSHHS overlay map in the Great Lakes region. There are many lakes in GSHHS that are classified as land in AMSR-E.

At high resolution, some vector paths in the coastline data have more points than the maximum allowed by some vector editing programs. Even the intermediate resolution version of the GSHHS Eurasian ocean coastline is made up of 34332 points, while currently the maximum number of points per path in Adobe Illustrator is 32000. This means that when the coastline is plotted it will either need to be simplified to reduce the number of points below 32000, or need to be split into smaller polygons. Since simplifying vector paths introduces it's own set of problems (the simplified polygons may have self-intersections or may intersect with polygons in other levels,) it happens more often that the paths are split, and the output vector image file may end up having thousands of lines and polygons that are not grouped in any logical manner. There are two easy solutions to this problem. The first is to create each filled and stroked level as a separate output file, then assign each to a separate layer in a file in the vector editing program. The other possible solution is to give each level a distinct color. The layers can then be sorted out in Adobe Illustrator using the "select same color" options to select all items of a particular color (and therefore level) in order to group them in separate layers.

Once the vectors are sorted, it is easy to select particular features and adjust their color, lines width, and add special effects. Labels and other annotations are also added at this point. Figure 6 shows an example sea ice concentration image.

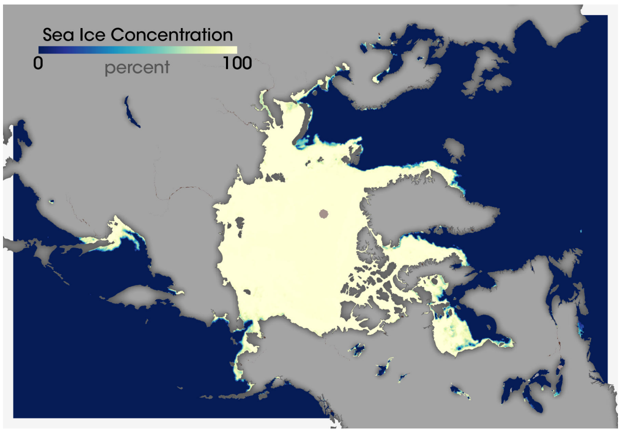

Figure 5. AMSR-E sea ice concentration image for December 19, 2010..

4. CONCLUSION

We have described the process of integrating graphic design software into a scientific visualization workflow. The integration process involves being aware of how data is represented in the design software and the ability to adjust the science processing tools to produce the required output formats. We have found that after the initial workflow setup, there is little increased resource usage in such a workflow, except for an increase in the number of intermediate files that may be needed to fully transfer the required metadata from the science tools to the design tools; and the results of such an integration are aesthetically pleasing, informative visualizations. One area still lacking in our workflow is the preservation of provenance in the final visualization – i.e. a record of the input data and processing steps used in creating a visualization. Such information would tie a visualization back to the data and processes used to create it, and allow others to more easily recreate it. In the context of visualizations created for non-experts, such information could increase interest in and use of NASA data by students and interested public. Currently some science visualization software tools such as VisTrails can record workflows, while some graphics editing software such as Photoshop can record some processing steps, but there is currently no way to integrate the two.

REFERENCES

D. Cavalieri, T. Markus, and J. Comiso. 2004. “AMSR-E/Aqua Daily L3 12.5 km Brightness Temperature, Sea Ice Conentration, & Snow Depth Polar Grids V013”, (December 19, 2010). Boulder, Colorado. http://nsidc.org/data/ae_si12.html

C. Ware, “Color Sequences for Univariate Maps: Theory, Experiments and Principles,” IEEE Computer Graphics and Applications, vol 8, p.p. 41-49, 1988.

P. Wessel, and W. H. F. Smith, “A Global Self-consistent, Hierarchical, High-resolution Shoreline Database”, J. Geophys. Res., 101, #B4, p.p. 8741-8743, 1996.