Appendix B

Data Profiling Template

All data should be identified within the source systems to understand the shape, consistency, and variability among the individual records. The goal of the integration team is to gather the systems, tables, fields, and other source inputs and then analyze each of these records either from the source or through input staging files. Although every project migration is different, data usually has a specific pattern that is worth noting in a data profiling document. Don McMunn (<http://www.ipcdesigns.com/data_profiling/>) has provided a data profiling tool that can aid in this effort. See Table 1 for several of the fields that I have found the most useful along with an Excel spreadsheet (located at xxxx) with several samples to get you started.

Table 1: Description of Profiling Fields

| Field | Description |
| --- | --- |
| TNAME | Name of table containing column being profiled |
| CNAME | Name of column in table being profiled |
| DATA\_TYPE | Name of the data type of the column |
| DATA\_LENGTH | Maximum length of the column in bytes |
| NUM\_TABLE\_ROWS | Total number of rows contained in the table at the time of the data profiling pass |
| NUM\_DISTINCT\_VALUES | Total number of unique values found in the column at the time of the data profiling pass |
| MIN\_DATA\_LENGTH | Shortest length of data value found in this column; mainly useful for string data types |
| MAX\_DATA\_LENGTH | Longest length of data value found in this column; mainly useful for string data types |
| NUM\_NULLS | Total number of NULL values found in the column at the time of the data profiling pass |
| DENSITY | A ratio of non-null values to total rows in the table |
| ALPHANUM\_COUNT | Number of rows containing ONLY an alphanumeric value that is not either a date or a numeric only value |
| DATE\_COUNT | Number of rows containing ONLY a valid date value in this column |
| NUMERIC\_COUNT | Number of rows containing ONLY numeric values in this column regardless of data type |
| MIN\_ALPHANUM\_VALUE | minimum alphanumeric value found in a string column; based on default page |
| MAX\_ALPHANUM\_VALUE | maximum alphanumeric value found in a string column; based on default page |
| MIN\_NUMERIC\_VALUE | smallest numeric value found in a numeric column |
| MAX\_NUMERIC\_VALUE | largest numeric value found in a numeric column |
| MIN\_DATE\_VALUE | oldest date value found in this column |
| MAX\_DATE\_VALUE | most recent data found in this date column; |
| NUM\_EMAIL\_INVALID | (if field is an EMAIL field), how many of these records are invalid? |
| NUM\_CONTAINS\_NONASCII | How many records contain characters outside the range of normal ASCII values (i.e., foreign and non-printable values such as ä or line feeds ^013) |
| PICKLIST\_VALUE | (if field is a PICKLIST field), what are the values? |