**Forest Service Lidar Indexing Project**

**Overall goal**

**Methods**

The indexing process was developed using the Region 6 lidar data stored on the T: drive. All run times are for these data. The indexing process consists of four basic steps:

* Recursively scan a parent folder to create a list of subfolders,
* Look for one or more target file types in each folder,
* For folders with target files, scan file headers and create an index,
* Postprocess the initial index to generate summary information and correct index problems (mainly CRS-related). Postprocessing varies depending on the folder structure and availability of CRS information.

The recursive scan can be accomplished using the DIR command (DIR /b /s>folderlist.txt). Our CIO contact ran a “special” script for us that took almost an hour. On VDI, the DIR command took less than three minutes. Running the DIR command locally with the T: drive connected via VPN took 6 hours.

The search for target files in each folder was done using a compiled C++ program that looked for the first file in a folder that matched the target types. I went with C++ because of the availability of the FindFirst() function that does not create a list of all files in a folder prior to checking for a file matching our target types. Running against local files was very fast. Running this locally with T: drive connected via VPN took 12 hours. Output from the search is a CSV file with the folder names and a flag indicating whether or not each target file type was found in the folder.

The scan of file headers was first attempted using python and/or R code running locally with the T: drive connected via VPN. This turned out to be very slow even when using a binary read of the file headers. From past experience running the same scan of headers on local drives, the binary reads were much faster compared to using PDAL (in python) or the lidR package (in R). Testing with the binary read function coded in C++ linked to R scripts using local files indicated a performance gain of 200-300 times compared to using a binary read in R or the lidR package. My hope was to run the same R/C++ code on VDI to take advantage of the faster connection to the T: drive. However, after installing Rtools from the software center on VDI (Rtools is needed to compile the C++ code that links to R to do the binary read), I found that permissions/policy would not allow the C++ compiler to run. In the end, I ran the R code from a local machine connected to the T: drive via VPN (ran for about 35 hours). While this seems excessive and should be much faster when run on VDI, the index files that were produced allowed me to move forward.

**Post-processing of the indexes varies depending on the data source.**

For the Region 6 data, I had to guess the CRS for ~60% of the index files. This is not particularly surprising given that our target data (non-3DEP) tend to be older so CRS information in point files, if included at all, is encoded as geoTIFF tags which do not provide a readily interpreted CRS “name”. I added a field (assignedCRS) to the dataframe associated with the boundingbox and boundary layers. The CRS WKT was added in this field when a CRS was “guessed”. I manually examined maps of all indexes with assigned CRS and only found seven projects/folders with problems. Five projects/folders had incorrect CRS (but don’t know how to figure out the correct CRS) and two had some point tiles with data in a different CRS than other tiles. This was manifested by a boundary polygon that spanned several states or extended into northern British Columbia.

**Results**

Region 6 data

Data for Region 6 is stored on the T: drive in this folder:

T:\FS\Reference\RSImagery\ProcessedData\r06\R06\_DRM\_Deliverables\PointCloud

The initial folder scan was done on 3/10/2025. The second scan looking for target file types was completed 3/31/2025 and found 272 folders containing LAS or LAZ files (8 folders had both file types). I did notice some differences in the total number of folders found (~43,000) between scans done by CIO and the DIR command so I assume that R6 is actively managing these data and some changes were made. After the scan of file headers, 279 indexes were created (folders with both asset types get two indexes).

Table 1. Summary information for Region 6 point cloud data.

|  |  |
| --- | --- |
| **Index files created:** | 279 |
| **Total data size (Tb):** | 59.4 |
| **Index files identified as 3DEP:** | 52 |
| **3DEP data size (Tb):** | 29.9 |
| **Non-3DEP data size (Tb):** | 29.6 |

**Index format**

A separate index is created for each folder and each file type (LAS, LAZ, COPC, or EPT). Index files are created by reading headers for all point data files in a folder. Index files are written in geopackage format with three layers:

* Boundingbox: overall bounding box for the tiles in the folder
* Boundary: polygon(s) that are the union of the tile bounding boxes. This will generally delineate areas covered by data but some portions of the tile bounding boxes may not actually contain point data.
* Assets: individual tile bounding boxes

Coordinate reference system (CRS) information is extracted from the point files and assigned to the index. If point files do not contain a CRS, a field (hasCRS) in the boundary and boundingbox layers is set to indicate that no CRS was found and the CRS WKT field (crs) is blank. If the CRS could be “guessed” based on the folder of file names, the corresponding CRS WKT is put in the assignedCRS field. For the Region 6 data, many folder names contain a string that identifies the CRS. However, the name is insufficient to fully guess the CRS. The name generally identifies the projection but not the horizontal datum.

The boundary and boundingbox layers share the same attributes (Table 1). The CRS of the index should be the same CRS described in the crs field or the assignedCRS field (when crs field is empty). For assets (Table 2), crs should match crs for the boundary and boundingbox layers. However, it is possible for some assets to have CRS information that is different from the overall CRS for the index. Only the first asset was examined when assigning the CRS for the index. If the first asset has no CRS information or CRS information that is different from other assets, there could be mixed CRSs for assets, or an incorrect CRS assigned to the index.

Entwine data is structured differently from simple point files. With entwine there is a json file in the parent directly that contains the layout of individual tiles corresponding to octree cells in the data structure. For the scan as implemented in Python, the json file is read to get general information and the header for the root volume tile (ept-data/0-0-0-0.laz) is read to get header information. Headers for individual assets (nodes) are not read. The R implementation with C++ header reading does not currently support entwine data.

Table 2. Field names, data types, and descriptions for boundary and boundingbox layers in point index geopackage files.

|  |  |  |
| --- | --- | --- |
| **Field name** | **Type** | **Description** |
| base | Character | Folder name. Does not include training slash. |
| pattern | Character | Search pattern (regex) for target files (“\.las”, “\.laz”, “\.copc”, “ept.json”) |
| assettype | Character | Asset type. Either “points” or “ept” |
| assetcount | Integer | Number of point tiles in folder for LAS/LAZ/COPC. For ept, should be 1. |
| assetsize | Long integer | Total size in bytes for all assets in folder. |
| totalpointcound | Long integer | Total number of points for all assets in folder. |
| hasCRS | Logical | Logical value indicating that CRS was found in assets. TRUE means CRS information was found and should be in the crs field. FALSE means that CRS was not found in assets. |
| minx | Double | Minimum X value for bounding box containing all assets. |
| miny | Double | Minimum Y value for bounding box containing all assets. |
| maxx | Double | Maximum X value for bounding box containing all assets. |
| maxy | Double | Maximum Y value for bounding box containing all assets. |
| crs | Character | WKT string representing CRS information found in assets. May be a simple projection or compound CRS. |
| assignedCRS | Character | CRS WKT assigned to index. Blank if assets contained CRS or no CRS was assigned |

Table 3. Field names, data types, and descriptions for assets layer in point index geopackage files.

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| filespec | Character | Full path to the asset file. |
| filesize | Long integer | Size, in bytes, of the asset. |
| pointcount | Long integer | Number of points in asset. |
| compressed | Logical | Logical value indicating file is compressed. |
| copc | Logical | Logical value indicating file is COPC format. |
| creation\_day | Integer | Creation day of year[[1]](#footnote-1). |
| creation\_year | Integer | Creation year1. |
| point\_record\_format | Integer | Point record format1. |
| point\_record\_length | Integer | Point record length1. |
| major\_version | Integer | LAS format major version1. |
| minor\_version | Integer | LAS format minor version1. |
| minx | Double | Minimum X value for bounding box containing all assets. |
| miny | Double | Minimum Y value for bounding box containing all assets. |
| minz | Double | Minimum Z value for bounding box containing all assets. |
| maxx | Double | Maximum X value for bounding box containing all assets. |
| maxy | Double | Maximum Y value for bounding box containing all assets. |
| maxz | Double | Maximum Z value for bounding box containing all assets. |
| crs | character | WKT string representing CRS information found in assets. May be a simple projection or compound CRS. |

1. Refer to LAS standard documentation for details. [↑](#footnote-ref-1)