Tsering Wangyal Shawa*

Guidelines for Creating Historical Geospatial Boundary Data

Keywords: image enhancing; georeferencing; historical boundary data; vectorisation

Summary: Princeton University Library's Map and Geospatial Information Center has recently started a project to create historical boundary data of African British Colonies. The maps for creating the boundaries are taken from the British Colonial Reports - Annuals, the British Colonial Reports, and the British Colonial Office List Map Supplement. The maps were scanned, image enhanced, and georeferenced, and boundary data was extracted from the maps. I have developed a methodology and workflow to extract historical boundaries and conflate the historical vectorized boundaries to the boundary data downloaded from the GADM website. Most of the historical boundaries will be nested within the GADM boundaries, if the GADM boundaries are within horizontal accuracy of the scanned map. This paper will introduce new methodologies of creating historical boundary data.

Introduction

Libraries have collected and accumulated large cartographic materials over time. At the end of the 1990s, libraries started scanning maps and making them accessible online. Some libraries have taken the next step, georeferencing maps and making those maps accessible online to their patrons. I think the next step libraries could take is to extract data from the georeferenced maps and make those data available to their patrons. Just as making scanned maps and georeferenced maps available to researchers helps their research, extracting the geographic data from maps and making those data available to patrons will open up even more research possibilities. The project I started recently to create historical geospatial boundary data of African countries that were under British colonial rule will not only help us to develop methodology and workflow for extracting historical geospatial data in a systematic way, but also will create historical administrative boundary data that can be made accessible to the public for non-commercial use. I believe the new methodology and workflow I have developed can be used by other libraries in extracting geographic data from their map collection and then like-minded libraries can work together in creating historical geographic data by distributing the data creation work and sharing the extracted data among the partner libraries.

This project involves scanning, georeferencing, and extracting administrative boundaries data from maps included in the British Colonial Reports - Annuals published from 1918 to 1939, the British Colonial Reports published from 1920 to 1966, and the British Colonial Office List Map Supplement published in 1948.

The goal of creating the historical geospatial boundary data is to allow scholars to do time-series analysis research on the African countries that were British colonies. I will use GADM (http://gadm.org/) administrative boundaries as reference data to create historical administrative boundaries

* Head, Map and Geospatial Information Center, Princeton University Library [shawatw@princeton.edu]

because these administrative boundary data are freely accessible to the public for non-commercial use and many scholars use this data for their research.
Let me describe the methodology and workflow I used in creating historical administrative boundary data.

Map preparation and scanning

First we remove the folded maps from the Reports and flatten them as much as we can. To flatten the map, we put a large self-healing cutting mat on the map and put heavy books on top of the mat for a few days. Once the map is properly flattened, then we scan it. If a map is fragile or in bad shape, we first repair it and then put the map in mylar before it is scanned. The original maps removed from the Reports are kept in the Map Collection and the printed copies of the original maps are folded and put back in the Reports. We also add the online links of scanned maps to the Reports' catalog record in the main library catalog system.
We use our PUMap database to enter some basic information about the map before we scan it. The basic information includes title, author, publisher, date of publication, copyright status, BibID (bibliographic ID from our main catalog), ImageID (old unique image file name generator), and bounding box in decimal degrees, etc. Once the data are entered, the database will generate a unique ARK (Archival Resource Key) ID. We use this ARK ID as our scanned map file name. We scan our maps at 400 dpi with an index of 256 colors and save them as TIFF images (the Princeton Map and Geospatial Information Center's map scanning standard). Later the TIFF images are converted to JP2 images for online viewing. If the scanned map is copyrighted, we enter "*true*" in the copyright field which will make the JP2 image not accessible online.

Image enhancing

Before the map is georeferenced I enhance the scanned map image for the vectorization process. I have used GIMP software which is open source image processing software, and can be downloaded for free.
First I convert the 256 color indexed image to RGB which will allow me to use the *Brightness and Contrast* tool and also the *Adjust Color Curve* tool. There is no magic brightness or contrast number to select because it depends on how the map was scanned. I play with brightness and contrast numbers, and when I see the right balance of distinct colors, I accept the setting. After the color brightness and contrast are adjusted, I use the *Adjust Color Curves* tool to change the image tonality. Once these processes are done, then I check to see how many distinct colors are on the map. If I see 10 distinct colors, including the background color of the map, then I enter 10 in the index color conversion tool. The reason for reducing the scanned map index color is to make the vectorization process much easier. The indexed color scanned map is saved as a TIFF file.

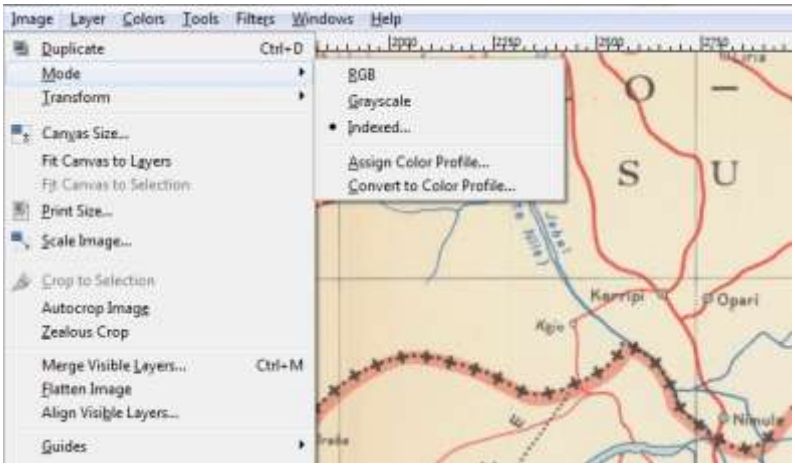

Figure 1: Converting indexed color image to RGB in GIMP software.

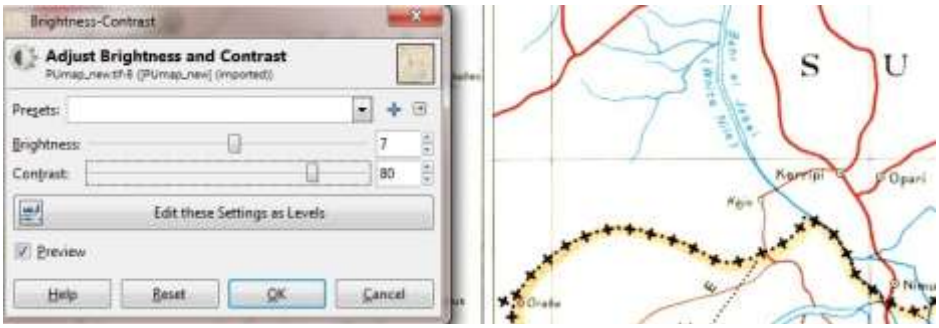

Figure 2: Adjust brightness and contrast.

Figure 3: Indexed color conversion.

Georeferencing a scanned map

Georeferencing is the process of aligning a scanned map to a known coordinate system. I use two different methods to georeference a map; it depends on the type of map. If the map has proper latitude and longitude grid lines on it, or if I know the latitude and longitude of a particular site on the map,

then I manually enter latitude and longitude values after selecting the intersection of latitude and longitude grid lines using the *Add Control Points* tool. However, if the latitude and longitude grid lines are not marked on a map, then I overlay known geographic data on the scanned map and add control points on the scanned map based on the same feature I see on both geographic data and the scanned map and extract latitude and longitude values from the geographic data.

In the first workflow of georeferencing a scanned map based on a known latitude and longitude, we will use ArcGIS software to georeference our scanned map because within this software there is an extension called ArcScan which will allow us to automatically convert the scanned map image to vector data. ArcGIS is a well-known GIS software package and it is fairly easy to use. Georeferencing is one of the tools in ArcMap software and it can be accessed through the *Toolbars*. When I have access to the Georeferencing toolbar I add control points based on intersections of latitude and longitude grid lines marked on the scanned map by entering the longitude value on X and latitude value on Y. To get better accuracy on our georeferenced map, I distribute the control points equally over the map, if possible.

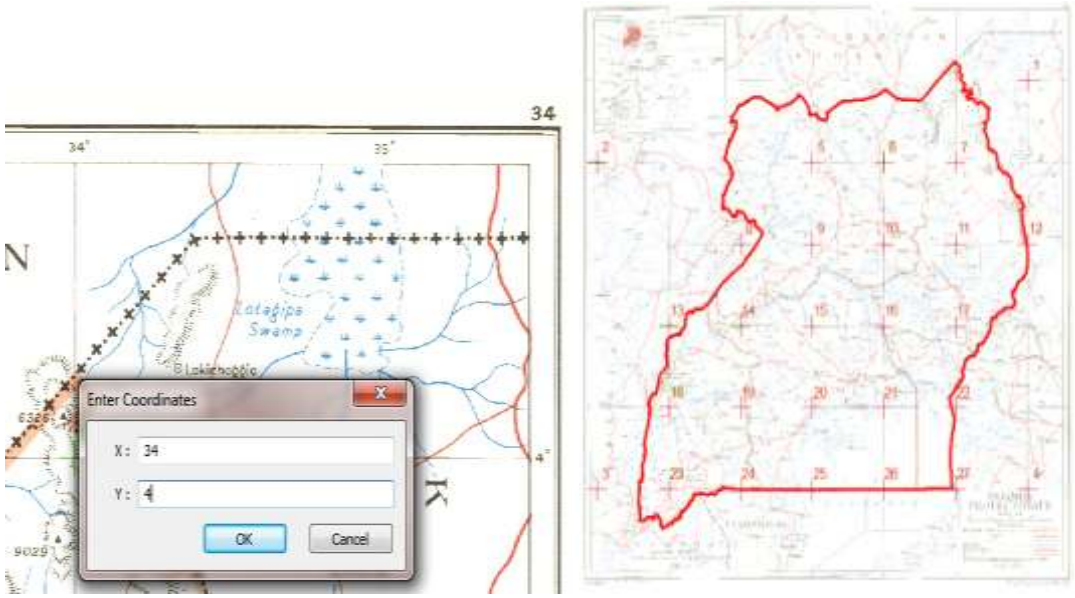

Figure 4: Coordinate values and control point marks on the georeferenced map.

I use different transformation methods to fit our scanned map on GADM boundary data: Polynomial (there are three different polynomials), spline, and adjust transformation[1]. The first order polynomial transformation, which is also called the affine transformation method, optimizes global accuracy; however it does not guarantee local accuracy. On the other hand, the Spline transformation method optimizes local accuracy but not global accuracy. The adjust transformation method optimizes for both global and local accuracy; however, it requires more control points to use this transformation method. I use either the first polynomial or the adjust transformation method. The transformation method is based on the fitting of our scanned map to the GADM boundary. Once I have entered all the control points and picked the transformation method, I save the control points data and finalize

[1] http://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/fundamentals-for-georeferencing-a-raster-dataset.htm

the georeferencing by selecting *Update Georeferencing* from the georeferencing toolbar. I use the control point data saved earlier to create a metadata record because it stores the total RMS (root-mean-square) error value, and also, if someone wants to use a different fitting or transformation method on the georeferenced map they can use the control point data I have saved.

I use the second method of georeferencing workflow if I don't have proper latitude and longitude grid lines on the map, and plan to use boundary and other geographic data that correspond to the scanned map. Before starting the georeferencing process, I try to find the map projection information of the scanned map. I know that maps published by individual countries or agencies will use different map projections. This is because when a spherical object is projected on a flat surface there will be some distortions in distance, direction, shape, and area. Different map projections preserve different distortions. It is, therefore, easier to georeference a map if you know the projection of the map. If I find the projection information on the map, I use the same map projection on the GADM administrative boundary data so that the shape of the GADM boundary looks the same as the scanned map. If no map projection information is found on the map then I guess the projection. Once the GADM administrative boundary is projected to the same map projection as the scanned map then I make sure the display of GADM boundary data on the ArcMap's *Data View* is displayed more or less in the same extent as the scanned map. The advantage of doing this is that when I click on the *Fit to Display* tool in ArcMap it will overlay the scanned map much better on the GADM boundary data, and therefore it will be easier to georeference the scanned map. When adding control points on the scanned map, I make sure the control points are distributed equally over the map, if possible, and fit the scanned map as closely to the GADM boundary as possible. Once I have added all the control points and transformed the scanned map, I save the control points data and update the georeferencing.

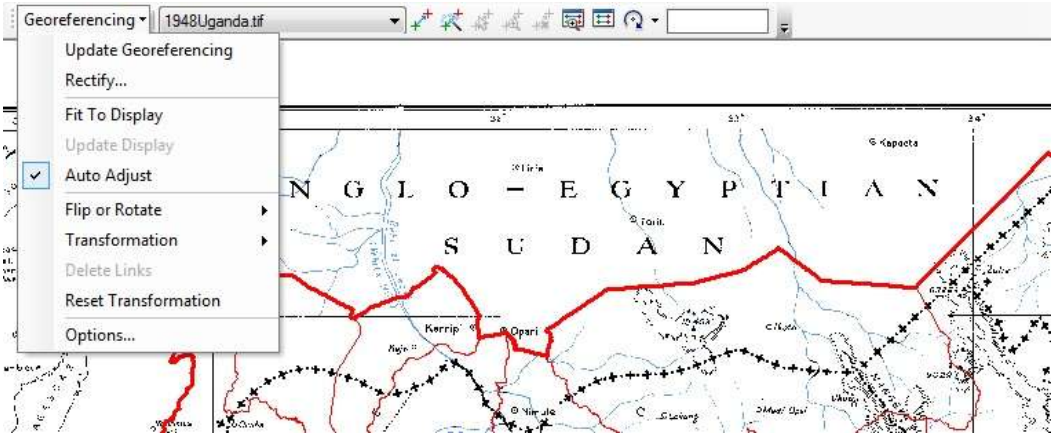

Figure 5: Screen shot to fit to display tool.

Creating vector data: automatic process

Automatic vectorization will reduce the manual processing of converting raster to vector data. I use ArcScan software in ArcGIS to vectorize the scanned map. This software allows only a binary color image that is equal or less than an 8-bit pixel depth to automatically vectorize a scanned map. The enhanced scanned map image has a 10 color index; I use the *Identify* tool in ArcMap to find out the color of the scanned map boundary. Once I identify the color, I use

the *Reclassify* tool to convert the 10 color index image to a two color image. This can be done by assigning new values in the *Reclassify* tool: the scanned map boundary color to 1 and the rest of the scanned map colors to 2.

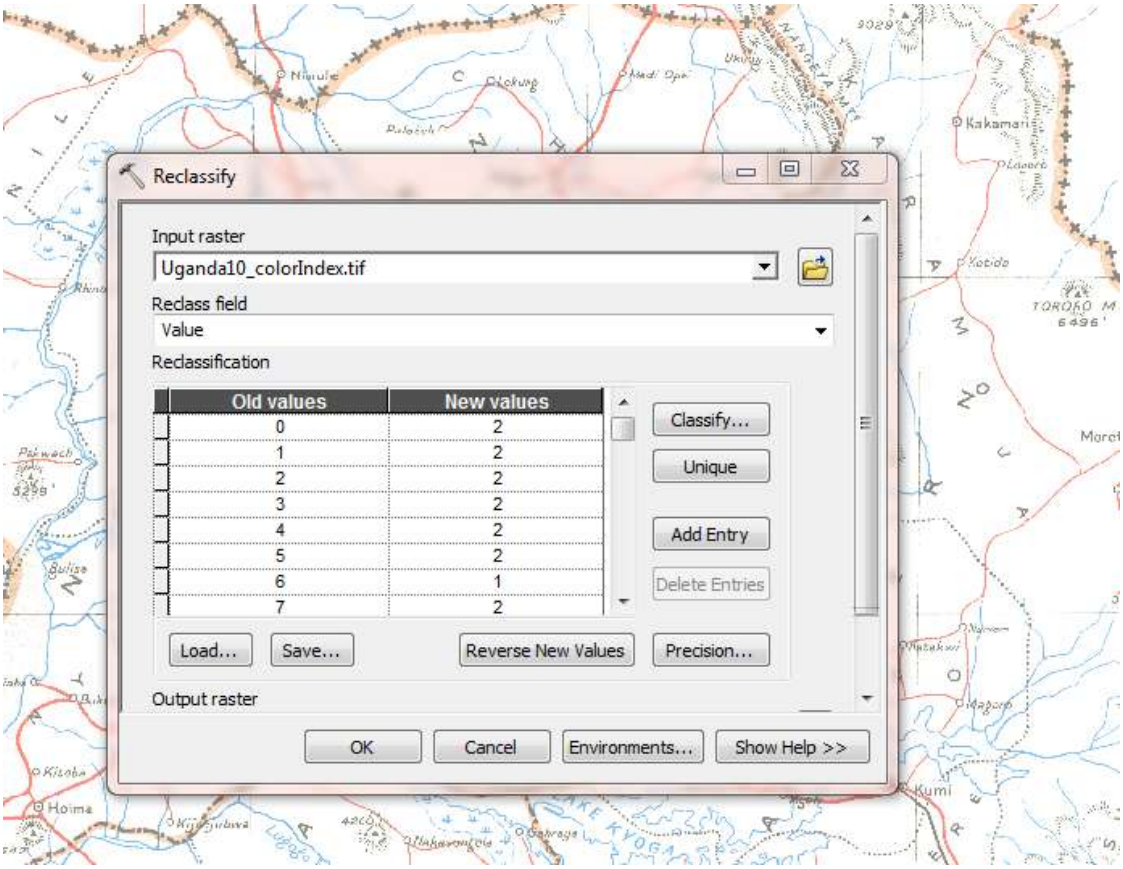

Figure 6: Reclassify tool.

When the scanned map image is converted into a two color map, the software will allow me to use the ArcScan tool to remove or clean parts of the image that are not boundary but share the same color as the boundary. The goal here is to remove all pixels that are not boundary. There are different Raster cleaning tools available in ArcScan. I can use the interactive raster cell selection tool to select raster cells and remove them from a map or I could use *Erase* and *Magic Erase* tools from the *Raster Painting* tool to select and erase cells. ArcScan has an option to draw lines if boundaries are not clear. After removing all pixels except the boundary, the scanned map image is saved.

Figure 7: Map reclassify into a two color map.

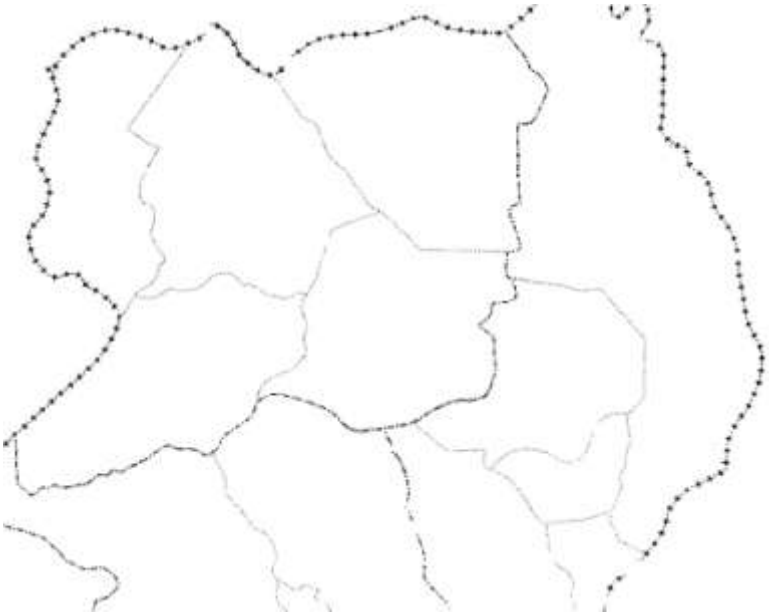

Figure 8: The cleaned scanned map.

In order to vectorize the boundary of the cleaned scanned map, I create a new shapefile to store the vectorized data. I give the same coordinate system as the georeferenced scanned map, and make this new shapefile editable.

The next step is to convert the scanned map boundary that is in raster file into vector file format. In the Vectorization Options window in ArcScan, the color I want to use for vectorization is Foreground because ArcScan will extract only the foreground color. I leave the vectorization method as Center-line. After the Vectorization Options is selected then I change the vectorization settings. Since the scanned map (Uganda map) administrative boundaries have gaps between lines I have set the Gap Closure Tolerance to 100. This setting can be changed base on the gap between the lines.

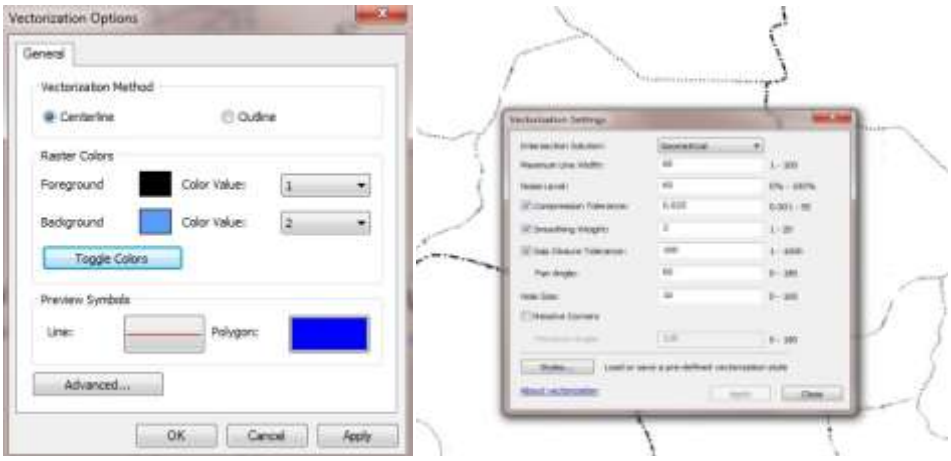

Figure 9: Screen shot of vectorization options and vectorization settings.

After selecting the right *Vectorization Options* and *Vectorization Settings*, I preview the vectorized lines by selecting *Show Preview*[2]. This process is important because it will help me to see how well the conversion of the scanned map boundaries to vectorized data will work. If certain lines in *Show Preview* are not converted properly, then I clean those areas in the scanned map image. After I am satisfied with the *Show Preview* lines, I generate boundaries from the scanned map by selecting *Generate Feature* in ArcScan. This will create a polyline based on the pixels of the scanned map.

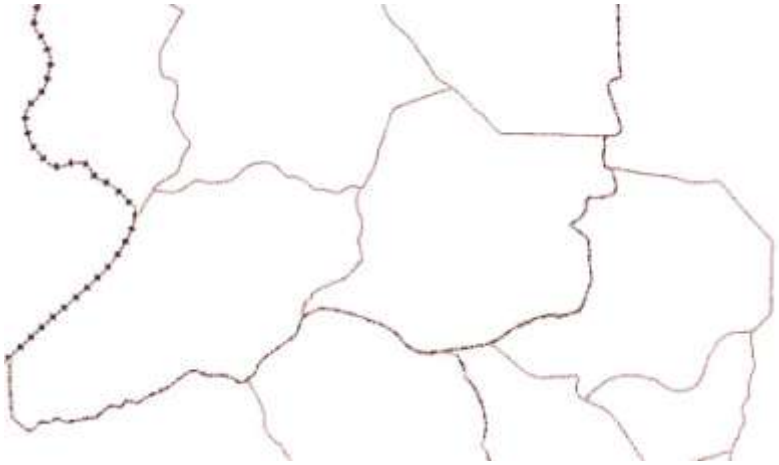

Figure 10: Vectorized feature from the scanned map.

[2] http://help.arcgis.com/en/arcgisdesktop/10.0/pdf/arcscan-tutorial.pdf

Methodology of Selecting and Creating Historical Boundary

The polyline extracted from the scanned map will not be used as the final historical boundary data because the extracted boundary will not match exactly with GADM boundaries although the boundaries may not have been changed. This is because while both the scanned map and GADM boundary data have scales, they are usually different. The boundaries extracted from the scanned maps are generally mapped at a much smaller scale and this mismatching of boundary line is usually due to differences in mapping scale and the use of different generalization techniques by cartographers. Larger scale maps are more detailed and less generalized than smaller scale maps. The scale of the map also plays a role in the accuracy of the map. Large scale maps are more accurate than smaller scale maps. The GADM boundary data do not have metadata to check the source of their data and the scale or the resolution of the data. When I compared the GADM Uganda country level boundary to DCW (*Digital Chart of the World, very common free global geographic data. The DCW data were extracted from 1:1,000,000 scale maps*) boundary data, it seems the GADM boundaries are less generalized then the DCW data. This means the GADM data was created or extracted from a better scaled (RF) map than 1:1,000,000.

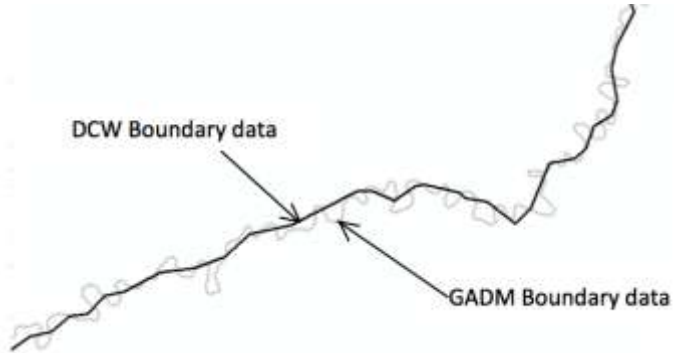

Figure 11: Compare line generization between the DCW and GADM data.

The RF (Representative Fraction) scale of the scanned Uganda map that I used as an example for this georeferencing exercise is roughly 1:2,002,176 scale. This Uganda map has no RF scale written on it, however, I calculated by measuring the scale bar with a ruler. A general rule is: if the scale bar unit is given in Kilometers then use the Centimeter unit on your ruler to measure the distance and if the unit is in Miles, then use inches; once you measure the distances, then you can use the following simple formulas to convert graphical scale units to RF scale.

The Uganda scanned map scale is in miles. I used the ruler to measure the scale and it was roughly 2.5 inches to 79 miles. Since there are 63360 inches in a mile, I can use this formula;

(79 x 63360)/2.5 = 2,002,176. This value is the R.F. Scale of the Uganda scanned map.

If the scale unit is in kilometers, and when you measured the scale of the map, you found it is 2.5 cm to 10 kilometers, then first you need to figure out the conversion of kilometer to centimeter. There are 100 centimeters in 1 meter and there are 1000 meters in 1 kilometer. In other words, there are 100,000 centimeters in 1 kilometer. We can use this simple formula;

(10 x 100000) /2.5 = 400,000 and the RF scale of this map is 1:400,000

According to the United States map accuracy standards[3], a map drawn at scale of 1:24,000 should have ± 40 feet (12.19 meter) horizontal accuracy and a map drawn at scale of 1 inch to a mile (1:63,360) should have ± 105.6 feet (32.19 meter) horizontal accuracy, or a map drawn at scale of 1 cm to 1000 meters (1:100,000) should have ±166.67 feet (50.80 meter) horizontal accuracy. This means two different scaled maps will not match exactly, if you view them at a much larger scale.
Since I could not find the colonial maps horizontal accuracy standard to measure inaccuracy, I will use the United States Army's 512th Engineering Detachment map accuracy[4] estimating guidelines as proxy. They use the following simple formula to estimate product accuracy:

> Accuracy = (Scale * Accuracy Descriptor)/1,000
> TPC (Tactical Pilotage Chart) map, scale 1:500,000 and accuracy descriptor was 2mm.
> (500,000 * 2)/1000 = 1000 meters (1 kilometer)
> ONC (Operational Navigation Chart) map, scale 1:1,000,000 and accuracy descriptor was 2mm.
> (1,000,000 * 2) /1000 = 2000 meters (2 kilometers)

According to the above formula, the accuracy of the Uganda Colonial Report map is:
Uganda Colonial Report map scale 1:2,002,176 and we will use the same accuracy descriptor (2mm)

> (2,002,176 * 2)/1000 = 4004 meters (4.004 kilometers).

I use the approximate horizontal accuracy of a map based on the scale as a guide in creating the final historical geospatial boundary data. First, I open an attribute table of newly created polylines from the scanned map and create a new field/column called Category and enter "*Colonial*" as the name of the category. I then convert GADM administrative level 1 boundary data from polygon to polyline; the level 1 administrative boundary corresponds to the scanned map district boundary. I open the attribute table of the GADM polyline and add a new field/column called Category and enter "*GADM*" as the name of the category. I merge the polyline extracted from the scanned map and GADM and then symbolize the merged data using the unique value as category. This will display two categories: Colonial and GADM. I pick red for GADM and blue for Colonial to show two distinct colors.
Second, to guide me in deciding which boundaries to keep and which to delete from the merged data, I create a buffer around the scanned map boundaries. The distance of the buffer is based on the horizontal accuracy of the scanned map. For example, the RF scale of the Uganda map is 1:2,002,176 and the horizontal accuracy of the map is within ± 4000 meters. I can use the 4000 meters distance to create a buffer and this buffer data will help me decide which boundaries to delete and which to keep. If the extracted boundaries from the scanned map and the GADM boundaries are within the buffer area then I delete the extracted boundary and if the GADM boundary is outside the buffer area then I keep the extracted boundary and delete the GADM boundary. Using this method will help us to create boundaries that will be conflated to the present GADM boundaries so that most of the boundaries will be nested within the GADM boundaries. If the national boundaries have not changed since the publi-

[3] http://pubs.usgs.gov/fs/1999/0171/report.pdf
[4] http://www.dsg.eb.mil.br/seminario_eua/arquivos/E040%20Digital%20Data%20Quality%20and%20Accuracy.pdf

cation of the map then I will accept the GADM boundary and delete the extracted national boundary. After deleting the polylines from the merged data that I don't want to keep, I save the edit.

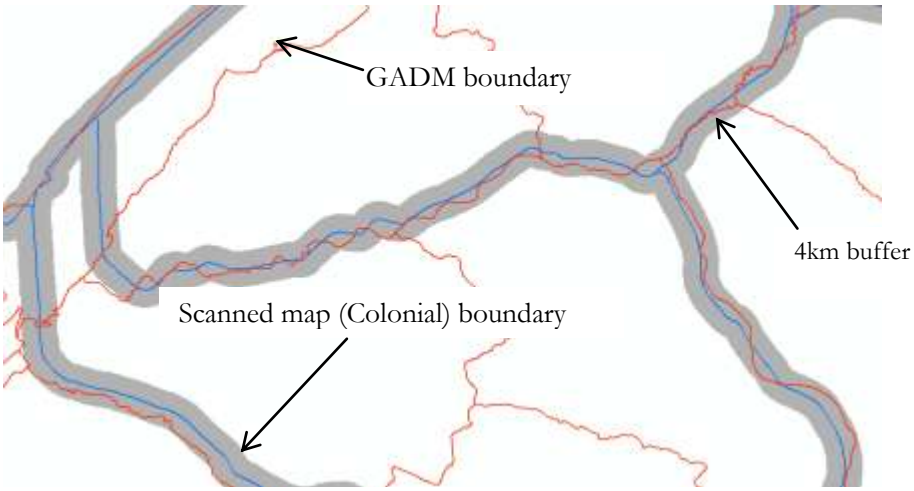

Figure 12: 4 kilometer buffer distance around the scanned map boundary

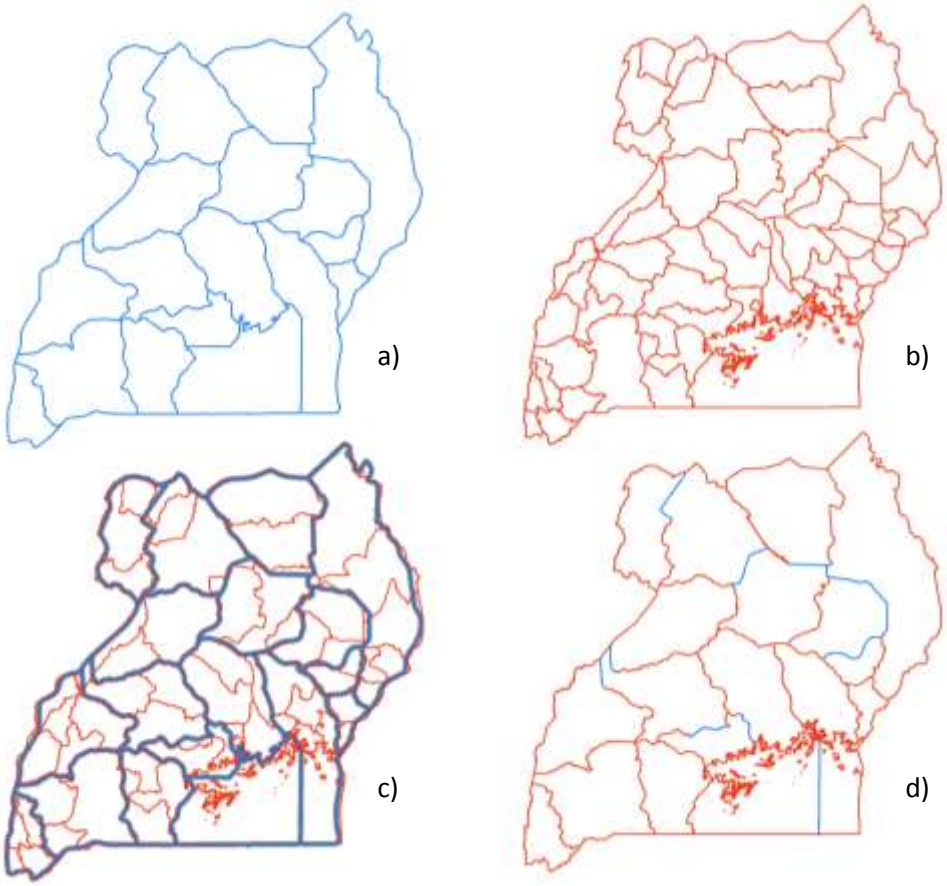

Figure 13: a) Boundary extracted; b) GADM level 1 (district) boundary; c) Boundary extracted from the scanned map and GADM level 1 boundaries merged on 4 km buffer; d) Final 1948 Uganda district boundaries conflated to GADM district boundaries.

Third, convert the merged data from polyline to polygons. Open the attribute table of the polygon and add three new fields called "*District*", "*Province*", and "*Year*" and enter each historical district and

province name, and year. This new polygon data is the final historical boundary data. We will make the finalized historical boundary, vectorized data from the scanned map, cleaned merged data, georeferenced scanned map, and control points data to our patrons. This will give our researchers more flexibility to use our data.

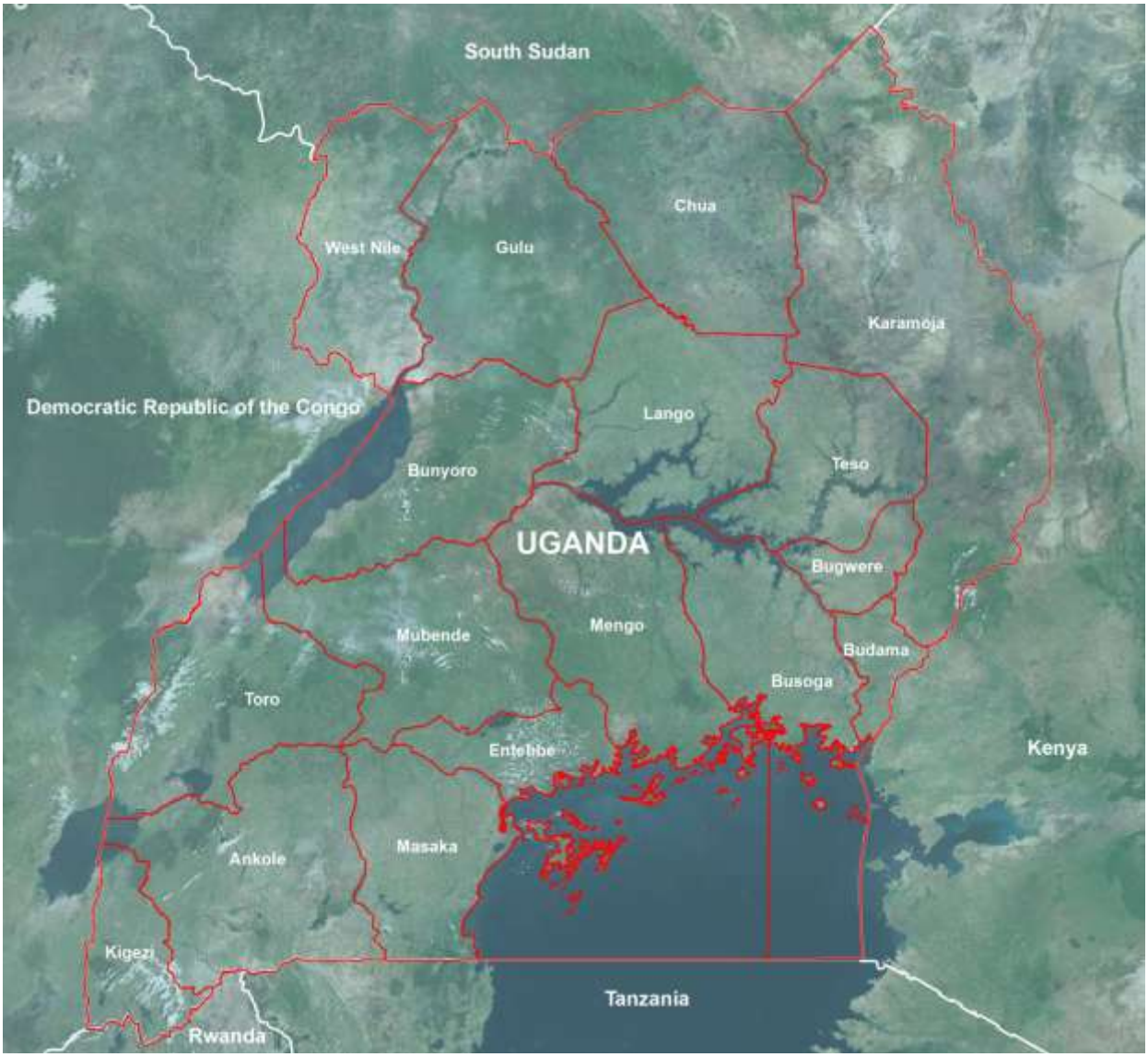

Figure 14: Finalized vectorized district boundaries of Uganda, 1948 overlaid on satellite image and other GADM country boundaries.

Table

Uganda_1948

FID	Shape *	District	Province	Year	Area
6	Polygon	Entebbe	Buganda	1948	6163646918.99
11	Polygon	Masaka	Buganda	1948	30680775845.400002
12	Polygon	Mengo	Buganda	1948	15010760286.9
13	Polygon	Mubende	Buganda	1948	13811338037
1	Polygon	Budama	Eastern	1948	2667065129.54
2	Polygon	Bugwere	Eastern	1948	3317145917.65
4	Polygon	Busoga	Eastern	1948	16884342423.700001
8	Polygon	Karamoja	Eastern	1948	30334067679.299999
14	Polygon	Teso	Eastern	1948	10358224872.9
3	Polygon	Bunyoro	Northern	1948	14496140572.200001
5	Polygon	Chua	Northern	1948	17162987077.799999
7	Polygon	Gulu	Northern	1948	16441349983.700001
10	Polygon	Lango	Northern	1948	14031974541.4
16	Polygon	West Nile	Northern	1948	10977021757
0	Polygon	Ankole	Western	1948	16107175318.5
9	Polygon	Kigezi	Western	1948	5336771501.66
15	Polygon	Toro	Western	1948	13834709597.200001

0 (0 out of 17 Selected)

Figure 15: Attribute table of finalized 1948 Uganda district boundaries.

References

Arteaga, M.G. (2013). Historical Map Polygon and Feature Extractor. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction*. Orlando, Florida. In digital form, http://dx.doi.org/10.1145/2534931.2534932

Bruce Godfrey, Hayley Eveleth (2015). An Adaptable Approach for Generating Vector Features from Scanned Historical Thematic Maps Using Image Enhancement and Remote Sensing Techniques in a Geographic Information System. *Journal of Map & Geography Libraries*. Vol. 11, Iss.1. In digital form, http://dx.doi.org/10.1080/15420353.2014.1001107

Prescott, J.R.V. (1987). *Political frontiers and boundaries*. London; Boston: Allen & Unwin.

Shawa, T. W. (2003). What is the best resolution to scan a map? Baseline: *A Newsletter of the Map and Geography Round Table*, 24(6), 6.

Shawa, T. W. (2007). Building a System to Disseminate Digital Map and Geospatial Data Online. *Library Trends*, Volume 55, Number 2, Fall 2006, PP. 254-263.

US Army. (2010). Quality and Digital Data Accuracy. *Image Intelligence Seminar: Command and Control and Geographic Information at Directorate of Geographic Service*, Brasilia, Brazil. In digital form, http://www.dsg.eb.mil.br/seminario_eua/arquivos/E040%20Digital%20Data%20Quality%20and%20Accuracy.pdf

USGS. (1999). *Map Accuracy Standards:USGS Fact Sheet 171-99*. In digital form, http://pubs.usgs.gov/fs/1999/0171/report.pdf