Legacy Data Interpretation and Migration for TERRA
Northern Region Land Type Association Legend and Spatial Data
Henry Shovic
Gallatin National Forest
February 8, 2004

Objectives

1. Prepare R1 Land Type Association Legacy data for migration to TERRA, the corporate database for legend data.
2. Prepare the TERRA legend and the spatial data for synchronization
3. Limit work to $6,000 (20 person days.)

Criteria

- Maximize usability of legacy attributes
- Minimize errors in transfer
- Maximize consistency of data
- Minimize data loss during preparation and transfer

Issues

- Legacy data were primarily in text format, compiled with some standardization, but by numerous individuals, and only generally edited for consistency.

- Legacy data were not spatially linked, i.e. the legend in text did not match spatial data. In fact, the map unit names were not even present in the spatial data.

- Legacy data were at levels higher than "landtype," the level of data for which TERRA was primarily designed. The entire project pre-dated TERRA development (pre-1997.)

- Data sources were too large for efficient manual interpretation. There were initially over 663 total "map unit descriptions" in the text, with 152 duplicates names. There were over 900 components in these descriptions. Each text description was over 200 words, for a total of 258 pages of text. There are 34,000 polygons in the spatial data.

Methods

The original legend data were provided in 25 ADOBE PDF files. These were converted to RTF (rich text format) files, and from there to WORD documents. They were then concatenated and all introductory and miscellaneous data removed. Inter-application VISUAL BASIC programming was used to concatenated, clean, and parse data to WORD documents and from WORD documents to ACCESS tables. ACCESS queries and programming were used to "clean" the data. Manual interpretation and ACCESS queried were used to provide quality control. Six VISUAL BASIC programs were written to perform various stages of parsing. Twenty-eight SQL queries were written to provide quality control, "clean" data for use with lookup tables in TERRA, and to populate final tables. About 10 hours of manual review was performed, for a total of about 2,000 manual corrections in field values.

The data was prepared for migration using TERRA-provided tools (The TERRA MIGRATION TOOLS, and the TERRA MIGRATION TEMPLATE.) TERRA personnel oversaw the final migration.

Results

- Objective 1: Prepare R1 Land Type Association Legacy data for migration to TERRA, the corporate database for legend data.

There were over 200 text blocks having identical legend names. There were small differences between these blocks and having no better method for differentiating between them, the first occurrence was used, and latter blocks were ignored.

After elimination of duplicates and parsing through four VISUAL BASIC programs, a total of 439 map units were derived from the text data. The final legend included 48 more map units (see below.) A total of 762 individual components were parsed from these descriptions. Thirty-nine fields were instantiated with 15,555 total values. About 10 hours of manual review was performed, for a total of about 2,000 manual corrections in field values (< 1% of total field values.)

Some fields (primarily in the vegetation subject area in the EUI_COMPONENT table) were filled, but not used in the final tables. This was because:

- there was too little correspondence between descriptors and standardized values accepted by TERRA
- there was little data available
- or naming conventions were variable.

In many cases, potential natural vegetation (PNV) was not described so much as was a list of existing major vegetation species. Though these data are not in the TERRA vegetation and soil fields, they were parsed and captured in a temporary table for future processing.

All data from the original text document were captured in memo fields for future processing. For the map unit table (EUI_MAPUNIT) these data are referenced in the COMMENTS field. The actual text data are in tables on the Gallatin National Forest, because the TERRA COMMENT field has a 2000 character limitation, too short for the entire text. However, for the component table (EUI_COMPONENT) the COMPONENT_CONCEPT field was large enough for source component text, including soils and vegetation data. These data, though not in the appropriate fields, can be accessed through "intelligent" queries using text-processing functions in ACCESS queries or application-based VISUAL BASIC modules, and can be analyzed to provide better estimates of PNV.

Three tables were completed: EUI_PROJECT, EUI_MAPUNIT, and EUI_COMPONENT. All map units have at least one component, with a maximum of four. The 48 map units without legend data (described below) have no component data, but have entries in the map unit table. There are a total of 487 entries in EUI_MAPUNIT, with 439 having mapped attribute data, with a total of 762 component descriptions in the EUI_COMPONENT table.

- Objective 2: Prepare the TERRA legend and the spatial data for synchronization

During review I found the published spatial data (titled: r1ltas) had no attribute to use for synchronization with legend data. This was because, the legend's Map Unit Descriptions were based on a landform and geology type and either a Section or Sub-Section Designation. Spatial attributes only included the landform and geology type as a single attribute, with BOTH Section and Sub-Section included as two other attributes. Therefore, a query was written to attach each Map Unit Symbol to the appropriate landform, geology type, and either section or sub-section, depending which one was present in the legend after processing. This was based on the assumption that spatial data of some sort were used when describing the various landform/geology type/section/sub-section combinations in the text document. After final processing I found two items in the legend not present in the spatial data, and 48 polygon names with either no section or subsection present; or a section or sub-section not in the legend text. The former were left in the legend, and the latter were not attributed further, but the names were added to the legend to retain synchronizability. The revised spatial layer as provided is in geodatabase format. It can be converted to a coverage or shapefile for use in ARCVIEW. This revised layer is now synchronizable, i.e., every polygon has a corresponding legend entry in EUI_MAPUNIT. The corresponding field used in my geodatabase is LTAName. The geodatabase feature class is called r1ltas_geodata in the geodatabase called R1LTAgeodatabase.mdb.

One other characteristic of these spatial data and the accompanying legend deserves comment. The original data were developed on a Forest basis, then a region-wide legend was developed for landform and geology type. Sections and Sub-Section attributes were then added by Forest personnel. There is no indication that these individuals were looking at the entire range of regional range of polygons having that landform, geology type, and Section or Sub-Section when they described the component. Therefore different map units could have identical attributes, but because they occur in separate Sections or Sub-Sections, they may have different names. This may possibly reflect in the legend, but was not tested here.

- Objective 3: Limit work to $6,000 (20 person days.)

This project to date cost $6,000. About $1,000 of unfunded additional work was done to finalize the results.   Additional work to provide vegetation and soils data by component and to revise the spatial layer would cost an additional $6,000.

Suggestions for future work

Some suggested improvements in the attribute data are:

- Parse and Interpret PNV data from memo data for component PNV fields
- Parse and interpret soils data from memo data for component soil fields.
- Synchronize spatial data with TERRA attribute data

Some suggested improvements in the spatial data are:

- Verify accuracy of legend data on a spatial basis using example areas having Landtype level mapping.
- Attach attributes to the 48 polygon types having no legend data
- Revise slope and elevation ranges using spatial data
- Verify Section and Sub-Section for each polygon
- Fill in the 17 un-mapped polygons
- Verify soils data using STATSGO (Regional soils spatial data)
- Use SILC (Regional satellite-based vegetation spatial data) to attach existing vegetation data to map units.

METADATA for geodatabase: r1ltas_geodata and interchange file: r1_lta_terra.e00

The field used in my geodatabase is LTAName. The geodatabase feature class is called r1ltas_geodata in the geodatabase called R1LTAgeodatabase.mdb.

Projection for both coverage and geodatabase: UTM zone 12, meters, NAD27
ARCINFO vs 7