Best Practices in Data Profiling and Cross-System Data Discovery

Philip Russom
Senior Manager, TDWI Research
August 27, 2008
Agenda

• Introduction to the Issues
  – A Definition of Data Profiling
  – The Four Practice Areas of Data Profiling
• Ten Best Practices for Data Profiling
• Summary & Recommendations
• Bibliography
Introduction to the Issues

• TDWI’s POSITION –
  – *Detailed profiling of the data in source and target systems is a prerequisite to successful projects for data integration, data quality, data warehousing, master data management, and so on.*

• PROBLEM –
  – *Many technical users scrimp on data profiling by doing it rarely or shallowly, by profiling only known systems or small pieces of them, and by settling for profiles that not very insightful or actionable.*
  – *These myths and poor practices result in project overruns, the exclusion of important data, incomplete and inaccurate profiles, and severe productivity losses.*

• SOLUTION –
  – *Augment data profiling best practices to include cross-system data discovery, regular data monitoring & remediation, and the collaborative improvement and sharing of profiles.*
  – *Automate these practices with a tool or tools that can automate discovery, data mappings, and data dependency definition.*
A Definition of Data Profiling

• Here’s a definition adapted from Wikipedia, so we all start on the same page:
  – *Data profiling is the process of examining the data available in an existing data source (e.g., a database or file) and collecting statistics and information about that data. The purpose of these statistics may be to find out whether existing data can easily be used for other purposes, provide metrics relevant to data quality and standards, assess the risk involved in integrating data for new applications, and assess whether metadata accurately describes the actual values in the source database.*

• Well, there’s more to it, since there are four related practices:
  – *Data discovery, data profiling, data monitoring, and collaboration around data profiles*
  – *But it’s convenient for “data profiling” to mean all four of these.*
The Four Practice Areas of Data Profiling

• Data Discovery
  – *Identify new sources. Discover cross-system data overlaps, transformations, and inconsistencies. Find related data structures distributed across multiple IT systems. Find best source of data for a given purpose.*

• Data Profiling
  – *Develop a data inventory. Document dependencies across data elements. Serve up profiles that are directly usable in a variety of tools and projects.*

• Data Monitoring & Remediation
  – *Assess the state of data today. Track its evolution. Notify data stewards of changes and exceptions.*

• Collaboration
  – *Share profiles and related artifacts. Let both tech & biz users access, annotate, and improve them.*
Ten Best Practices for Data Profiling

• Just do it!
• Profile data thoroughly.
• Discover and profile new data sources.
• Profile data across multiple IT systems.
• Map data as you discover and profile it.
• Re-profile data as it evolves.
• Re-profile data periodically via data monitoring.
• Collaborate through data profiles.
• Support many practices with data discovery, profiling, and monitoring.
• Use a data profiling tool.
Just do it!

• Common excuses for skipping profiling altogether:
  – “We know which sources to tap and we know their data so well that we needn’t profile them.”
  – “We have a tight deadline, and data profiling is not a deliverable.”

• Lack of profiling puts projects at risk.
  – *Surprises pop up, and projects conclude late and over budget.*

• Just do data profiling, as a prerequisite and critical success factor for projects in data integration, data quality, MDM, BI, and so on.
Profile Data Thoroughly

• Don’t scrimp on data profiling
  – You can profile data early or late in a project timeline.
  – Earlier is definitely better for project success.

• Fallacious but common attitude:
  – “Running some ad hoc queries and SQL scripts tells us all we need to know.”

• Be thorough
  – Rely on data discovery to find all the potential sources
  – Distinguish similar data sources
    • E.g., “customer” will be defined differently in different systems
  – Document data meanings, relationships, mappings, etc.
Discover and Profile New Data Sources

• Don’t just profile known data sources.
  – Analyze data sources you may have forgotten, overlooked, or been unaware of.
  – Assume that new sources are appearing and old ones are changing, else you’ll miss opportunities.
  – Seek tool automation that assesses sources you’re exploring and suggests data meaning and relations.

• Keep digging, and you’ll find better sources.
  – Don’t be content with just any source for a data point.
  – Don’t stop when a source is found, ignoring other, better sources.
Profile Data across Multiple IT Systems

• Don’t profile just parts of tables or databases.
  – Statistics describing the occurrence of values and data types in a table column are useful, but not enough.
  – Instead, cross-system data discovery can operate on whole databases and reach across multiple systems.

• Don’t just inventory data elements.
  – Discover cross-system keys, business rules, data mappings, data exceptions, transformations, and so on.
  – This is essential for projects that are inherently cross-system, like MDM, CDI, DQ.
  – Cross-system data discovery also applies to database consolidation/migration, especially w/legacy systems.
Map data as you discover and profile it.

• Don’t just inventory data elements.
  – Profiles constitute a valuable inventory, best documented in a data glossary.
  – The glossary should also catalog relationships.
  – And it should catalog or develop mappings.

• Mappings are a big chunk of the work in developing data integration and data quality solutions.
  – Mapping is tedious and prone to error when manual.
  – Mapping is time consuming.
  – Avail yourself of any available software automation to make mappings faster and more accurate.
Re-profile Data as it Evolves

• Data profiling is not a one shot deal.
  – *Databases are alive. They grow, shrink, and evolve.*
  – *System owners change data structures and data definitions without telling you.*

• Thus, you must re-profile data sources periodically.
  – *Assess their current state.*
  – *Discover changes in sources that require adjustments to solutions for data quality, integration, MDM, BI, etc.*

• This is true whether you are working with:
  – *Source databases you’ve handled before*
  – *Targets databases that demand improvement*
Re-profile Periodically via Data Monitoring

• Monitoring differs from profiling and discovery:
  – *Data profiling and discovery tend to be development tasks that touch many potential data sources & targets.*
  – *Data monitoring re-profiles data touched daily as part of a deployed solution for DQ, DI, MDM, etc.*

• Data monitoring enables important tasks:
  – *Continuous improvement of data’s quality*
    • Key to DQ projects, plus high-value databases (as in BI)
  – *Identification of exceptions and changes*
    • Link monitoring to processing and managing these
  – *Automation via a software tool for daily data handling (whereas discovery and profiling tend to be manual)*
A quick aside about Data Remediation

• Data processing finds anomalous data that needs attention
  – *True of processes for data integration, data quality, data profiling, data monitoring, and so on.*
  – “Exceptions” are often logged in files that are rarely processed fully
  – Little follow up to see that exceptions are corrected, standardized, etc.
• Remediate data, don’t just log exceptions
  – *Data profiling & (especially) monitoring should lead to data remediation*
  – Else, how will data & associated business processes improve?
  – *Remediation is also key to data stewardship and governance*
• Look for tool functions that automate remediation
  – *Most remediation is manual, making its slow & inaccurate*
  – Remediation functions may be built into tools for DI, DQ, DP, monitoring
  – *Ideally, remediation functions should automate a workflow for discovering, documenting, routing for correction, and checking up on exceptions (to see that they got addressed).*
Collaborate through Data Profiles

• Poor practices limit collaboration via data profiles:
  – “We’ve always just documented source-data profiles in Microsoft Offices files, which we email to everyone.”
  – Documentation is the usual product of manual profiling.
  – Doc is hard to update, share, & apply to multiple projects.

• Tool automation enables collaboration:
  – Share data profiles, glossaries, mappings, etc. via an easily accessed repository.
  – Let technical and business people annotate the profiles, to enrich them and make them more broadly applicable.
  – Reuse profiles across multiple projects for greater productivity, accuracy, and consistency.
Support Many Practices with Data Discovery, Profiling, and Monitoring

• Data management practices
  – Data quality and data integration are prime beneficiaries
  – MDM, CDI, DW, BI and other inherently cross-system apps
  – Operational database migrations, consolidations, upgrades, collocations, and synchronizations

• Data management organizations
  – Broad programs for data governance, data stewardship, enterprise data architecture
  – Some competency centers need profiling specialists

• Business initiatives that are data intensive
  – Mergers and acquisitions, reorganizations, data as an enterprise asset programs, etc.
Use a Data Profiling Tool

• Mostly manual methods are inhibiting data profiling:
  – *Ad hoc queries, SQL scripts, visual perusals of data*
  – *Documentation that’s hard update or apply directly*

• Extending data profiling into discovery, monitoring, remediation, and collaboration requires tool(s).
  – *Manual discovery lacks automated mappings and cross-system relationship definition*
  – *Manual monitoring & remediation isn’t even a possibility*
  – *Manual collaboration just means hurling docs via email*

• Use the data discovery, profiling, and monitoring functions built into tools for DQ, DI, MDM, etc.

• Or augment these with a dedicated data profiling tool.
Recommendations

• Move up to the next level of data profiling.
  – *Do it more often and thoroughly, plus share the results*
  – *Embrace data discovery, monitoring, and collaboration*

• Realize that profiling is not a one-shot deal.
  – *Monitoring keeps profiles current & discovers changes*
  – *Remediate exceptions, don’t just find and log them*

• Eschew myopia!
  – *Discover new sources, learn new things about old data*

• Collaborate over data profiles.
  – *Solicit input from various technical & business people*
  – *Share profiles broadly for consistent data use and reuse*

• Prefer tool functions over mostly manual methods.
  – *So profiles are directly applicable to project deliverables*
  – *So you can profile often, thoroughly, and collaboratively*
Bibliography

• “Data Profiling Myths that need Busting,” a TDWI Research Roundup article by Philip Russom, online at: http://www.tdwi.org/Research/display.aspx?id=9038

• “Unifying the Practices of Data Profiling, Integration, and Quality (dPIQ),” a TDWI Monograph by Philip Russom, online at: http://www.tdwi.org/Research/display.aspx?id=8004
Cross-System Data Analysis and Quality
Exeros: Cross-System Discovery, Audit and Remediation

Corporate Data Landscape/Data Lineage

- RDBMS
- Flat Files
- Mainframe
- Spreadsheets
- Etc.

Discover → Audit → Remediate

Business Rules

Exceptions

Data relationship management
Exeros X-Profiler

What is unique?
Only solution on the market that:
• Automatically discovers primary-foreign keys
• Performs cross-source overlap analysis

X-Profiler
Column level profiling
Automatic discovery of primary-foreign keys
Cross-system overlap analysis of up to 20 systems simultaneously

Useful for
MDM ROI and Source Data analysis
Data Quality Assessments
X-Profiler

- Column level profiling
- Automatic discovery of primary-foreign keys
- Cross-system overlap analysis of up to 20 systems simultaneously

Useful for

- MDM ROI and Source Data analysis
- Data Quality Assessments

What is unique?

Only solution on the market that:

- Automatically discovers primary-foreign keys
- Performs cross-source overlap analysis
Exeros Discovery: Transformation & Discrepancy Discovery

What is unique?
DisCOVERS cross-system business rules and data exceptions by examining data values

Exeros Discovery:
Data analyst workbench
Automates discovery of:
cross-system business rules
metadata
data inconsistencies
Includes all X-Profiler capabilities
Discrepancy discovery of a snapshot in time

Useful for:
Detailed data mapping between 2 data sources
Exeros Discovery: Transformation & Discrepancy Discovery

What is unique?
Discovers cross-system business rules and data exceptions by examining data values

Exeros Discovery:
Data analyst workbench
Automates discovery of:
cross-system business rules
metadata
data inconsistencies
Includes all X-Profiler capabilities
Discrepancy discovery of a snapshot in time

Useful for:
Detailed data mapping between 2 data sources
Exeros Discovery: Transformation Discovery

Data analyst workbench
Automates discovery of:
- cross-system business rules
- metadata
- data inconsistencies

Detailed data mapping between 2 data sources
Includes all X-Profiler capabilities
Discrepancy discovery of a snapshot in time

What is unique?
Discovers cross-system business rules and data exceptions by examining data values

1% of Rows Map Incorrectly
- Row 6: Demo1: Expected Value = 3 Actual Value = 1
- Row 125: Demo1: Expected Value = 0 Actual Value = 3
- Etc…

Exeros Discovery: Data analyst workbench
Automates discovery of:
- cross-system business rules
- metadata
- data inconsistencies

Detailed data mapping between 2 data sources
Includes all X-Profiler capabilities
Discrepancy discovery of a snapshot in time

What is unique?
Discovers cross-system business rules and data exceptions by examining data values
What Complex Business Rules are Discovered from the Data?

- Scalar
  - One to one
  - Substring
  - Concatenation
  - Constants
  - Tokens
- Conditional logic
  - Case statements
  - Equality/Inequality
  - Null conditions
  - In/Not In
  - Conjunctions
- Joins
  - Inner
  - Left Outer
- Aggregation
  - Sum
  - Average
  - Minimum
  - Maximum
- Column Arithmetic
  - Add
  - Subtract
  - Multiply
  - Divide
- Reverse Pivot
- Cross-Reference
Validator: Ongoing Audit & Remediation of Critical Business Data Flow

Dataset 1

Dataset 2

data relationship management
Validator: Ongoing Audit & Remediation of Critical Business Data Flow
Validator: Ongoing Audit & Remediation of Critical Business Data Flow

Dataset 1

Validator

Dataset 2

EMPLOYEE_ID | AREA_CODE | PHONE_NUMBER
--- | --- | ---
'23' | '419' | '545-8796'
'1' | '206' | '555-9875'
'2' | '386' | '555-9482'
'3' | '704' | '555-3412'
'4' | '704' | '555-8122'
'5' | '710' | '555-4848'
'6' | '710' | '555-7773'
'7' | '710' | '555-5598'
'8' | '206' | '555-1189'
'9' | '710' | '555-4444'
'10' | '704' | '513-6894'
'11' | '704' | '513-5362'

EMPLOYEE_ID | AREA_CODE | PHONE_NUMBER
--- | --- | ---
'23' | '419' | '545-8769'
'1' | '206' | '555-9875'
'2' | '386' | '555-9482'
'3' | '704' | '555-3412'
'4' | '704' | '555-8212'
'5' | '710' | '555-4488'
'6' | '710' | '555-7737'
'7' | '710' | '555-5589'
'8' | '206' | '555-1819'
'9' | '710' | '555-4434'
'10' | '704' | '513-6984'
'11' | '704' | '513-5326'
Validator: Ongoing Audit & Remediation of Critical Business Data Flow

<table>
<thead>
<tr>
<th>EMPLOYEE_ID</th>
<th>AREA_CODE</th>
<th>PHONE_NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>'23'</td>
<td>'419'</td>
<td>'545-8796'</td>
</tr>
<tr>
<td>'1'</td>
<td>'206'</td>
<td>'555-9857'</td>
</tr>
<tr>
<td>'2'</td>
<td>'386'</td>
<td>'555-9482'</td>
</tr>
<tr>
<td>'3'</td>
<td>'704'</td>
<td>'555-3412'</td>
</tr>
<tr>
<td>'4'</td>
<td>'704'</td>
<td>'555-8122'</td>
</tr>
<tr>
<td>'5'</td>
<td>'710'</td>
<td>'555-4848'</td>
</tr>
<tr>
<td>'6'</td>
<td>'710'</td>
<td>'555-7773'</td>
</tr>
<tr>
<td>'7'</td>
<td>'710'</td>
<td>'555-5598'</td>
</tr>
<tr>
<td>'8'</td>
<td>'206'</td>
<td>'555-1189'</td>
</tr>
<tr>
<td>'9'</td>
<td>'710'</td>
<td>'555-4444'</td>
</tr>
<tr>
<td>'10'</td>
<td>'704'</td>
<td>'513-6894'</td>
</tr>
<tr>
<td>'11'</td>
<td>'704'</td>
<td>'513-5362'</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EMPLOYEE_ID</th>
<th>AREA_CODE</th>
<th>PHONE_NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>'23'</td>
<td>'419'</td>
<td>'545-8769'</td>
</tr>
<tr>
<td>'1'</td>
<td>'206'</td>
<td>'555-9875'</td>
</tr>
<tr>
<td>'2'</td>
<td>'386'</td>
<td>'555-4982'</td>
</tr>
<tr>
<td>'3'</td>
<td>'704'</td>
<td>'555-4312'</td>
</tr>
<tr>
<td>'4'</td>
<td>'704'</td>
<td>'555-8212'</td>
</tr>
<tr>
<td>'5'</td>
<td>'710'</td>
<td>'555-4488'</td>
</tr>
<tr>
<td>'6'</td>
<td>'710'</td>
<td>'555-7737'</td>
</tr>
<tr>
<td>'7'</td>
<td>'710'</td>
<td>'555-5589'</td>
</tr>
<tr>
<td>'8'</td>
<td>'206'</td>
<td>'555-1819'</td>
</tr>
<tr>
<td>'9'</td>
<td>'710'</td>
<td>'555-4434'</td>
</tr>
<tr>
<td>'10'</td>
<td>'704'</td>
<td>'513-6884'</td>
</tr>
<tr>
<td>'11'</td>
<td>'704'</td>
<td>'513-5326'</td>
</tr>
</tbody>
</table>
Validator: Ongoing Audit & Remediation of Critical Business Data Flow
Validator: Ongoing Audit & Remediation of Critical Business Data Flow
Case Study: Movement & Validation of a Loan

- Family buys home through their “Local Bank”
- Local Bank is a Lender and Servicer

- Loan guarantor guarantees the principle and interest on the loan
- Guarantor may also buy the loan from Local Bank

- Local Bank sells servicing rights to other servicers Bank2

Guarantor

Bank 2
Case Study: Mortgage Guarantor
Back Office Audit and Remediation of Loans

Data movement introduces data quality and consistency problem
Consistency degrades when data is copied from one company to another
Consistency degrades when data is copied around within a company
Current reconciliation approach is slow, expensive, semi-automated

Sample Data in Guarantor and Bank2 (IDs have been masked)

<table>
<thead>
<tr>
<th>Guarantor</th>
<th>LNDR_NO</th>
<th>GRNTR_LN_NO</th>
<th>AMRT_TERM</th>
<th>NOTE_RT</th>
<th>MTG_PMT</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guarantor</td>
<td>3594335</td>
<td>458397234</td>
<td>342</td>
<td>6.1180</td>
<td>501.07</td>
<td>…</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank2</th>
<th>LNDR_NO</th>
<th>GRNTR_LN_NO</th>
<th>AMRT_TERM</th>
<th>NOTE_RT</th>
<th>MTG_PMT</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank2</td>
<td>3594335</td>
<td>458397234</td>
<td>354</td>
<td>6.1180</td>
<td>501.07</td>
<td>…</td>
</tr>
</tbody>
</table>

Loan data must be validated between Bank2 and Guarantor
Case Study: Exeros Discovery & Validator Solution

Cross-system loan validation:

1 day to setup and configure
70 business rules discovered using Exeros Discovery
Rule automatically populated into Validator product
Validated 1MM rows in under 25 minutes
Provides remediation environment to manage exceptions
Exeros Discovery: Analyst workbench that automates cross-system data analysis and mapping

- Cross-system overlap and inconsistency analysis of up to 20 systems simultaneously
- Detailed discovery of business rules, transformation logic and inconsistencies between 2 data sources

Exeros X-Profiler: Cross system overlap analysis subset of Exeros Discovery

Exeros Validator: Operational platform that provides ongoing audit and remediation of business rules across your distributed data landscape
Questions??
Contact Information

• If you have further questions or comments:

Philip Russom, TDWI  
prussom@tdwi.org

Todd Goldman, Exeros  
todd@exeros.com

•