Data Profiling Best Practices

Overview

This white paper provides an overview of best practices with data profiling activities, in particular focusing on two areas:

- **Why use data profiling technologies?** – Examines the benefits of data profiling solutions and the circumstances where best deployed.

- **Data profiling process** – Examines the best scenarios for deploying and using Group 1 Software's data practices for data profiling.

Why Use Data Profiling Technologies?

As recent as a few years ago, the area of data analysis – understanding the quality and structure of data assets within an application – was a relatively ill-defined area within an organization's IT strategy.

Traditional approaches to data analysis are usually dependent upon a combination of inputs – documentation, individual knowledge, and ad hoc data base query tools – which are used to select aspects of a data source. Such approaches are often time consuming and incomplete, as analysis tends to be concentrated in known areas of the data.

Data profiling tool sets, like Profiler Plus, allow organizations to accurately and efficiently analyze and diagnose the quality of their data. By completing a process of analyzing complete data sources as one process, organizations capture a complete understanding of their data assets.

Deployment of Data Profiling Technologies

Profiler Plus can be deployed into any project, or activity requiring data analysis. These activities fall into one of two main core categories:

- Data quality management

- Data integration

Data Quality Management

Data quality management initiatives focus on the process of ensuring that an organization's data assets are of sufficient quality to meet its needs. There are three main areas of interest:

>> **Completeness** – Does the organization have data assets that are incomplete or missing? For example, do all customers have associated addresses?

>> **Accuracy** – Is the organization's data assets sufficiently accurate to meet internal (business processes, decision making, etc) and/or external (regulatory, third parties) requirements?

>> **Integrity** – Are the organization's data assets consistent across the enterprise? – for example, does the list of suppliers in a companies ERP system match those in the finance application? – Do the relationships between different data assets make sense?

The most practical approach in any data quality management initiative is to be goal oriented. Analyzing all data assets is not an efficient use of time. Instead, efforts should focus on the information assets the organization believes to be of the greatest importance to their business. In addition, the metrics by which these data assets are assessed (quality criteria, quality tolerance levels, etc.) must also be defined. Typically, one would expect these to be determined by the key users/owners of the data, typically business or operational units.

By establishing which assets are assessed and defined as quality criteria, Profiler Plus assists the analysis process in several ways, including:

>> Allows users to determine what attributes to examine up front and include in the analysis process.

>> Enhances the ability to rapidly and automatically analyze data. Traditional manual approaches are time consuming, error prone, and expensive. Profiler Plus enables users to rapidly analyze data and compare it to the business requirements.

>> Allows the centralized management of data issues and documentation.

>> Gives visibility as to where certain issues are occurring.

Data Integration

Data integration initiatives focus the consolidation of one or many sources of data into a new technology application. These activities are usually the result of a new business initiative or system implementation, such as Customer Relationship Management or management information initiative.

By their very nature, data integration projects require a solution that allows an organization to physically move data from one application to another. However, differences in application design and the usage of information often mean that the data must be manipulated (through transformations, or translations) as part of the process.

Mapping specifications are developed to transform data and load it into the target application. The accuracy of these specifications is critical to the integration process. It is especially important as the cost of resolving inaccurate data-mappings grows exponentially as the project nears completion.

One key cause for this issue is an initial poor understanding of the source data to be moved. Data integration projects have a number of commonly occurring risks, including:

>> Definition of the target application is not finalized.

>> Understanding of the source data is based on out of date original documentation.

>> Insufficient analysis of the source data, as traditional data analysis techniques are costly and time consuming.

>> Data integration staff are not users of the data.

>> Scope of source data to be integrated increases as understanding of the source data improves and/or target system requirements change.

Profiler Plus addresses many of these issues by providing a thorough understanding of the targeted source data.

Data Profiling Process

The methodology described takes a centrist approach to the problem and needs adjustment based on the goals and desires of the specific project at hand and the features and functions of the tools used. For example, if the tool supports loading samples within its functionality, the extract programs does not need to concern themselves with this aspect. On the other hand, if the data profiling tool does not support sampling, this functionality (if required) must be included in the extract programs requirements.

The approach consists of five steps or phases:

1. **Prepare for the project**
2. **Prepare for the analysis**
3. **Extract and Format the data**
4. **Sampling**
   a. Load a Sample of the data
   b. Analysis of the sample
   c. Adjust the extracts and formats of the data
   d. Produce deliverables
   e. Delete the samples
5. **Analysis**
   a. Load the Data
   b. Perform the Analysis
   c. Produce deliverables

Prepare for the Project

This data profiling methodology is not intended to take the place of the project plan methodology used to create the Project Initiation Document. However, the major components listed maintain continuity and reference for the "profiling" aspects of the project. Remember, the Project Initiation Document allows you to manage expectations.

The Project Initiation Document serves four major purposes:

>> The data profiling project works within the boundaries established by the Project Initiation Document.

>> The data profiling project produces the deliverables as outlined in the Project Initiation Document.

>> The data profiling project communicates via the policies and procedures as identified in the Project Initiation Document. This identifies the communications to those closely involved in the project as well as those of general interest in the project.

>> Assigns work load. Typically one or more entities are assigned to a single analyst. In the case of very wide entities or short time frames, attributes of one entity are assigned to multiple people.

The Project Plan or Project Initiation Document contains the following sections:

>> Background
>> Scope
>> Deliverables
>> Project team
>> Communications management
>> Change management plan
>> Risk management
>> Cost management
>> Task list
>> Gantt chart
>> Network chart

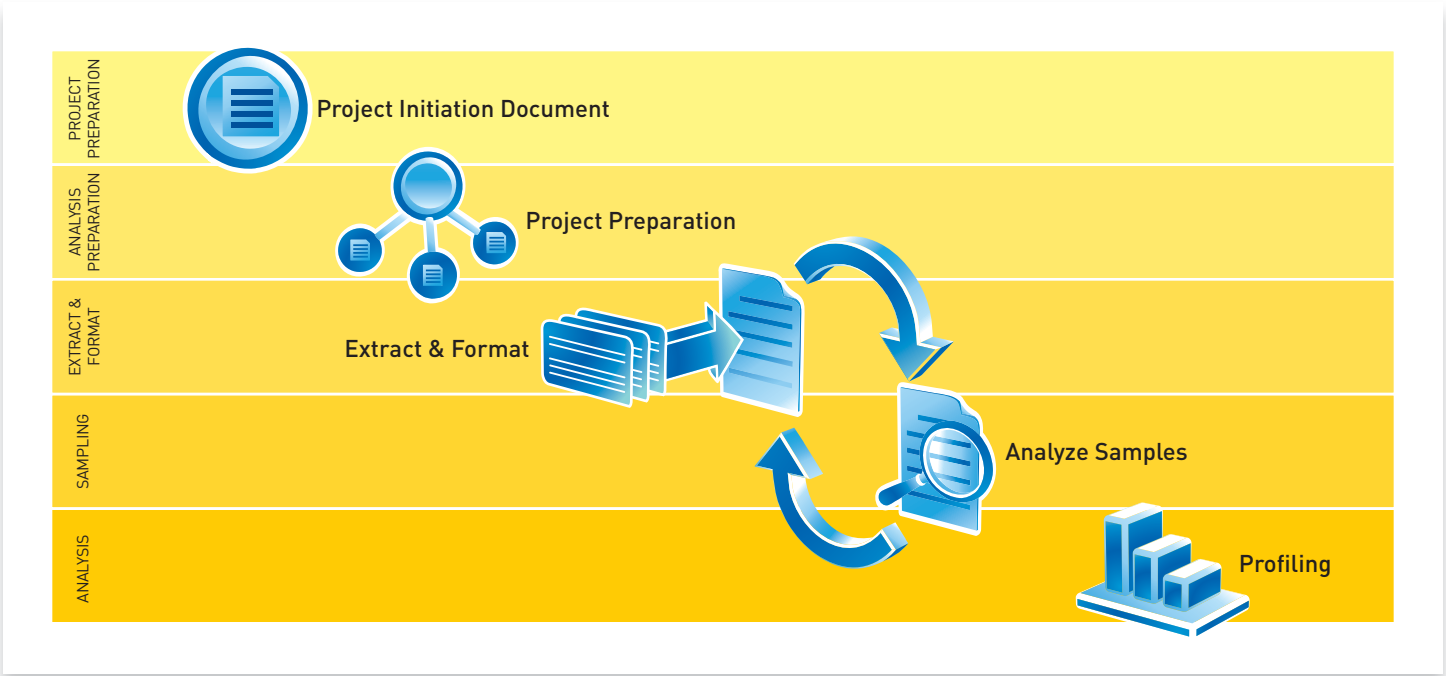
Analysis Preparation

It is important to gather, train, and otherwise familiarize the analysis team with the materials available to assist in the analysis, and setup of the data profiling tool for use.

Exactly where you start depends on the initiative being undertaken. For most conversion projects, the target is known and it is most effective to drive towards that goal. For a data quality initiative, there is no target and the focus needs to be focused on the correct data, in the correct column, and the consistency of that data.

Review Project Initiation Document

The Project Initiation Document contains the deliverables, time frames, and boundaries of the project. The team must be familiar with this critical document to avoid excessive time spent on irrelevant matters.



Profiling Overview

Current Documentation

It is critical to gather and gain familiarity with all pertinent documentation, including:

- >> Gathering current system documentation
- >> Targeting system documentation
- >> Compiling logons and passwords in existing metadata repositories

Team Training

The team must not only be trained in the use of the tool being used, but in the deliverables expected and the process and policies in place. Familiarity with other tools used ensures successful project completion.

Pay particular attention to functionality that requires additional effort in the extract and format process. For example, if the profiling tool does not allow you to dynamically parse and reanalyze data, this step must be done in the extract and reformat process.

Internal Setup/Decisions

A few decisions must be made on how the data is grouped internally. Profiler Plus supports virtual grouping of entities using the system construct. Two options are available for grouping data – by functional area or by physical system.

The team decides the scope of the activity, and the appropriate systems, entities, and attribute, to examine. This decision is driven by external requirements, including:

- >> Particular data within a company must meet compliance regulations.
- >> Information requirements of a target system within a migration project.

Profiler Plus allows the user to set a scope flag on all attributes, which is available from all attribute lists. The team selects all attributes within scope at the beginning of the exercise. During the

analysis process, users view this information and ignore out of scope attributes. This helps prevent "paralysis by analysis".

Users can view the attribute scope at any time by using the attribute viewer windows, at either project, system, or entity level. The filtering or grouping functionality available in these windows assist with seeing what is in and out of scope.

Activity Workflow

There are three main roles in analysis projects, and one individual may cover multiple roles.

>> Project manager: interested in issues, scope, progress.

>> Data analysts: analyze the data and look for candidate anomalies.

>> Business users/analysts: have particularly detailed knowledge about the data content identify whether data is correct.

It is important to ensure that all in scope data is analyzed, and that no analysis work "falls through the cracks". There are two perspectives here:

>> Each user must know their assigned entity.

>> The team, as a whole, must know which entities are assigned, and to whom.

Each user can view all the items assigned to them from the "Users->Username" node within the tree. Filtering by entity shows all entities.

To view the entities assigned to particular people, access the project entities or system entities windows. Filtering on "Assigned To = Blanks" reveals entities that have not been assigned.

During a data profiling exercise, there are two main types of activity that must be monitored to ensure that the project stays on track:

>> How much of the in scope data has been analyzed?

>> How are identified issues progressing?

Ultimately, the goal is to finish analyzing all in scope data and resolve all open issues.

To support the analysis issue, when analysts have finished analyzing data to their satisfaction, they can mark the attribute as analyzed. This is done either from the profile view or the attribute list view available at entity, system or project level.

It is then simple for team members to view the analyzed attributes using a filter or group on any of the attribute list views. This, combined with a filter on the scope flag, allows progress tracking and attribute analysis.

To support the progression issue, Profiler Plus allows the creation and assignation of notes, with reporting functionality to assist team management. Analysts create notes, and then assign them to specific individuals for resolution, along with priority and status information. Individuals log in to Profiler Plus to view their assigned notes, or alternatively, export their notes to HTML or Excel.

To see all the notes assigned to a given user, the user can open the "Items Assigned To" list from the users node, (for a given user), and filter on "Type=Notes". This information can also be accessed from the project notes window.

When an issue is resolved, the note is marked as closed.

Managers track the progress of note and analysis status by using the notes windows. By using filters, it is possible to see the issues that remain open, their priority, and their assignation. For example, a manager can filter on "Status=Open", and group by "Assigned to". This provides a grouped view by name of all open notes, making it easy to follow up and resolve outstanding notes.

To facilitate this process, when creating users, it is helpful to use the user description field so that analysts can easily choose the most likely candidate user to request research.

Extract and Format the Data

Extract and format consists of creating the extracts and any required format definitions required by Profiler Plus.

This activity is for data files:

>> That cannot be accessed directly via ODBC connections.

>> Where suitable production file formats do not exist.

>> Where data is not stored at the granular level required to support the analysis.

Create the Extract Program(s)

Creation of the extract program(s) is treated as any other program development. These programs must include requirements, code walk-thrus, test plans, and all the other items that create a well functioning system. An error in this area results in inaccurate data.

Load Preparation

Load preparation consists of preparing the data, file definitions, and systems to support the analysis. The steps involved in this preparation include:

>> Creating csv.

>> Creating flat file and flat file definition.

>> Creating appropriate ODBC connection, as required.

Csv	Each field, if separated by a comma, and text fields enclosed within quotes. Generally this type of file allows the first row to contain the name of the column.
csv File Definition	Some product require or allow you to create definition rules for csv files. It is helpful to add or change column names or add descriptions to the attributes.
Flat File Definition	Varies based on the data profiling product chosen. It varies from a flattened copybook or equivalent for the language used, to pre-defined formats specific to the tool itself.
ODBC Connection	Open DataBase Connectivity, a standard database access method developed by Microsoft Corporation. The goal of ODBC is to access any data from any application, regardless of which database management system (DBMS) is handling the data.

Sampling

Preparation for the sampling phase is based on the condition of the documentation, the knowledge of the analyst or the experts available, and the volume of data.

This sample load allows you to:

>> Determine any fields that are incorrectly defined in the associated file definition.

>> Determine fields that are misunderstood and need refinement.

>> Identify fields to be separated into component parts for analysis.

>> Identify fields that are out of scope.

>> Identify tables of values that need to be created and loaded into the system.

>> Allow determination of the correct level of granularity for the fields being analyzed.

Running the sample step allows significant savings on large files that may take hours to load and then reloaded once a basic understanding is achieved.

Load a Sample of the Data

Many types of sampling exist. It is best to choose one that provides the best results in the shortest period of time. Where possible, work within what is available in the data profiling tool. The first 100 or so samples generally provide enough information and save processing the entire input file.

Analysis of the Sample

Data analysts analyze the data by attribute. All the attribute profile information is available in the entity profile viewer. Assuming that the scope flag is set, only those attributes that are “In Scope” should be analyzed.

When the data analyst is finished analyzing the attribute, the “Analyzed?” flag is set to ‘Y’ to indicate that analysis is complete. If issues exist within the data, a note can be created by right clicking on

the attribute and selecting "Raise Note". This note will then be linked to that attribute to aid subsequent reporting. If the analyst feels that other attributes should be involved with the note, they can be associated in the "Attached to" tab.

Alternatively, the notes tab on the entity itself is used to report anomalies that the analyst considers to be entity specific.

Large free form text fields are generally not benefited by data profiling. However, it is not uncommon for these large fields to contain some cryptic codes in the beginning of the field. Consider splitting the first twenty positions into a separate field for analysis.

In some cases, it is desirable to split or combine fields prior to loading. Fields such as concatenated keys, telephone numbers, and addresses, can be adjusted prior to the load or data can be exported from the tool and re-imported.

Adjust the Extracts and Formats of the Data

Based on the results of the analysis, adjustments may be required to the data extract program(s) due to:

>> Misunderstood data – incorrect documentation or rusty expert knowledge.

>> Finer granularity desired – data is not at the atomic level.

>> Bypassing out of scope fields.

The intent of the sample analysis is to identify the gross discrepancies that are aided by refinements in reformatting.

Profiler Plus supports multiple profiles for an entity, so analysis can be concentrated around data sets with particular sampling characteristics.

Produce Deliverables

Various deliverables are developed from both in scope and out of scope initiatives.

Delete the Samples

Once data has been analyzed, the data profiles can be deleted. This does not remove issues or metadata, but it frees up space on the analysis server. However, the ability to drilldown to records and patterns does not exist.

Analysis

The actual analysis varies based on the knowledge of the business, the system, and the results desired. However, the approach below, while not all encompassing, sets the core requirements.

The analysis results are found by examining the single pane analysis windows for each entity profile.

Analysis Assistant

Profiler Plus auto generates analysis results for the user. These results are used as a sanity check, or to prompt further investigation. The results are found in the assistant tab, within the entity profile window. The list is as follows:

>> **Outliers**: indicates that there may be anomalous data for a given attribute. Further investigation of the frequency value pairs should be performed to see if anomalous values exist.

>> **Primary key**: informs the user whether the attribute is a candidate primary key, or is potentially corrupted.

>> **Code**: indicates that the field may be a code field. If it contains outliers, these may well be values that are not valid. Checking them with a business user is a sensible precaution.

>> **Indicator**: indicates that the field is used as an indicator field.

>> **Constant**: indicates that the field contains only one value.

>> **Pattern outlier**: indicates that there may be outliers within the patterns.

Blanks/Nulls/Low Values/High Values

Blanks, nulls, low-values, and high-values are strong indicators of the current condition. If the field is a key field, this generally represents an erroneous condition. While null usually represents a field not required by the business, the remnants (screens, reports, etc.) should be removed.

If this field is required by a new system, a way of populating it, pre or post conversion, must be identified.

Low Value	High Value
000-00-0000	999-99-9999
NULL	

If the field is required, it would be known that these records are in error and would require an approach to find the correct values.

Minimums/Maximums

Minimum and maximum values show that the data needs additional research.

System	Minimum	Maximum
System 1	000-00-00001	

Zeros in the first three, second two, or third four positions indicate an error.

Patterns

Patterns are one of the best error indicators.

System	Values	Pattern
System 1	123-45-6789	9(3)-9(2)-(4)
System 1	12-3456789	9(2)-(7)
System 2	123456789	9(9)

The above example shows several things. Assuming that there were no other variations in the formats:

- >> It is determined that both the SSN's and the TIN's are stored in a formatted fashion in System 1. If storage of this attribute is not formatted, a business rule is needed to remove the special characters.

- >> If both systems carried individual and corporation tax identifiers, a code or some other means of determining individuals from corporations is needed in System 2.

Duplicates / Inconsistencies

Some fields should not contain duplicates. Finding the number of occurrences quickly identifies the items that need to be researched and corrected.

System	Values
System 1	123-45-6789
System 1	123-45-6789

Finding the number of occurrences to be more than one within the same system (in this example) might indicate an error conditionthat requires further research.

Invalid Codes

For flags and codes such as state codes, country codes, and color codes, create a table of these codes and join to the content field. This allows missing, misspelled, or unexpected values to be identified.

Identify Keys

Where keys exist or are required, a unique field is critical.

System	Values
System 1	123-45-6789
System 1	123-45-6789

If this field were intended to be the key, the above occurrence would require further research/correction using the key testing functionality described below.

Key Testing

You may wish to determine whether duplicate values exist for collections of attributes. For example, the testing for composite keys. If duplicate pairs of values exist, drill down to the underlying records for further examination.

Join Testing

This allows testing to determine whether attributes from different tables join together. This is typically used when data is duplicated across systems, and orphan records begin to occur. Essentially, a join between two attributes may be tested and Profiler Plus will provide detailed information and outline any integrity issues. You can then drill down to the actual records to see the problems in detail, and raise a note if necessary.

Outputs

Data quality specifications:

- Complete analysis of data in scope
- Summary of analysis activities (using Profiler Plus' notes functionality)

Data integration specifications:

- Input into mapping specification


APPROVED
123.45678.9
PITNEY BOWES • PITNEY BOWES • PITNEY BOWES •

For more information about our products and services, please log onto our website at ***www.g1.com*** or call us today at ***888-413-6763***.

Pitney Bowes
GROUP 1 SOFTWARE

4200 Parliament Place, Suite 600
Lanham, MD 20706-1844
1-888-413-6763 • www.g1.com