**Hartwell Data - Reprojection Info**

28 Sep 2010

1. **Data Archive**
2. **Projection**
3. **Filenaming Convention**
4. **Chinese Character Set Encoding**
5. **Base data**

1. **Data Archive** http://dx.doi.org/10.7910/DVN/29302

2. **Projections** The original projection of this data was undefined. If opened in a GCS WGS1984 Lat / Long Projection, the geographic context appears to be correct, so for the purposes of this task we are making the assumption that the original projection was in Lat / Long Decimal Degrees. For the transformation, use the following:

input projection (assumed): GCS World WGS 1984 (Lat Long)

output projection: PCS Gauss Kruger Xian 1980 Zone GK 19 (meters)

3. **Naming Convention** Currently the files are in nested series of folders. The order of the folders from the top drilling down are arranged by Year, Admin Level, Administrative Unit Name. There is one exception in which an intermediary folder for the Jin Dynasty is placed at the Unit Type level, then is broken down into Unit Type and Administrative Unit Names within the Jin.

There is a filename syntax for the various levels (described briefly here), which will be used to assemble the new output filenames.

Syntax Description:
- YEAR folders are named with the prefix "c" (=china), four digit year, then suffix "_av" (=arcView).

- ADMIN LEVEL folders are named for the district division levels (or Dynasties) contained in the folder, followed by the suffix "_map" (for example, chin_map, folder contains a merged dataset for the extent of the entire Chinese dynasy existing at that year, while circ_map contains "circuit" (a type of admin division) polygons broken up into major "province" level layers, and cnty_map contains

county polygons broken up into the same province layers. There are a variety of sub-types that occur in these layers, reflected in the Administrative Unit Names abbreviated in the filenames.

- ADMINISTRATIVE UNIT NAMES are indicated in the filenames of each Shapefile as an abbreviation prefix, followed by the year of the layer and a suffix letter indicating the administrative unit type being depicted by the features in the Shapefile. For example, this is how to interpret the file called:

c1080_av\pref_map\gnd1080p.shp

China, Year 1080 (arcView Shapefile) \ Prefecture Level units \ Guangdong Province year 1080 Prefectures

The desired output (new filename) for this file would be:

"v5_1080_pref_gnd_1080_p.shp"

it's necessary to keep both "pref" and the final "p" because "pref" is being used as a container, and the actual unit type suffix might vary.

The syntax rules for creating the new file name:

a) iterate through the levels of folders, keeping the parts of the folder names and filename according the following ruleset

b) YEAR folder: strip off the prefix letter "c", keep the four digit year, strip off the suffix "_av", then use the result as the prefix of the newly output filename preceded by "v5_" and followed by an underscore "_"
example: (input folder name, "c1391_av" >> output filename section, "v5_1391_")
resulting new file name down to this level: "v5_1391_"

c) ADMIN LEVEL folder: keep everything up to the underscore, strip off "map"
example: (input folder name, "cnty_map" >> output filename section, "cnty_")
resulting new file name down to this level: "v5_1391_cnty_"

d) ADMIN UNIT NAMES from layer names: strip existing underscores if found in the third character position, then insert underscore "_" between prefix, year, suffix
example: (input filename, "zlz1391c.shp" >> output filename section, "zlz_1391_c.shp")

resulting new file name down to this level: "v5_1391_cnty_zlz_1391_c.shp"

note, some filenames have underscores in them. these should be stripped out first, so that there are not double-underscores in the output filename.

example 2: (input filename, "nj_1391c.shp" >> output filename section, "nj_1391_c.shp")

the result shouldn't be: nj__1391_c.shp with a double underscore between prefix and year.

note: prefix and suffix will always be alphabetic with the year in between. the use of the underscore seems to have been done to keep the prefix as three characters long. so for example, zlz is okay, but fj had to be padded out to fj_ to be three characters long. (hope this helps the parsing!) the suffix is always one character long. the year is always a four digit number, so for the Tang period the number will be 0741, not 741.

**4. Chinese Character Set Encoding**

The encoding is BIG5 Traditional Chinese.

**5. Base Data**

To the best of our knowledge, the original base data that Hartwell used to develop his historical administrative units was the draft 1990 PRC Counties dataset provided to him by Lawrence Crissman (http://www.acasian.com ).