Data Cleansing

Why is Legacy Data "Dirty" ?

- Dummy Values,
- Absence of Data,
- Multipurpose Fields,
- Cryptic Data,
- Contradicting Data,
- Inappropriate Use of Address Lines,
- Violation of Business Rules,
- Reused Primary Keys,
- Non-Unique Identifiers, and
- Data Integration Problems

Steps in Data Cleansing

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

Parsing

Input Data from Source File

Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL

Parsed Data in Target File

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Correcting

Corrects parsed individual data components using sophisticated data algorithms and secondary data sources.

Correcting

Parsed Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Standardizing

Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.

Standardizing

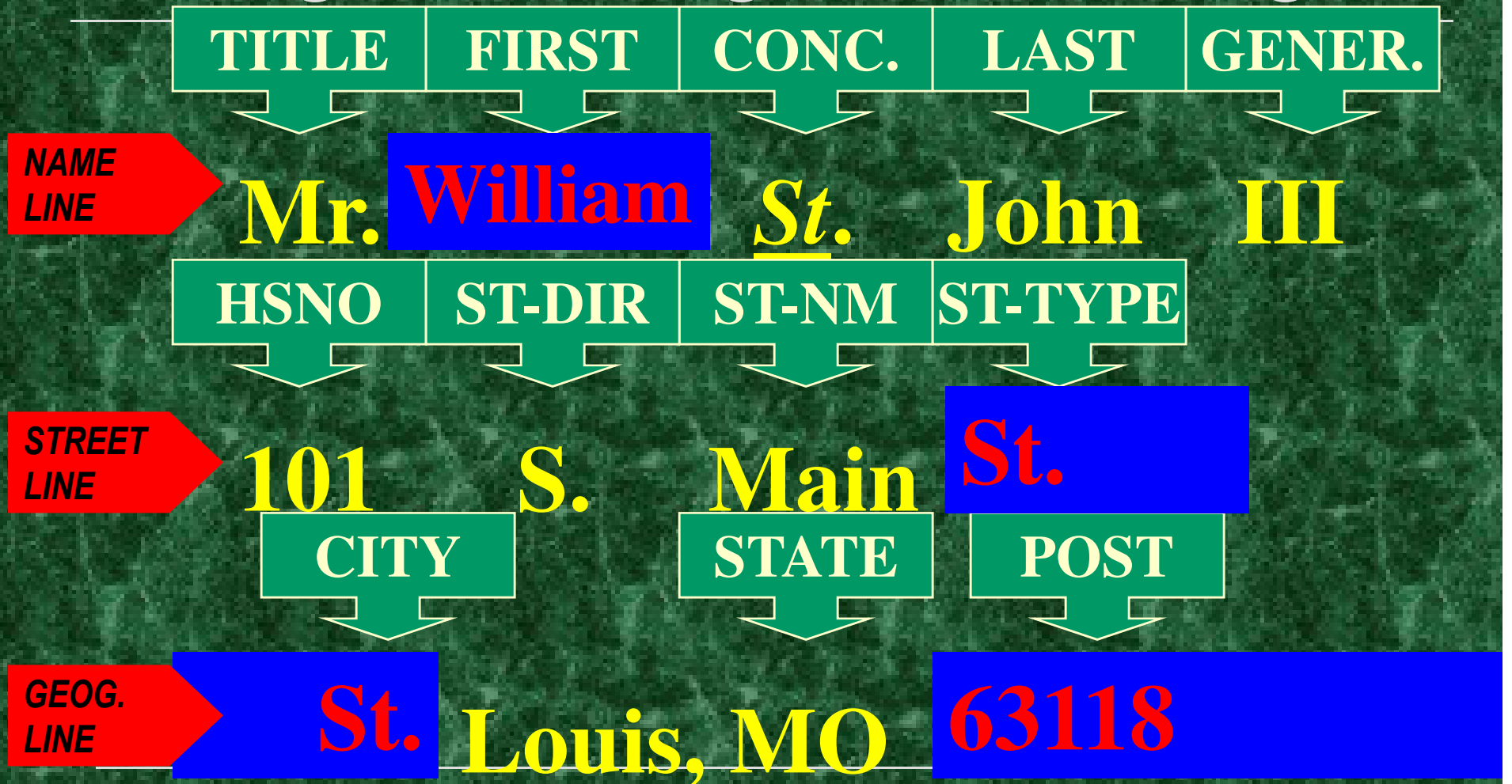
Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Corrected Data

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Parsing, Correcting, Standardizing



Matching

Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

Match Patterns

Business Name	Street	*Branch Type*	*Customer #/Tax ID*	City	*Vendor Code*	Pattern	Pattern I.D.
Exact	Exact	Exact	Exact	Exact	Exact	AAAAAA	P110
Exact	VClose	Exact	VClose	Exact	Blanks	ABAAA-	P115
Exact	VClose	Exact	Blanks	Exact	Exact	ABA-AA	P120
Exact	VClose	Close	Close	Exact	Exact	ABCCAA	S300
VClose	VClose	Exact	Close	Exact	Exact	BBACAA	S310

Matching

Corrected Data (Data Source #1)

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Corrected Data (Data Source #2)

Pre-name:	Ms.
First Name:	Elizabeth
1st Name Match Standards:	Beth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker-Lewis
Title:	
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr., Suite 2
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398
Phone:	708-555-1234
Fax:	708-555-5678

Consolidating

Analyzing and identifying relationships between matched records and consolidating/merging them into ONE representation.

Consolidating

Corrected Data (Data Source #1)

Corrected Data (Data Source #2)

Consolidated Data

Name:	Ms. Beth (Elizabeth) Christine Parker-Lewis
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Address:	12800 S. Butler Dr., Suite 2 Chicago, IL 60633-2398
Phone:	708-555-1234
Fax:	708-555-5678

Consolidating

Legacy Systems View (3 Clients)

Account No.
83451234

Policy No.
ME309451-2

Transaction
B498/97

The Reality – ONE Client