Informatica® SSA-NAME3
10.5

Population Override Manager

Informatica SSA-NAME3 Population Override Manager
10.5
September 2022

© Copyright Informatica LLC 1999, 2023

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2023-09-28

Table of Contents

Preface 6

Learning About Informatica SSA-NAME3. 6

What Do I Read If. 8

Informatica Resources. 9

Informatica Network. 9

Informatica Knowledge Base. 9

Informatica Documentation. 9

Informatica Product Availability Matrices. 9

Informatica Velocity. 10

Informatica Marketplace. 10

Informatica Global Customer Support. 10

Chapter 1: Introduction. 11

Overview. 11

Usage Warning. 12

Population Components that can be Overridden. 12

Edit-List Rules. 12

Frequency Tables. 13

Scalar Frequency Tables. 13

Matching Purposes. 14

z/OS Limitations. 14

Chapter 2: Starting Population Override Manager. 15

Step 1: Starting the Population Override Manager. 15

Step 2: Connecting to SSA-NAME3. 17

Step 3: Selecting a System and Population. 18

Step 4: Selecting Overrides. 18

Step 5: Selecting Field Overrides. 19

Step 6: Selecting Purpose Overrides. 20

Chapter 3: Purpose Operations. 21

Cloning a Purpose. 21

Make the required selections. 22

Clone the Purpose. 22

Editing a Purpose. 22

Select the Purpose to be Edited. 22

Edit the Purpose. 23

Saving the Purpose. 25

Renaming a Purpose. 25

Deleting a Purpose. 25

Chapter 4: Overriding Frequency Tables. 27
Frequency Tables. 27
Scalar Frequency Tables. 28
What Data to Use. 28
How to Override the Frequency Table. 28
How to Override the Scalar Frequency Table. 30

Chapter 5: Edit-List Overrides. 32
Using Edit-List Overrides. 32
The Search Bar. 32
Disabling a Rule. 34
Adding a New Edit-List Override. 35
Deleting an Override. 38
Saving, Testing and Committing Overrides. 38

Chapter 6: Rule Definitions. 40
Rule Definitions Overview. 40
Edit-List Processing Overview. 40
Edit-List Categories. 41
Common Categories for Word Rules. 41
Common Categories for Prefix-Suffix Rules. 42
Phrase Replacement Rules. 43
Character Replacement Rules. 43
Account Name Markers. 43
Compound Name Markers. 44
Delete Markers. 44
Major Markers. 44
Defining a New Category. 44
Rule Types. 45
Word Rule Types. 45
Prefix-Suffix Rule Types. 48
The Phrase Replacement Rule Type. 49
The Character Rule Type. 51
The Account Name Marker Rule Type. 52
The Compound Name Marker Rule Type. 52
The Delete Marker Rule Type. 52
The Major Marker Rule Type. 53
Edit-List Processing Phases. 53

Chapter 7: Implementing a Local Population. 55
Order of Precedence. 55
What to Do After Committing a Local Population. 55

Managing the Edit Rule Wizard. 56
Size of a Local Population. 56
Upgrading a Local Population using the latest Standard Population. 56
Using the SSAPRUPG Utility. 57

Index. 58

Preface

Welcome to the Informatica SSA-NAME3 Population Override Manager - a Java GUI tool. This tool allows a trained data analyst to override some of the Standard Population rules that are supplied with the product, or provided in the form of a Custom Population.

Learning About Informatica SSA-NAME3

This section provides details of documentation available with the SSA-NAME3 product.

Introduction to SSA-NAME3

Provides an overview of SSA-NAME3. It is written in a way that can be read by someone who has no prior experience of the product and wants a general overview of SSA-NAME3. It explains the problems SSA-NAME3 overcomes and provides an overview of how this is done. One chapter is dedicated to providing an overview for Application Programmers.

Getting Started

This manual is intended to be the first technical material a new developer or designer reads before installing or using the SSA-NAME3 software, regardless of the platform or environment. Its goal is to help a new user get the software installed and produce a working prototype application that calls SSA-NAME3 and executes searches against their own data.

To achieve this it provides a "script" to follow which includes pointers to pertinent sections of the other manuals.

Application & Database Design

This manual contains tips and techniques useful for setting up and optimizing a name search and matching application, including database issues, and illustrates best-practice techniques, common pitfalls, and strategies regarding the subject of name and address matching.

Installation Guide

This manual provides information on how to install the SSA-NAME3 product.

SSA-NAME3 Workbench User Guide

This is a guide to using the SSA-NAME3 Workbench - a Java GUI tool that helps a programmer understand and prototype SSA-NAME3 calls. The Workbench is also used for:

- Generating Sample Program Code;
- Executing SSA-NAME3 Calls;
- Testing different SSA-NAME3 run-time options;
- Producing debugging and support information for Informatica Corporation

Note: The Workbench in itself is not a search and match application. It assists the developer build a search and match application.

API Reference

The ultimate goal of an SSA-NAME3 implementation is for application programs to be able to call SSA-NAME3's API Functions to build keys and search strategies and to compute match scores and decisions.

This manual describes a typical program process flow for building an identity search application, and also lists in detail each of the API Functions. It describes the parameters required by these functions and the information returned.

Population Override Manager User's Guide

This is a guide to using the SSA-NAME3 Population Override Manager - a Java GUI tool that allows a trained data analyst to override some of the Standard Population rules that are supplied with the product, or provided in the form of a Custom Population. The types of rules that can be overridden using this tool are:

- Edit-list rules
- Frequency tables
- Scalar Frequency Tables
- Matching Purposes

Note: Use of this tool without proper training from Informatica should not be attempted, as improper use can adversely affect the reliability and performance of the search application(s).

Edit Rule Wizard User's Guide

This is a guide to using the SSA-NAME3 Edit Rule Wizard - a Java GUI tool that helps a business user safely add certain types of Edit Rules to the Standard or Custom Population without requiring specific knowledge of SSA-NAME3 or support from a programmer or data analyst. The types of rules that can be added using this tool are:

- Discard a word or phrase when searching and matching (e.g. a new "noise" word)
- Add a new replacement word or phrase when searching and matching (e.g. a new "abbreviation", "nickname" or "acronym")
- Add a new compound name marker word

Release Notes

The Release Notes contain information about what's new in this version of SSA-NAME3. It is also used to summarize any documentation updates as they are published.

What Do I Read If. . .

I am. . .

. . . a business manager

The INTRODUCTION TO SSA-NAME3 will address questions such as "Why have we got SSA-NAME3?", "What does SSA-NAME3 do"?

I am. . .

. . . a system designer or DBA

The INTRODUCTION TO SSA-NAME3 will address questions such as "What resources are needed to implement SSA-NAME3?". The APPLICATION & DATABASE DESIGN manual will lead you through many of the design considerations of name search and matching applications.

I am. . .

. . . installing SSA-NAME3

Before attempting to install SSA-NAME3 you should read the Getting Started document. This will describe the pre-requisites and help you plan the installation and implementation of SSA-NAME3. The actual installation steps for your platform are documented in the Installation Guide.

I am. . .

. . . an Analyst or Application Programmer

A high-level overview is provided specifically for Application Programmers in the INTRODUCTION TO SSA-NAME3 manual. Before attempting to develop programs that interface with SSA-NAME3, you should also read the GETTING STARTED and APPLICATION & DATABASE DESIGN manuals, as well as experimenting with calls in the WORKBENCH USER GUIDE.

When developing the application program(s), use the API REFERENCE manual which describes a typical application and the Function parameters.

Working example programs that illustrate the calls to SSA-NAME3 in various languages are available by using the Sample Program button on the Workbench.

I want to know. . .

. . . what SSA-NAME3 does

The INTRODUCTION TO SSA-NAME3 manual gives an overview of what SSA-NAME3 does and how it does it.

I want to know. . .

. . . how to setup the database

Refer to the APPLICATION & DATABASE DESIGN manual for tips and techniques on configuring the database to store SSA-NAME3 Keys and optimizing it for searching and matching.

I want to know. . .

. . . how to code a search application

The INTRODUCTION TO SSA-NAME3 manual contains a specific section designed to get application programmers familiar with the concepts of developing an SSA-NAME3 search and match application.

The API REFERENCE GUIDE details the Function calls required and their parameters. The SSA-NAME3 WORKBENCH USER GUIDE shows how to generate a sample program in a variety of programming languages.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Network

The Informatica Network is the gateway to many resources, including the Informatica Knowledge Base and Informatica Global Customer Support. To enter the Informatica Network, visit https://network.informatica.com.

As an Informatica Network member, you have the following options:

- Search the Knowledge Base for product resources.
- View product availability information.
- Create and review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit https://search.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

If you have questions, comments, or ideas about the product documentation, contact the Informatica Documentation team at infa_documentation@informatica.com.

Informatica Product Availability Matrices

Product Availability Matrices (PAMs) indicate the versions of the operating systems, databases, and types of data sources and targets that a product release supports. You can browse the Informatica PAMs at https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services and based on real-world experiences from hundreds of data management projects. Informatica Velocity represents the collective knowledge of Informatica consultants who work with organizations around the world to plan, develop, deploy, and maintain successful data management solutions.

You can find Informatica Velocity resources at http://velocity.informatica.com. If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that extend and enhance your Informatica implementations. Leverage any of the hundreds of solutions from Informatica developers and partners on the Marketplace to improve your productivity and speed up time to implementation on your projects. You can find the Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through the Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
https://www.informatica.com/services-and-training/customer-success-services/contact-us.html.

To find online support resources on the Informatica Network, visit https://network.informatica.com and select the eSupport option.

CHAPTER 1

Introduction

This chapter includes the following topics:

- Overview, 11
- Usage Warning, 12
- Population Components that can be Overridden, 12
- z/OS Limitations, 14

Overview

The default rules which govern how the SSA-NAME3 key-building, search strategy and matching routines operate for a given population of data are contained within files called Standard Populations. The SSA-NAME3 product includes Standard Populations for over 50 countries and languages. These are the files with an extension of `.ysp`.

Occasionally, customers with unusual needs may have a Custom Population built for them by Informatica Corporation, either using an existing Standard Population as a base, or from scratch if dealing with a totally new type of data. These files have an extension of `.ycp`.

These Population rule files are not viewable as text files and can only be accessed via the SSA-NAME3 callable routines (i.e. from an application program), or via one of the packaged GUI clients:

- The Developer'sWorkbench
- The Population Override Manager
- The Edit RuleWizard

While the SSA-NAME3 Standard Populations will satisfy the majority of customers' needs, occasionally a user will have the need to override, disable, or add new rules because of the specific local content of their data.

The Population Override Manager is a Java GUI tool that allows a trained data analyst to override some of the Standard Population rules that are supplied with the product, or provided in the form of a Custom Population.

An example of this need might be to add a new rule to cater for some specific types of noise words found in the user's data, but not commonly found in data representative of the population. For example, **XXX** has been used by a company's data entry operators to denote that an address has an unknown street number. In this example, a data analyst trained in the use of the Population Override Manager could add a new noise word rule for **XXX**. This activity will result in the generation of a Local Population.

Usage Warning

When the Population Override Manager GUI is started, the user is shown the following warning message:

Important: Improper use of this program can result in damage to the search and matching rules currently in use in your organization. The features of this program should only be used by a specialist trained in the activity of building and maintaining rules for SSA-NAME3 populations. This specialist should be aware that his or her actions can have adverse effects on existing/production systems. Do you want to continue?

The reason for this message is to make it clear that care should be taken when adding, removing or overriding rules in a Population.

The data analyst doing this job should be trained by Informatica Corporation in the use of this tool.

Some considerations are:

- Adding a rule to fix a specific case for a specific search may adversely affect the quality of other types of searches. For example, One user doesn't like the fact that IVAN is a nickname for JOHN and asks for the rule to be removed. This may suit the *customer lookup* search, but now the fraud search potentially misses valuable matches. (This case may be better handled by using different Match Levels.)
- Adding certain types of rules requires that the SSA-NAME3 keys should be re-built and re-indexed.

Population Components that can be Overridden

While there are many components that make up a Population rule set, only the following components currently accept overrides from the Population Override Manager.

- Edit rules
- Frequency tables
- Scaler Frequency tables
- Matching Purposes

Edit-List Rules

An Edit-list is a lookup table containing rules ("Edit rules") about the commonly occurring words and phrases in a population of data.

During Key Building, Range Building and Matching, the components (words, codes and numbers) of a name or address are checked against entries in the Edit-List. If an exact string match is found the associated rule is invoked.

Edit-List rules are used to address commonly occurring variations and noise in name and address data. They are not used to address phonetic and orthographic error - this is handled by other components of the Population rule set.

Edit-List rules are used to address the following types of noise and variations: noise words, personal titles, company titles, nicknames, alternate forms of words, abbreviations, phrase replacements (e.g. acronyms), prefix and suffix variations, street types and more.

Each field type that can be used for building keys or search ranges has an Edit-List associated with it. In fact, each field type will have its own Edit-List. In most Standard Populations, this applies to the following three field types:

- Person Names
- Organization Names
- Street Addresses

In addition, some fields that are not used for Key or Range building, but are used for Matching, also have Edit-Lists. Examples of these are:

- ID numbers
- Telephone Numbers
- Postal Codes

These types of fields are called Efields. Efields benefit from Edit-Lists to overcome such problems as when an ID number that contains all 9's should be considered a "null" number. Therefore, a rule would be required to treat all 9's as a noise word.

Frequency Tables

The SSA-NAME3 Frequency tables are used in the Key and Search Range building functions of SSANAME3. They are used by these Algorithms to generate keys that allow for improved performance for common name searches. As such, they are only used for field types that are available for key or search range building. These are typically:

- Person Names
- Organization Names
- Street Addresses

Most Standard Populations include pre-packaged Frequency tables built from data representative of that particular population. For the majority of SSA-NAME3 users, these packaged Frequency tables are adequate for providing the right balance of reliability and performance.

However, in some applications where the volume of data is very large, where high performance is critical, or where the data content is thought to be very unrepresentative of the data used to build the frequency tables for the Standard Population, there may be some benefit in overriding the Frequency tables using the organization's own data.

In other cases, where the search reliability is critical (i.e. there is a very high level of risk associated with missing matches), it may be beneficial to remove the Frequency table from the Standard Population. This will typically lower the performance of the search, but can improve reliability at the edge for a search (i.e. where the boundary exists between suspect matches and non-matches).

The Population Override Manager allows the data administrator to override or remove a Frequency table.

To discuss your particular needs in this area, call an Informatica Corporation Technical Representative.

Scalar Frequency Tables

The SSA-NAME3 Scalar Frequency Tables are used to generate a run-time scale "value" for a particular search criteria and strategy. This scale value is an estimate of the percentage of the file that may be retrieved for a particular search, given the search data, field type and search level used.

An application program, via the SSA-NAME3 API `ssan3_get_ranges` Control parameter `SEARCH_LIMIT`, may impose a "scale warning limit" for a particular search such that if that scale is exceeded, the program will

receive a return code along with the search key ranges and be able to go back to the searcher with a message that the search may be too expensive, prior to executing the database I/O.

As Scalar Frequency Tables take up a significant amount of memory, they are not packaged in the Standard Populations. If a user wishes to make use of the `SEARCH_LIMIT` control, then a Scalar Frequency Table will need to be generated from the organization's own data for the required Field(s). This is done using the Population Override Manager.

Matching Purposes

All Standard Populations include a number of standard Matching Purposes. A Matching Purpose determines how the Match API (ssan3_match) interprets the user data passed to it. The Matching Purposes packaged in the Standard Populations have been designed to cater for the most common Matching needs when comparing entities such as names, addresses, etc. Occasionally it may be desirable to modify one or more of the settings defined by one of these Standard Matching Purposes. The Population Override Manager provides the ability to Clone (Copy) an existing purpose. This new Purpose can then be modified to suit the user's specific needs. Total count of Cloned and Standard purposes is limited to 36.

z/OS Limitations

The Population Override Manager and Edit Rule Wizard can not currently be used with an SSANAME3 server that is running under IBM or z/OS. However, they can be used with an SSA-NAME3 server that is running under Unix System Services on the z/OS mainframe.

CHAPTER 2

Starting Population Override Manager

This chapter includes the following topics:

- Step 1: Starting the Population Override Manager, 15
- Step 2: Connecting to SSA-NAME3, 17
- Step 3: Selecting a System and Population, 18
- Step 4: Selecting Overrides, 18
- Step 5: Selecting Field Overrides, 19
- Step 6: Selecting Purpose Overrides, 20

Step 1: Starting the Population Override Manager

- You have installed SSA-NAME3.
- You have started the SSA-NAME3 server.
- Standard Population on which the Population Override Manager will operate is available.

  Note: The Population Override Manager requires the SSA-NAME3 server to be running.

To start the Population Override manager, complete the following steps.

1. On a Win32 platform, select **Start** > **Informatica** > **Population Override Manager**.

2. Click **Enter Population Override Manager** to launch Population Override Manager.

   The table below has list of menus available on the the Population Override Manager entry screen.

Menu	Description
Introduction to SSA-NAME3	You use this option to launch a document that provides a short introduction to SSA-NAME3. It can be read by someone who has no prior experience of the product and wants a general overview about the SSA-NAME3. It explains the problems SSA-NAME3 is designed to overcome and the approaches it uses. There is also a chapter dedicated to providing an overview for Application Programmers.
Documentation	You use this option to launch document *POPULATION OVERRIDE MANAGER USER'S GUIDE*.
All Systems Information	You use this option to launch document that lists the available systems and populations that have been installed, along with information about when these were last updated.
Enter Population Override Manager	You use this option to launch **Population Override Manager**
Exit	You use this option to close **Population Override Manager**

Step 2: Connecting to SSA-NAME3

To connect to SSA-NAME3, complete the following steps.

1. In the **SSA-NAME3 Population Override Manager- Connection** window, enter **Host Name** and **Port Number**, where the server is running.

   Note: The default port used by SSA-NAME3 is 1665.

2. Click **Connect**.

Step 3: Selecting a System and Population

To select a System and Population, complete the following steps.

1. In the **SSA-NAME3 Population Override Manager- Connection** window, select **System** and **Population**.

   Note: The System name is the name of **pr** sub-directory where the Standard Populations (files with an extension of `.ysp`) reside. The System called **default** is where the installer copies the Standard Population files from the installation CD. You may also have one or more Custom Populations (files with an extension of `.ycp`). It is recommended that you copy the Standard Population Files you will be using into a new **pr** subdirectory. The name of that sub-directory could be a name relating to the business system or project name. It is also recommended to have different directories/systems for different phases in the project development (For example,. "Dev", "Test", "QA", "Production").The available Populations will be those you chose to install. Select the most suitable Population for your data.

2. Click **OK**.

Step 4: Selecting Overrides

To select the type of Overrides, complete the following steps.

1. In the **SSA-NAME3 Popoulation Override Manager- Overrides Selection** window, select one of the Overrides type.
   - Field Overrides
   - Purpose Overrides

2. Click **OK**.

Step 5: Selecting Field Overrides

You have selected **Field Overrides** in the **SSA-NAME3 Popoulation Override Manager- Overrides Selection** window.

To work with the Field Overrides, complete the following steps:

1. In the **SSA-NAME3 Popoulation Override Manager- Field Selection** window, select **Field**.

   Note: The **Field** drop-down list contains the fields in the Population that have Edit-Lists associated with them.

2. Select one of the options mentioned below:
   - Override Frequency-Table
   - Override Scalar-Frequency-Table
   - Override Edit Rules.

   Note:

   - Fields that are used for key and search range building (For example, **Person_Name**, **Organization_Name** and **Address_Part1**) will also allow overriding of their Frequency Table and Scalar Frequency Table.
   - By default, the **Override Edit-Rules** radio button is selected.
3. Click **OK**.

Step 6: Selecting Purpose Overrides

You have selected Purpose Overrides in the **SSA-NAME3 Popoulation Override Manager- Overrides Selection** window.

To work with Purpose Overrides, complete the following steps:

1. In the **SSA-Name3 Select Purpose** window, select **Purpose**.

   Note:

   - You can Clone Standard Purposes.
   - You can edit, rename and delete only Cloned Purposes.

2. Choose one of the options below and perform appropriate actions as mentioned in the *Purpose Operation* section.

   - **Clone**
   - **Edit**
   - **Rename**
   - **Delete**

CHAPTER 3

Purpose Operations

This chapter includes the following topics:

- Cloning a Purpose, 21
- Editing a Purpose, 22
- Renaming a Purpose, 25
- Deleting a Purpose, 25

Cloning a Purpose

Having chosen the Clone Purpose function from the dialog described in the section Purpose, the following dialog will be displayed:

Make the required selections

Purpose to be Cloned

Select the name of the existing Purpose to be cloned. The default value of this field is the name of the Purpose selected on the Purpose Selection Dialog.

Type of Clone Operation

There are 2 possibilities. **Clone all Match Levels** will make an exact copy of the existing purpose including all Match Levels (Typical, Loose & Conservative). The resultant Purpose will behave identically to the original Purpose.

Create one combined Match Level will create a purpose which will contain a single Match Level (Typical). This single Match Level will be a combination of the Typical, Loose & Conservative Match Levels of the original Matching Purpose. Which Level is used for each field is determined by the Field Table at the bottom of the dialog.

New Purpose Name

Enter the name of the new Purpose. This name may be a maximum of 32 characters and must not contain spaces.

Select Required Fields and Levels

It is possible to exclude fields from the new Purpose. This allows the user to determine which fields will be present in the new Purpose. To exclude a field from the new Purpose, uncheck the checkbox next to the field concerned in the Include Field column.

Also, if you have chosen the **Create one combined Match Level** option above, you may wish to choose the Match Level to be used for each field which will be present in the new Purpose. Do this by selecting the required Match Level for each field from the list of Match Levels appearing in the **Match Levels** column.

Clone the Purpose

Once you have made the desired selections above, click the **Clone Purpose** button. The new Purpose will be created and will be available for use by the **ssan3_match** API.

Editing a Purpose

Select the Purpose to be Edited

On the **Purpose Selection** Dialog, select the name of the Purpose to be edited. Then click the **Edit** button. Note that Purposes from the original Standard Population may not be edited. Therefore, before it is possible to edit a Purpose, at least one Cloned Purpose must be created.

Edit the Purpose

Click the **Edit** button to begin editing the Purpose. The following dialog is displayed.

This dialog presents all the user-modifiable settings within the selected purpose. The dialog consists of three main panels - the **Purpose Limits** panel, the **Field Information** panel and the **Purpose Description** panel.

Purpose Information Panel

This panel contains the Accept & Reject Limits for all Matching Levels in this purpose. These will normally be Typical, Conservative & Loose. If this Purpose was cloned using the Create one combned Match Level option, then there will be a single Typical Matching Level.

Accept-Limit: This field is required and must be in the range 0-100

Reject-Limit: This field is required and must be in the range 0-100

Refer to the API Reference manual for further details on Accept & Reject Limits.

Field Information

On the left hand side of this panel is a list of the fields which make up the chosen Purpose. Initially, the first field in the list is selected. The user may select a different field simply by clicking on it. On the right hand side are the settings which may be modified for the currently selected (highlighted) field. The available settings are:

Required: If selected, indicates that this field must be present during the call to *ssan3_match*. If **Required** is selected and this data item is not supplied to *ssan3_match* then an error will result. **Required** is mutually exclusive with **Optional**.

Optional: If selected, indicates that this field may be absent during the call to **ssan3_match**. Note that the absence of an **Optional** field will not degrade the score. **Optional** is mutually exclusive with **Required**.

Weight: This setting determines the weight of this field relative to the other fields in the purpose.

For example, when matching a Person_Name field with WEIGHT=2 and a DATE field with WEIGHT=1, then the Person_Name field will contribute two-thirds or 66% of the final overall score for the match.

Best Group: This is an optional setting. If specified, this must be a single alpha character which identifies the Best Group to which this field is to belong. One or more fields may belong to the same Best Group. When calculating the score, the best score in the group is used. The lower scores are discarded and the weight for each discarded field is set to 0. As only one field will contribute to the final score, then all members of the group should have the same weight specified.

Required Group: This is an optional setting. If specified, this must be a single alpha character which identifies the **Required Group** to which this field is to belong. One or more fields may belong to the same **Required Group**. The usage of this field is described in the following Count section.

Count: This setting is used in conjunction with Required Group above. Count is a numeric value which indicates the number of fields which should be present in this group. If **Count** fields of the group are not present, then, for all fields in this **Required Group**, all scores are set to 0 and all weights are set to maximum.

If **Required Group** is not present, then the Count setting is ignored.

Score Contribution - Contribution Threshold: These two settings allow control over whether the score for this field should contribute to the overall score. If the calculated score for this field is above or below **Contribution Threshold**, then this field is excluded from the final score calculation, ie. the weight for this field is set to zero.

The default behavior is to allow the field's score to contribute to the final score, in which case the **Contribution Threshold** setting is ignored.

Score Accentuation - Accentuation Threshold: These two settings allow finer control over the score for this field. If the calculated score for this field is below **Accentuation Threshold**, then decrease its score to 0. Alternatively, if the calculated score for this field is above **Accentuation Threshold**, then increase its score to 100.

The default behavior is to always use the field's original score, in which case the **Accentuation Threshold** setting is ignored.

Reference Record: This setting determines which Record will be used as the reference while calculating the score for this field. The options are:

- Always use the File Record as the Reference
- Always use the Search Record as the Reference
- Use the Record with the most number of words as the Reference
- Use the Record with the least number of words as the Reference

For example, when matching the names KEN JOHN PEEL and JOHN PEEL, if JOHN PEEL is chosen as the Reference Record then these two names will score 100 even though there is an extra word in the first name, whereas if KEN JOHN PEEL is chosen as the Reference Record then these two names will score 66.

Null Value Effect - Null Value Score: This setting determines how a null value supplied for this field will contribute to the overall score. The default behavior is that a null value will cause this field to be disregarded when calculating the overall score. Alternatively, a null value will cause a score equal to **Null Value Score** to be used for this field when calculating the overall score.

Purpose Description

This is a free form comment area provided for the user to describe the intended use of the Purpose and record the history of any changes made to the Purpose.

Saving the Purpose

When you have made the required changes to the Purpose, click the **Save Purpose** button. This will save the new Purpose Settings into the Population. The updated Purpose is immediately available for use.

Renaming a Purpose

On the **Purpose Selection** Dialog, select the name of the Purpose to be renamed. Then click the **Rename** button. The following dialog will appear:

Enter the **New Purpose Name** and click **OK**. The Purpose will be renamed and a confirmation dialog will be displayed.

Note: Purposes from the original Standard Population may not be renamed. Only Cloned Purposes may be renamed.

Deleting a Purpose

On the **Purpose Selection** dialog, select the name of the Purpose to be deleted. Then click the **Delete** button. The following dialog will appear:

Click **Yes** to confirm that the Purpose should be deleted. The Purpose will be deleted and a confirmation dialog will be displayed.

Note: Purposes from the original Standard Population may not be deleted. Only Cloned Purposes may be deleted.

CHAPTER 4

Overriding Frequency Tables

This chapter includes the following topics:

- Frequency Tables, 27
- Scalar Frequency Tables, 28
- What Data to Use, 28
- How to Override the Frequency Table, 28
- How to Override the Scalar Frequency Table, 30

Frequency Tables

The SSA-NAME3 Frequency tables are used in the Key and Search Range building functions of SSANAME3. They are used by these Algorithms to generate keys that allow for improved performance and reduced selectivity for common name searches. As such, they are only used for field types that are available for key or search range building. These are typically:

- Person Names
- Organization Names
- Street Addresses

Most Standard Populations include pre-packaged Frequency tables built from data representative of that particular population. For the majority of SSA-NAME3 users, these packaged Frequency tables are adequate for providing the right balance of reliability and performance.

However, in some applications where the volume of data is very large, where high performance is critical, or where the data content is thought to be very unrepresentative of the data used to build the frequency tables for the Standard Population, there may be some benefit in overriding the Frequency tables using the organization's own data.

In other cases, where the search reliability is critical (i.e. there is a very high level of risk associated with missing matches), it may be beneficial to remove the Frequency table from the Standard Population. This will typically lower the performance of the search, but can improve reliability at the edge for a search (i.e. where the boundary exists between suspect matches and non-matches).

The Population Override Manager allows the data administrator to override or remove a Frequency table.

Scalar Frequency Tables

The SSA-NAME3 Scalar Frequency Tables are used to generate a run-time scale "value" for a particular search criteria and strategy. This scale value is an estimate of the percentage of the file that may be retrieved for a particular search, given the search data, field type and search level used.

An application program, via the SSA-NAME3 API **ssan3_get_ranges** Control parameter SEARCH_LIMIT, may impose a "scale warning limit" for a particular search such that if that scale is exceeded, the program will receive a return code along with the search key ranges and be able to go back to the searcher with a message that the search may be too expensive, prior to executing the database I/O.

The Scalar Frequency Tables are populated with the high-frequency SSA-NAME3 fuzzy keys generated from a population of names.

As Scalar Frequency Tables take up a significant amount of memory, they are not packaged in the Standard Populations. Therefore, careful consideration should be given need for the SEARCH_LIMIT control in balance with the memory requirements.

If a user wishes to make use of the SEARCH_LIMIT control, then a Scalar Frequency Table will need to be generated from the organization's own data for the required Field(s). This is done using the Population Override Manager.

What Data to Use

To be effective and valuable, the Frequency and Scalar Frequency tables must be generated from data that is truly representative of the production names or addresses that will be encountered in the system.

You will need a file or files containing representative data for each field (Person_Name, Organization_ Name, Address_Part1) that is to have its frequency table overridden.

The file(s) need only contain a maximum of 250,000 records provided this is the full data set, or else has been randomly selected from the full data.

The Population Override Manager frequency table override facility allows the user to specify an offset and length into a file where the column containing the raw data (names or addresses) is to be found.

How to Override the Frequency Table

On the following screen, choose the Field name and select the **Override Frequency Table** option.

Click **OK**, the following screen is displayed.

Enter the file name that contains the names, or if overriding the `Address_Part1` field, the file that contains the address data up to, but not including, locality/state/postcode information. If the file contains the relevant data in a certain column, enter the appropriate field offset and field length. If the data is stored in Unicode, choose the appropriate Unicode encoding.

The above screen shows a file called `person_names.txt` containing the person name data at offset 12 for a length 50.

Click **OK** to start the frequency table generation process - the following windows display the progress and completion messages.

Click **OK** on the "Completed" message window to return to the field selection screen. At this point, an updated Local Population has been created containing the new frequency table.

How to Override the Scalar Frequency Table

On the following screen, choose the Field name to generate the Scalar Frequency Table for, and select the **Override Scalar Frequency Table** option.

Click**OK**, the following screen is shown:

Enter the file name that contains the names, or if overriding the Address_Part1 field, the file that contains the address data up to, but not including, locality/state/postcode information. If the data is stored in Unicode, choose the appropriate Unicode encoding.

If the file contains the relevant data in a certain column, enter the appropriate field offset and field length.

Maximum Keys defines the maximum number of entries in the table. If this field is left as "0", all keys generated from the file of names will be used in descending frequency up until the value specified in Minimum Frequency is reached. A value other than "0" will cause the table to be truncated to that number of keys, in descending frequency. This allows some additional control over the size of the generated Local Population. However, if the size of the generated Local Population is not an issue, this value should be left as "0".

Minimum Frequency defines the frequency value above which keys will be added to the table. To choose a value for this setting, first establish the approximate size of the database of names that will be being searched in the production system, then determine the percentage of this file above which a search warning should be generated. For example, the production database contains 1,000,000 entries and the requirement is to receive a search warning if more than 500 candidates (0.05% of the file) are returned for a search.

Now, translate that 0.05% to the size of the input file to the Scalar Frequency Table generation process. So, for example, if there are 50,000 names in the input file, 0.05% is 25. Therefore, enter 25 as the Minimum Frequency.

Note: Setting this value too low will increase the size of the Local Population significantly.

CHAPTER 5

Edit-List Overrides

This chapter includes the following topic:

- Using Edit-List Overrides, 32

Using Edit-List Overrides

When launched, the Edit-List override screen as shown appears:

The Search Bar

The search bar is used for searching for existing rules about a specific word or phrase.

For example, if you want to search for all of the Edit rules for the word Tom. Enter Tom in the search bar and click **Search**.

The following screen appears:

Note: The search will return all rules related to the Tom rule, including rules related to rules related to Tom, thus performing a drill-down search.

If you select the **Hex** checkbox, the following screen appears with the hexadecimal equivalent of the string:

Disabling a Rule

To disable a rule, the rule must first be found via a Search.

Continuing with the Tom example, if you want to disable the **Secondary Name rule TAMI =>THOMAS**, do the following.

Select that particular rule from the search results - the row will be highlighted yellow.

Now click the **Disable Rule** button. This will add the **Secondary Name rule TAMI =>THOMAS** rule to the Override Edit-List Rules pane at the bottom of the screen:

The rule is marked as disabled by the presence of the "-" in the **Action** column.

Note: This specific example demonstrates why it is necessary to have formal training in the use of the Population Override Manager. Because most Secondary Name rules are symmetric, to disable the **TAMI =>THOMAS** rule logically also requires the disabling of THOMAS)TAMI.

Adding a New Edit-List Override

To add a new Edit-List rule, first check if there is an existing rule for the word or phrase you are interested in by doing a Search.

If a rule exists for a word and you want to change its behavior, disable the rule first before adding a new rule. To add a new rule, click on the **Add Override** button. This will show the **Add Rule** screen.

You will need to select what type of rule you wish to add from the radio-button list. The available rule-types are:

- Word Rule
- Prefix-Suffix Rule
- Phrase Rule
- Character Rule
- Account Name Marker
- Compound Name Marker
- Delete Marker
- Major Marker

For a full description of these Rule Types and what rules are available in each category, refer to the section below on Rule Types.

After selecting the appropriate Rule Type, you will be prompted to enter the actual word, phrase or character string, and depending on the rule type, a replacement word, phrase or character string. Some rules also require selecting an additional option.

The example below shows a new **NN - Replacement** rule being added for **WLLM** to **WILLIAM**.

Notice that it is also possible to add a comment. This will show up when searching for the rule in the future.

After entering the new rule, click **OK** and the rule will be added to the Override Edit-List Rules pane at the bottom of the main screen.

Deleting an Override

New override rules will remain on the Override Edit-List Rules pane until they are successfully committed, when they are merged with the Edit-List and removed from this pane. While they are on this pane, they can be removed by positioning on the rule to remove and click the **Delete Override** button.

After override rules have been committed, they can only be removed by following the instructions in the section Disabling a Rule above.

All of the current uncommitted overrides can be removed by using the **Discard Changes** button at the top of the screen.

Saving, Testing and Committing Overrides

Before committing overrides to the Edit-List, they must first be tested. Before testing the rules, they must first be saved.

So, the sequence is as follows:

- Add new overrides
- Save Overrides
- Test
- Commit Changes
- Click **Save Overrides** button, the following window will be displayed allowing the entry of some general comments about the rules added, removed or changed in this session.

- Testing the overrides runs them through the Edit-List syntax checker. This checks for such things as duplicate rules and logical inconsistencies. If an error is found, it is displayed with a message similar to:

- In this case, we have added a duplicate rule for WLM)WILLIAM.
- Fix the error and re-test.
- Once the overrides have been tested successfully, they can be committed. This is the time to make sure the rules are what is required, as once they are committed, it may be time consuming to remove them.
- When ready, click **Commit Changes** button

CHAPTER 6

Rule Definitions

This chapter includes the following topics:

- Rule Definitions Overview, 40
- Edit-List Processing Overview, 40
- Edit-List Categories, 41
- Defining a New Category, 44
- Rule Types, 45
- Edit-List Processing Phases, 53

Rule Definitions Overview

This chapter describes the different rules that can be defined and overridden in an Edit-List.

An Edit-List consists of multiple sections:

- Category definitions
- Word Rules
- Prefix-Suffix Rules
- Phrase Rules
- Character Rules
- Account Name Markers
- Compound Name Markers
- Delete Markers
- Major Markers

Edit-List Processing Overview

Before we look at the various Edit-List Rule types, there are some fundamental Edit-List processing concepts that should be understood.

Prior to being processed against the Edit-List, names and addresses first go through a character "cleaning" phase. In this phase, special characters, delimiters and double spaces are removed. In many cases, accents

are also removed. For non-Latin character sets, other rules may apply - check with Informatica Corporation technical support if you have questions about non-Latin character cleaning.

This means, for most Rule Types, do not enter words that contain special characters or delimiters. If it is necessary to define a rule about a word that contains special characters, use the character rule type.

A word may be defined with multiple rule types, provided the words are processed in different Edit- List Processing phases (see the end of this chapter for a description of Edit-List Processing Phases).

Edit-List lookup processing uses exact matching. In other words, for an Edit-List rule to be triggered, the word, number, code or phrase in the name must match exactly the word, number, code or phrase in the Edit-List.

When defining replacement rules, the recommended approach is to replace the shorter word with its longer variation. This is because there is generally more chance of error and variation in the longer word, and as the Edit-List lookup processing uses exact matching, it makes sense to use the word that has the lesser chance of error in the lookup process.

Always, when defining new Edit-List rules, put the effort into making the rules complete. For example, if adding an abbreviation rule for the new word, **UNIVERSITY**, don't limit your thinking to the most obvious (for example, **UNI**). Ask your end-users, or analyze your data, to see what other variations may be in common user (for example, **VARSITY**, **UNIV**, **UNVSTY**).

If you have discovered a rule which you think should be in the Standard Population, please advise Informatica Corporation who will evaluate its inclusion in the next release.

Edit-List Categories

A Category is the means for classifying and categorizing Edit rules. Each Edit rule is defined as being a member of a category and the category is defined to have a specific behavior.

A Category has a two-character name, a rule type and a description or comment associated with it. A Category is referred to and displayed using a combination of its name and comment.

Edit-list Categories may differ in some populations and between Edit-Lists in the same Population, however the majority of Populations have a common set of Categories. The Category associated with a particular rule is displayed when searching for rules in the Population Override Manager and also when adding a new rule.

Following are the common Categories associated with each Edit-List section. This is not an exhaustive list, as some Edit-Lists may have more or less Categories, and new Categories can be added via the Population Override Manager.

Common Categories for Word Rules

Following are the common categories for word rules.

Name Edit-Lists

Category	Rule Type	Examples
NW - NoiseWord	Word is deleted	THE, AND
CT - Company Word Delete	Word is deleted	INC, LTD, CO

Category	Rule Type	Examples
CS - Company Word Skip	Word is marked Skip	DEPARTMENT, ASSOCIATION
SS - Word Skip	Word is marked Skip	MR, MRS, DR, JR
PT - Personal Title Delete	Word is deleted	MR, MRS, DR, JR
NK - Nickname Replace Diminutives	Word and its diminutives are replaced	CATH(E,IE,Y) => CATHERINE
NN - Nickname Replace	Word is replaced	MIKE => MICHAEL
RR - Word Replace	Word is replaced	SVCS => SERVICES
SN - Secondary Lookup	Word generates additional search ranges	AL => ALBERT, ALFRED

Street Edit-Lists

Category	Rule Type	Examples
NW - NoiseWord	Word is deleted	THE, AND
RR - Word Replace	Word is replaced	AVE => AVENUE
DI - Directional Skip	Word is marked Skip	NORTH, SOUTH
SX - Street Word Left Delete	Word is deleted, Word to left is marked Major	RD, AVE
SY - Street Word Left Skip	Word is marked Skip, Word to left is marked Major	HIGHWAY, BOULEVARD
SG - StreetWord Right Delete	Word is deleted, Word to right is marked Major	RUE, JALAN
SH - StreetWord Right Skip	Word is marked Skip, Word to right is marked Major	CORNER
PB - Post Box Word Skip	Word is marked Skip	BOX, POB

Common Categories for Prefix-Suffix Rules

Following are the common categories for prefix and suffix rules across all Edit-List types.

Category	Rule Type	Examples
PP - Prefix Join	Word is joined to following word	MAC
PS - Prefix	String is split from front of word	ELECTRO
PR - Prefix Replace	String at front of word is replaced	MC => MAC
PD - Prefix Delete	String is deleted from front of word	
EE - Suffix Join	Word is joined to preceding word	VILLE

Category	Rule Type	Examples
ES - Suffix	String is split from end of word	
ER - Suffix Replace	String at end of word is replaced	BOROUGH => BURGH
ED - Suffix Delete	String is deleted from end of word	JNR

Phrase Replacement Rules

Phrase rules exist in only one category, called Phrase Rules. The rule type is one of 'replace'. Thus, the only action taken on Phrase Rules is to replace the phrase.

Examples of the use of Phrase Rules include:

Usage	Example
Acronyms	IDS => IDENTITY SYSTEMS
Concatenation	WORLD WIDE => WORLDWIDE
Abbreviation	SVC STN => SERVICE STATION
Noise	RECORD DELETED => [null]

Character Replacement Rules

Character replacement rules exist only in the character rules category. Use the character rules to replace or remove strings or words that include special characters or delimiters.
Note: Character rules are case-sensitive.

The following table describes examples for the use of character rules:

Usage	Example
Replacement	& => AND
Noise	C/O => [null]

Account Name Markers

Account Names are names that include two or more individuals or entities in an abbreviated form. An example of an account name is **JOHN AND MARY SMITH**.

The Account Name Marker is the word that joins the individual names; in the above example, AND. Account Name Markers exist in a category of their own.

The recognition of Account Names by SSA-NAME3 allows it to build more selective keys and perform more accurate matching.

Compound Name Markers

Compound Names are names that include two or more individuals or entities in their full form. An example of a compound name is: **IDENTITY SYSTEMS DBA Informatica Corporation**

The Compound Name Marker is the word that joins the individual names; in the above example **DBA (DOING BUSINESS AS)**. Compound Name Markers exist in a category of their own.

The recognition of Compound Names by SSA-NAME3 allows it to build more selective keys and perform more accurate matching.

Delete Markers

Delete markers are used to delete everything before, after or between the marker definition. They are a more aggressive form of "noise" removal than the Edit Word rules.

Major Markers

A Major Marker is a string of one or two characters that helps identify the position of the "major" word in a name or address.

The major word itself is used when building certain search strategies, usually the Narrow search, and sometimes the Typical search, such that more emphasis is placed on the major word in the search.

The Major word is also used in certain Matching Purposes, especially the Household Purpose, to put more emphasis on the family name.

There is a default method of choosing the major word in a name or address based on an internal setting which specifies whether the major word is most likely at the Left or Right end of the name.

This default setting can be overridden at run time by using the NAMEFORMAT Control.

Other ways of overriding this default behavior is by using Edit word rules, discussed more below, and by using this Major Marker category, in the case where special characters have been used to identify words in the name.

Defining a New Category

A New Category can be defined for Word rules and Prefix-Suffix rules only.

This may be useful in the some situations such as:

- To classify override rules added locally (i.e. via the Population Override Manager), separately from the Standard Population rules.
- If the user has identified a true new category of rules to add.
- If, after discussions with an Informatica Corporation technician, it has been decided that special weighting of certain new categories of words in Matching is required (also requires a Custom Population).

When adding a new category, the user will be prompted to enter a Category Name, Comment and Rule Type. The Rule Type is used in the Edit-List to determine what action to take when that particular category of word is found in the data.

See the following section on Rule Types for a more detailed description of the various rule types that can be added forWord and Prefix-Suffix rules.

Rule Types

This section describes in detail the various Rule Types.

Word Rule Types

A word used in a Word Rule is a character string without embedded spaces, special characters or other delimiters. It can be a word, initial, number or code. The following Rule Types are supported for words.

Delete

Some words are considered to have no value when appearing in a name and some actually impede proper identification. It is preferable to remove these words from the name altogether. Those words will vary from place to place and sometimes even between applications (In person's names Titles and suffixes like **Jnr** or **1st** are a good example; in company names words like **Ltd, Corporation** etc.). For this reason it is necessary to remove those words from the name and use the remaining parts only.

Words of type Delete will be deleted from the name and play no part in the building of a key, key range or in matching.

Replace

It is common that some words appear in different forms such that theWord Stabilization process cannot possibly match them, typical examples would be 1st and First. As this routine will replace all forms with one preferred word for key building purposes, it will be possible to correctly retrieve the desired records.

Nickname

This is similar to the Replace Rule Type, but is specific to certain types of person nicknames. This rule will replace the word, and any variations of the word that have one of the following diminutive endings (EE, EY, IE, EI, IA, AI, E, I, Y, A, O, IEE or any double letter) with a new word.

When processing names for key building, range arrays or matching, if one of the above endings is found on a word, it is temporarily stripped from the word, the preceding consonant de-duplicated, and the result is used to look for a rule in the Edit-List.

For example, consider the nickname BILLY. There are many possible spellings for this word – BILLIE, BILLI, BILL, BILLEE etc. The Nickname rule type creates a 'stem' form by removing the diminutive ending and de-duping the consonant on the end. In this case, the result will be BIL. Thus, having a Nickname rule in the Edit-List for the word BIL with a replacement word of WILLIAM will ensure all the variations are replaced with WILLIAM.

Nickname rules can apply to all words in a name, therefore bear in mind when defining such rules the possible ambiguities and their prevalence in your data. For example, if the following Nickname rule is defined:
NK ROB>ROBERT<

This will work well for ROBBIE and ROBBY, but if ROBE is a surname in your population then it will also be changed to ROBERT.

Skip

There are a couple of reasons why one may want to define a word as a Skip word, however both require assistance from Informatica Corporation and the building of a Custom Population.

Preventing a Word from being used in the Major Position of a Key

For key building, it is possible to specify that no keys or key ranges are built which would have a Skip word in the Major position. This is an internal Algorithm option that must be set by an Informatica Corporation technician in a Custom Population for the customer, as it is not turned on by default.

Setting this option can result in significantly less keys, depending on the number of Skip words defined in the Edit-List and their frequency in the data. Use of this option may reduce disk space and improve performance, however, may also result in some loss in reliability. It is normally only a consideration for the high-volume user.

Reducing the Weight of Words in the Matching process

The weight of Skip words can be lowered in the Matching process. This only applies when two Skip words match. For example, when matching:

ABC SYSTEMS

DEF SYSTEMS

In the default Name Matching Purposes, all words would be weighted the same. However, with an internal Algorithm set on, and SYSTEMS defined as a Skip word, it is possible to assign a lower weight to that word pair and thus reduce the score.

This may be of interest in some special circumstances when overmatching is occurring due to too many skip type words matching.

Mark

Occasionally, an application may have a need to identify if a word is present in a name or address; for example, for the purpose of name classification (is this a company or a person name?).

It is not a standard feature of SSA-NAME3 to provide such classification "out-of-the-box", as it is not required functionality for effective key-building, range-building or matching.

However, if a user has such a requirement, then by using the Mark Edit-List rule type in conjunction with the ssan3_info API call, such results can be achieved.

This requires defining all of the words that would be used to classify the names to the Edit-List in a Category that uses the Mark rule type. Then, after a call is made to ssan3_get_keys, an ssan3_info call can be made using the ITEM=results.categories parameter to return the word categories found in the name. The category returned will be a two-character, internal form of the name.

Major Left Delete

A word categorized as Major Left Delete marks the word to its left as a Major word, and is then deleted. If the word is the first word in the name then the default rule for choosing the Major word is used (i.e. based on an internal setting which specifies whether the major word is most likely at the Left or Right end of the name.)

The major word itself is used when building certain search strategies, usually the Narrow search, and sometimes the Typical search, such that more emphasis is placed on the major word in the search.

The Major word is also used in certain Matching Purposes, for example in the Household Purpose to put more emphasis on the family name.

Major Left Keep

Same as the Major Left Delete rule type, except that the word is not deleted but rather marked as a Skip word.

Major Right Delete

A word categorized as Major Right Delete marks the word to its right as a Major word, and is then deleted. If the word is the last word in the name then the default rule for choosing the Major word is used (i.e. based on an internal setting which specifies whether the major word is most likely at the Left or Right end of the name.)

Major Right Keep

Same as the Major Right Delete rule type, except that the word is not deleted but rather marked as a Skip word.

Secondary Name

The Secondary Name rule type is a special type of Replace rule. Words defined as Secondary Names receive special treatment in the generation of search ranges and also in matching. However, Secondary Name definitions do not cause extra keys to be stored in the database.

Secondary Names are used for a number of purposes:

- Improving selectivity for searches containing "ambiguous" words. For example, defining the following Secondary Name rules:

  BERT >HERBERT

  BERT >GILBERT

  BERT >NORBERT

  BERT >BERTRAM

  BERT >ALBERT

  HERBERT >BERT

  GILBERT >BERT

  NORBERT >BERT

  BERTRAM >BERT

  ALBERT >BERT

  Does not cause the replacement values to happen in the keys, but in the search ranges. In this example, it means that a search for BERT will also generate search ranges that look for HERBERT, GILBERT, NORBERT, BERTRAM and ALBERT, and a search for any of the latter will also search for BERT. However, a search for ALBERT will not, for example, return a search range containing NORBERT. Thus, selectivity is improved.

  It is important that such Secondary Name rules are defined symmetrically, or the results will not be reliable. So, if a rule is added for TINA => CHRISTINA, the reverse rule CHRISTINA => TINA should also be added.

- Better handling of synonyms.
  It is appropriate to define certain types of synonyms as Secondary Name rules instead of Replace rules. For example, take the synonyms VEHICLE and AUTOMOBILE. If a direct Replace rule is used to replace one with the other, for example VEHICLE) AUTOMOBILE, then the word VEHICLE is lost forever and AUTOMOBILE is used in both keys and search ranges. Thus, a misspelling of VEHICLE will not find names that have the correct spelling, or other spelling variation of VEHICLE in them (as it will not fire the VEHICLE) AUTOMOBILE Replace rule).

  By defining the VEHICLE) AUTOMOBILE replacement rule as a Secondary Name rule, a search that contains the word VEHICLE will find names the contain that word and it's phonetic variations, as well as names that contain AUTOMOBILE and its phonetic variations. Note, it is still important to add important abbreviations as Direct replacement rules (e.g. AUTO)AUTOMOBILE). Remember to define these types of Secondary Name rules as symmetric also.

- Adding temporary replacement rules that do not require the SSA-NAME3 Key index to be re-built. Because Secondary Name rules only affect key ranges for searching, and not the keys that are stored in the database, it is possible to define new Secondary Name rules on the fly, and get immediate results without rebuilding keys. Therefore, even if a Direct Replacement rule would be best, if immediate benefit is required, then defining it as a Secondary Name rule will suffice. If and when some regular housekeeping is done to rebuild the SSA-NAME3 Keys index, these "temporary" Secondary Name replacement rules can be changed to Direct replacement rules.

Secondary Name rules are also used in Matching.

- In the above example, BERT will score well with HERBERT, GILBERT, NORBERT, BERTRAM and ALBERT; however ALBERT and NORBERT will not score as highly.
- Another use for Secondary Names in Matching is for defining geographical proximity in Addresses. For example, say the zip codes 02077 and 02120 are geographically next to each other and likely to have a certain amount of overlap in real addresses (for example, in the use of "vanity" addresses). Defining the following Secondary Name rule: 02077>02120, means that addresses with either zip code will score well against each other.

Do Not Stabilize

Words defined in a Category that have the Do Not Stabilize rule type will not undergo word stabilization. Word stabilization is a part of key and range building that addresses phonetic and orthographic error.

In most situations, it is advisable for words to be stabilized because this overcomes spelling and typing error in the word. In some cases, however, a number of very common words can stabilize to the same form and create a selectivity problem in large data volumes for searches involving those words. For example, the words JOHN, JAMES, JOAN, JAN and JANE all stabilize to the same word. Thus, when searching for JOHN SMITH, JOAN SMITH, JANE SMITH, JAN SMITH and JAMES SMITH will all be returned.

Provided the risk of missing a match has been properly examined, it may be decided to separate these in the keys by defining them as Do Not Stabilize.

In cases where the risk of missing a match is high, use of this rule is not recommended.

Prefix-Suffix Rule Types

Prefix Join

In names, it may be common that certain prefixes sometimes appear joined to a word, and sometimes separated. For example, MAC DONALD or EL DORADO Vs MCDONALD or ELDORADO). When such common variations occur, the accuracy of matching may be improved by defining Prefix Concatenate rules. In the examples above, the prefix words MAC and EL would be used in the rules.

A prefix word, when found, will be concatenated to the word following it in the name unless that word is a single character or a code.

The resultant word will then be used in a re-lookup of the Edit-List to check for other rules that may apply.

It is important to think carefully when defining such rules. For example, in some populations VAN may be a good candidate for a Prefix Concatenate rule, as it is commonly a prefix. However, in other populations, VAN is a common family name, and this processing would not be appropriate.

Prefix Split

A prefix defined with a Prefix Split rule type, when found at the beginning of a word, will cause that prefix to be split from the word. Both parts are then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the split prefix word has no other Edit-List rules, it is made a Skip type word.

If the Prefix Split word is found alone in a name (i.e. not part of a word), it is simply left as it is and marked as a Skip type word.

Prefix Replace

A prefix defined with a Prefix Replace rule type, when found at the beginning of a word, will cause that prefix to be replaced with the replacement value defined in this rule. The combined word is then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the Prefix Replace word is found on its own, it is still replaced and re-processed through the Edit-List.

An example of a Prefix Replace rule might be MC) MAC.

Prefix Delete

A prefix defined with a Prefix Delete rule type, when found at the beginning of a word, will cause that prefix to be deleted. The resulting word is then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the Prefix Delete word is found on its own it is also deleted.

Suffix Join

In names, it may be common that certain suffixes sometimes appear joined to a word, and sometimes separated. For example, MATRA VILLE Vs MATRAVILLE). When such common variations occur, the accuracy of matching may be improved by defining Suffix Concatenate rules. In the examples above, the suffix word VILLE would be used in the rule.

A suffix word, when found, will be concatenated to the word preceding it in the name unless that word is a single character or a code.

The resultant word will then be used in a re-lookup of the Edit-List to check for other rules that may apply.

Suffix Split

A suffix defined with a Suffix Split rule type, when found at the beginning of a word, will cause that suffix to be split from the word. Both parts are then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the split suffix word has no other Edit-list rules, it is made a Skip type word.

If the Suffix Split word is found alone in a name (i.e. not part of a word), it is simply left as it is and marked as a Skip type word.

Suffix Replace

A suffix defined with a Suffix Replace rule type, when found at the end of a word, will cause that suffix to be replaced with the replacement value defined in this rule. The combined word is then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the Suffix Replace word is found on its own, it is still replaced and re-processed through the Edit-List.

An example of a Suffix Replace rule might be `BOROUGH =>BURG`.

Suffix Delete

A suffix defined with a Suffix Delete rule type, when found at the end of a word, will cause that suffix to be deleted. The resulting word is then used in a re-lookup of the Edit-List to check for other rules that may apply.

If the Suffix Delete word is found on its own it is also deleted.

The Phrase Replacement Rule Type

The phrase replacement process is similar to that performed by the Word Replace logic except that phrase replacement rules can detect and replace multiple words, initials, numbers or codes.

A phrase, and its replacement, each have a maximum length of 50 characters, and each word in a phrase or its replacement has a maximum length of 24 characters.

Both the phrase and its replacement can be one word; however, if both the phrase and its replacement are one word then theWord Replace rule would typically be used.

Phrase Replacement Logic

If a particular Phrase replacement rule is a sub-set of another Phrase replacement rule, and the last word in both rules is the same, the longest phrase definition which matches is used.

For example, if the Edit-List has the following Phrase definitions:

Phrase: SOFTWARE AMERICA

Replacement: XXX

Phrase: SEARCH SOFTWARE AMERICA

Replacement: YYY

then the name :

RESEARCH DIVISION SEARCH SOFTWARE AMERICA

would be produce the following result:

RESEARCH DIVISION YYY

because *SEARCH SOFTWARE AMERICA* is the longer match of the two entries.

Notice that no *XXX* replacement will ever take place in his case, because once the replacement is made and the name becomes *RESEARCH DIVISION YYY*, then *SOFTWARE AMERICA* no longer exists.

After the replacement, the resulting word or phrase is used in a re-lookup of the Edit-List to see if any other rules apply.

There are a number of reasons for defining Phrase Replacement rules.

Acronyms

It is recommended that in most cases the acronym is replaced with its long form. The reason is that there is a greater chance of error in the long form,

For example,

Phrase:IBM

Replacement: INTERNATIONAL BUSINESS MACHINES

Remember, when defining new Edit-List rules, try to make the rules complete to maximize the benefit. In the above example, it would also be important to define a Phrase Replacement rule for:

Phrase:I B M

Replacement: INTERNATIONAL BUSINESS MACHINES

Phrase:INTL BUSINESS MACHINES

Replacement: INTERNATIONAL BUSINESS MACHINES

Note: It is not necessary to define a rule for I.B.M., as the full-stops are removed prior to Edit-List processing.

This example also provides an opportunity to discuss when one might want to use a single word in a phrase and replacement rule. For example, if the word INTERNATIONAL was needed in a number of phrase replacement rules, one can define it once as a single word, and it will be processed before the other rules. Thus defining,

Phrase:INTL

Replacement: INTERNATIONAL

Would alleviate the need for the rule:

Phrase:INTL BUSINESS MACHINES

Replacement: INTERNATIONAL BUSINESS MACHINES

Concatenation

For example,

Phrase:NOT KNOWN

Replacement:UNKNOWN

Phrase:WORLD WIDE

Replacement:WORLDWIDE

Abbreviation

For example,

Phrase:SVC STN

Replacement:SERVICE STATION

Phrase:PTY LTD

Replacement:PROPRIETARY LIMITED

Noise

For example,

Phrase:RECORD DELETED

Replacement:

Phrase:SEE FILE

Replacement:

The Character Rule Type

Normal Edit-List processing occurs on a name after it has been passed through a character cleaning phase; however, sometimes it is necessary to apply Edit rules before this cleaning has taken place - this is the use of the character rules.

For example, with no character rule, the name,

C/O

enters normal Edit-List lookup processing as:

C O

For the purpose of this example, C/O is assumed to be a common abbreviation for CARE OF. However, in a name C O could easily be a person's initials or company acronym. To avoid the confusion, a character rule can be defined for C/O, as such:

Substitute: C/O

with Word: CARE OF

Because the character rule is invoked prior to the character cleaning phase, which also performs case stabilization, such rules must be defined as case sensitive. Thus, if *C/O* is expected to be encountered also as lower case *c/o*, or mixed case *C/o*, then extra rules need to be defined to cater for these.

Flags

The following flags may be optionally defined for a character rule.

Delimited After: A delimiter (space, special character, start or end of name) must be found immediately following the string for the rule to be triggered.

Delimited Before: A delimiter (space, special character, start or end of name) must be found immediately prior to the string for the rule to be triggered.

Delimited After and Before: A delimiter (space, special character, start or end of name) must be found both immediately before and after the string for the rule to be triggered.

For example adding the following rule to the Edit-list definition file,

Substitute:T/A

with Word:TRADING AS

Flag:None

will mean the following word will trigger the rule:

DESCENT/ASCENT

and be changed to:

DESCENTRADING ASSCENT

If this is not desired, then add the appropriate delimiter flag. In this case, the rule will now look like:

Substitute:T/A

with Word:TRADING AS

Flag:Delimited Before and After

The Account Name Marker Rule Type

For SSA-NAME3 to recognize Account Names it needs both Account Name Patterns defined to the Algorithm, as well as Account Name Marker words defined to the Edit-List.

The Standard Populations have these defined for the most common patterns and markers.

Because the Population Override Manager does not give the user access to the definition of the Account Name Patterns, if you feel that the embedded patterns do not adequately cover your data, please discuss this with an Informatica Corporation technician who may suggest a Custom Population.

All that needs to be defined for an Account Name Marker is the actual marker value.

The Compound Name Marker Rule Type

Unlike Account Names, Compound Names do not require special patterns defined to the Population. Therefore, new Compound Name Markers can be added without need for a Custom Population.

All that needs to be defined for a Compound Name Marker is the actual marker Value.

The Delete Marker Rule Type

The definition of a Delete Marker rule requires the entry of a Value and a Flag. The Value is the marker.

The Flag is one of the following:

Delete After: Everything after the marker (including the marker) is deleted. For example entering *REF* as the marker value will cause *REF* and everything after *REF* to be deleted in the following name.

JOHN SMITH REF FILE NO 23456

Delete Before Everything before the marker (including the marker) is deleted. For example entering NAME as the marker value will cause NAME and everything before NAME to be deleted in the following name.

PERSONS NAME: JOHN SMITH

Note: The colon will be deleted by normal character cleaning.

Delete Between In this case, a pair of delimiters is entered as the marker value. Everything between the delimiters is deleted.

For example entering ()(brackets) as the marker value will cause everything inside the brackets, including the brackets, to be deleted.

JOHN SMITH (OLD RECORD)

The Major Marker Rule Type

The definition of a Major Marker rule in the Population Override Manager requires the entry of a Value and a Flag. The Value is the one- or two- character string that is the marker.

The Flag is one of the following:

Delimited: The major word can be found between the delimiters entered. For example, entering {} as the Value defines braces as the delimiter markers. A word thus found enclosed in braces will be flagged as the major word in the name. If more than one word is found enclosed in the delimiters, the default setting discussed above, Left or Right, will be used to choose the major word.

Head: The major word can be found prior to this character or characters in the name. For example, entering a comma (,) as the marker means the entire head of the name (all characters up to the comma) is flagged as being the major word in the name. If more than one word is found prior to the marker, the default setting discussed above, Left or Right, will be used to choose the major word.

Left: The major word can be found immediately to the left of the marker. For example, entering a star (*) as the marker means the word to the left of the star will be chosen as the major word. If there are no words left of the marker, the default setting discussed above, Left or Right, will be used to choose the major word.

Right: The major word can be found immediately to the right of the marker. For example, entering a backslash (\) as the marker means the word to the right of the backslash will be chosen as the major word. If there are no words right of the marker, the default setting discussed above, Left or Right, will be used to choose the major word.

Tail: The major word can be found after this character or characters in the name. For example, entering a percent (%) as the marker means the entire tail of the name (all characters after the percent) is flagged as being the major word in the name. If more than one word is found after the marker, the default setting discussed above, Left or Right, will be used to choose the major word.

Edit-List Processing Phases

During Key Building, Range Building and Matching, the components of a name are checked against entries in the Edit-List. If an exact string match is found, the associated rule is triggered.

A word may belong to multiple Edit-List sections, as long as the words are processed in different Edit- List Processing phases.

The sequence of Edit-List processing is as follows.

Phase	Description
1	Compound & Account Name Rules
2	Major & Delete Marker Rules
3	Character Rules
4	Phrase Replacement Rules
5	Word & Prefix-Suffix Rules

Edit Word Processing Phase (5)	Rule Type
5.1	Mark
5.2	Prefix join
	Delete
	Major right, delete
	Major right, keep
	Word not Stabilized
	Postfix join
	Replace
	Skip
	Major left, delete
	Major left, keep
5.3	Prefix delete
	Prefix split
	Prefix replace
5.4	Postfix split
	Postfix delete
	Postfix replace
5.5	Nicknames with diminutives

CHAPTER 7

Implementing a Local Population

This chapter includes the following topics:

- Order of Precedence, 55
- What to Do After Committing a Local Population, 55
- Managing the Edit Rule Wizard, 56
- Size of a Local Population, 56
- Upgrading a Local Population using the latest Standard Population, 56

Order of Precedence

The action of committing overrides to a Standard or Custom Population will create a *Local Population*. Local Populations have a file extension of *.ylp*.

When any of the SSA-NAME3 Clients (the Workbench, Population Override Manager or the Edit Rule Wizard), or an Application program invokes the SSA-NAME3 callable routine, passing it a System and Population name, the callable routine will look in the folder identified by the system name and load the population with that population name.

There is an order of precedence in the loading of this Population.

If a YLP (Local Population) is present, it will be loaded. If no YLP is present, but a YCP (Custom Population) is present, it will be loaded. Otherwise the YSP (Standard Population) with that name will be loaded.

What to Do After Committing a Local Population

Once committed, a Local Population *(.ylp)* is ready to use. It simply needs to be copied to the location where the *SSAPR* environment variable points to the system or application in question.

Remember that many types of rules that can be added to the Edit-List require the SSA-NAME3 Keys to be rebuilt in the database before the changes are effected. So, depending on the rules added, it may be necessary to have these keys re-built prior to releasing the YLP for use by a search application.

Any type of rule that either deletes, replaces, splits or concatenates words, numbers, codes or strings, will require the SSA-NAME3 Keys to be rebuilt.

The only exception is the Secondary Name Rule Type, which is a replace rule that only affects search ranges, and thus does not require the SSA-NAME3 Keys to be rebuilt.

Managing the Edit Rule Wizard

If the Edit RuleWizard is in use in your organization, it will also be capable of creating a Local Population.

The Population Override Manager and Edit Rule Wizard clients have been designed to use the SSANAME3 server so that Population maintenance conflicts can be avoided. Essentially, when either of the Clients loads a Population for work, that Population is "locked" for single-user access.

When an Edit Rule Wizard user commits a set of new Edit rule additions, these will also create a Local Population (.ylp). However, the difference is that none of the rules that can be added by the Wizard require the SSA-NAME3 Keys to be rebuilt.

Thus, a Local Population committed by the Edit Rule Wizard is ready for use as soon as it has been committed.

Note: Standard or Extended Keys should be implemented if the Edit Rule Wizard is being used.

Size of a Local Population

The first time a Local Population is generated from a Standard Population, its size will change to be approximately three times the size of the original Standard Population. Thereafter, its size will stabilize (unless adding a Scalar Frequency Table which can add between 1MB and 6MB to the size of the Local Population depending on the settings chosen).

The reason for the increase in size is that it:

- Keeps a copy of the original Edit List rules
- Creates an Edit List that is a merge of the original Edit List rules and the Local Rules
- Generates a version of the Edit List that requires no translation at run time.

Upgrading a Local Population using the latest Standard Population

Over time, Informatica Corporation releases updated versions of the Standard Populations.

To take advantage of the rules in a later Standard Population while at the same time retaining the local rule overrides added via the Population Override Manager or Edit Rule Wizard, perform an upgrade on the Local Population using the ssaprupg command line utility.

This utility can be run in test mode or in commit mode.

In test mode, the utility will test the upgrade of the Local Population without committing it. This is useful to see if any rule conflicts exist that need to be corrected. For example, a certain noise word may have been added via the Edit Rule Wizard and therefore exist in the Local Population. However, Informatica Corporation may have added that same noise word to the latest version of the Standard Population. During the upgrade test process, this will result in a rule conflict that must be manually fixed before continuing. (To fix the conflict in this example, start the Population Override Manager and disable the conflicting noise word rule in the Local Population).

The utility should only be run in commit mode once all rule conflicts have been addressed.

Using the SSAPRUPG Utility

ssaprupg usage is as follows:

```
ssaprupg -sSystem -pPopulation [-c] [-1LogFile] [-2ErrFile]
```

Where:

-s is the name of the SSAPR subdirectory

-p is the name of the population (without the file extension)

-c will cause the upgraded Local Population to be committed

-1 specifies the name of a file that will contain log output

-2 specifies the name of a file that will contain error output

To use the ssaprupg utility:

1. If it is not already there, copy your Local Population file to the default `SSAPR` subdirectory (e.g. `%SSATOP%\pr\default`) where the latest Standard Population resides.
2. Open a Command Prompt window.
3. Change to the `%SSATOP%\bin` directory and set up the SSA-NAME3 environment variables by running `env\nm3s.bat`
4. Change to the default SSAPR directory (e.g. `%SSATOP%\pr\default`). This is not mandatory as the location will be found using the combination of the SSAPR environment variable and the -s parameter
5. Run the ssaprupg utility in test mode. For example: `%SSABIN%\ssaprupg -sdefault -pusa -1log.txt -2errs.txt` will perform a test upgrade of the `usa.ylp` population using the `usa.ysp` population in the same directory.
6. If there are any rule conflicts, the errors will be written to the file specified by the -2 parameter. For example: `e0: dup NN+NN: ANI` reports that there is a duplicate NN (nickname) rule for the word `ANI`.
7. Fix any rule conflicts.
8. Once the utility runs without producing any errors, the upgraded Local Population can be committed by running, for example: `%SSABIN%\ssaprupg -sdefault -pusa -c -1log.txt` This action when completed successfully will make a copy of the existing Local Population as `usa.ylp-old` and overwrite the Local Population with the upgraded version.

INDEX

A

abbreviation 40, 45, 51
Account Name Marker 43
Account Name Marker Rule Type 52
Acronyms 49

C

Categories 41
Character Replacement Rules 43
Character Rule 51
Character Rule Type 51
Compound Name Marker Rule Type 52
Compound Name Markers 44
Concatenation 49
Custom 11, 18
Custom Population 11, 18

D

Defining a New Category 44
Delete 45
Delete Marker rule 52
Delete Marker Rule Type 52
Delete Markers 44
Do Not Stabilize 45

E

Efields 12

F

Frequency tables 13

L

Local 11
Local Population 11

M

Major Left Words 45
Major Marker 44
Major Marker Rule Type 53
Major Right 45
Managing 56
Marked Words 45

N

Nickname 45
Noise Word 49

O

Order of Precedence 55
Overrides 32

P

Phrase Replacement 49
Phrase Replacement Rules 43
Population Override Manager 11
Prefix Delete 48
Prefix Join 48
Prefix Replace 48
Prefix Split 48
Prefix-Suffix 42
Prefix-Suffix Rule Types 48
Processing Overview 40

R

Replace 45
Rule Types 45
rules 12

S

Scalar Frequency Tables 13
Secondary Names 45
sections 40
Skip word 45
ssaprupg 57
Standard 11, 18
Standard Population 11, 18
Suffix Delete 48
Suffix Join 48
Suffix Replace 48
Suffix Split 48

U

Upgrading 56
Usage Warning 12

W

Word Rule Types 45
Word Rules 41