Imagery and Raster - Distributing and Updating

Imagery is a critical and fundamental base layer for geospatial applications and geographic information systems. The Geographic Information Network of Alaska has been a leader in the distributor of imagery, specifically satellite data and base map data for Alaska and to a lesser extent the broader Arctic area since 2001. GINA has worked closely with federal, state, local government, commercial, and public users trying to consume imagery datasets into their GIS systems in their environment. Through this effort GINA has developed a strong understanding of the wide variety of data formats and distribution techniques needed to support Alaskan and Arctic users. This document will focus on the high level recommendations for raster data distribution and updating.

UAF-GINA Recommendations for Updating Raster formats

Raster imagery consists of pixels (data), a projection, and information about the geographic extent of the data. This information defines a raster covering an area of interest. What follows are GINA's recommendations for the main components of a raster, and strategies for optimizing storage and update procedures. First steps are for a user-defined projection and extent, as dictated by the area of interest.

GINA's Recommendations for Raster Storage

The ideal storage format has these properties:

- Space efficient
- Compressed
    - Allows the use of lossy or lossless compression, reducing the size of the data when possible.
    - For satellite imagery, a lossy compression scheme like JPEG with a 70% quality setting is suggested.
- Allows efficient random reads (and writes)
    - The storage format needs to allow applications to request subsets of the imagery quickly and efficiently.
- Provides overviews
    - The storage format needs to provide a way to store and access "zoomed out" versions of the data quickly and efficiently. Generally this is done by storing reduced resolution versions of the data,

Determining Projection

The ideal projection depends entirely on the area of interest, and will vary based on the location, and the extent (see below). Most users expect to view the world as "North Up" so ideally the projection would also accommodate that preference.

For arctic waters around Alaska, GINA suggests EPSG:3572.

pre-generated, so high quality downsampled data can be quickly accessed. For example, additional copies of the data can be stored at zoom levels of 50.0%, 25.0%, 12.5%, 6.25%, 3.125%, 1.5625% to present smaller “thumbnail views” of the original data.

Defining Extent

The “extent” is a minimum bounding rectangle defined by coordinate pairs. The extent identifies the area of interest at a resolution that is useful to the user.

- Provides a valid data mask
    - For datasets where there is not valid data for the entire extent, there needs to be a way to determine if a given pixel location has valid data or not. This is often done by defining a “no data value” or a no data mask.
      Lossy compressed data requires a “no data” mask, as compression artifacts can push valid data into the “no data” defined area, and vice versa.
- Allows data to be “reasonably” sized
    - The format should allow datasets larger than 4 gigabytes to be stored in a single file.
- Allows alternate color spaces
    - JPEG compression schemes work best in the YCBCR color space. YCBCR is used for efficient color storage and transmission in digital photography and video.

GINA’s Recommended Raster Format

GINA recommends using tiled, JPEG compressed Geotiff, with the bigtiff extensions and a no data mask. These can used singly or in an indexed set, depending on how the data is structured. Data with large areas of contiguous no-data are best stored as sets, while data that is rectangular is best stored as one large geotiff.

For some specialized applications, for example when serving data to a high volume web mapping site, explicitly tiled formats may be required to minimize the cost of each individual request.

GINA’s Strategy for Raster Updates:

How updates are approached for raster data depends a lot on how frequently the updates occur, and the size of the expected updates. For example, the USGS topographic maps for the Alaska are updated very infrequently, while near real time satellite imagery is updated every hour.

For Infrequently Updated Datasets:
First, determine the “expected” size of the update. If it is more that 1/10 the total size of the dataset, then just replace it outright, without any updating process.

Otherwise, slice the updated region into “tiles”, and extract the data for the updated region as in a tiled manner, including the overviews. Update the dataset in place by applying the update “tiles” to the source data. Checksums (md5sums) should be maintained for the update tiles, the dataset to be updated, and the final product to verify the process is complete and successful.

For Frequently Updated Datasets:

For datasets updated frequently, it is best to just focus on making the transmission format as space efficient as possible, and convert the transmission format into the form actually used by the application. The exact details of this process would depend a lot of the form and characteristics of the data. For 3 banded imagery, an efficient strategy would be to:

1. Cut out a defined region, at a defined scale.
2. Convert that data to a highly compressed form in a space efficient container.
    a. A YCBCR color space standard JPEG at a fairly low quality setting for example.
    b. Include separate mask, if that is needed, either as a WKB geometry, or a compressed 1bit mask.
3. Transmit the data, with a minimal set of compressed metadata
    a. Which defined region, which defined scale, date, etc.
    b. Checksum
4. Transform the transmitted data into the format used by the application.
5. Add to the application / make the app aware of the data.