THE POWER TO KNOW.

SAS® Enterprise Content Categorization 12.1

User's Guide

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2012. *SAS® Enterprise Content Categorization 12.1: User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Enterprise Content Categorization 12.1: User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, August 2012

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

Contents

About This Book

Audience

SAS Enterprise Content Categorization Studio enables multiple users to work together collaboratively using SAS Content Categorization Studio. This product includes support for LITI concepts and for collaborative operations. LITI concepts perform run-time disambiguation of matches based on their context. The following types of users can perform these operations:

- Taxonomists develop the categories and concepts that comprise the taxonomy for your enterprise.
- Subject matter experts write the category rules and concept definitions for this taxonomy. These rules also include LITI concept rules.
- Testers test the rules and analyze the rule matching results in input testing documents.

You could be assigned one of these functions, or all of them.

SAS Enterprise Content Categorization Studio enables you to use this software with other SAS products. This manual focuses on tasks that define and configure the collaborative operations and LITI rule-building for SAS Enterprise Content Categorization Studio.

Prerequisites

Here are the prerequisites for using SAS Enterprise Content Categorization Studio:

- SAS Content Categorization Studio loaded onto your machine
- Appropriate server permission for all users, assigned by the database and project administrators
- Read *SAS Content Categorization Studio: User's Guide* and *SAS Enterprise Content Categorization Studio: Administrator's Guide*

Conventions

This manual uses the following typographical conventions:

Convention	Description
`TGM_ROOT`	The root directory where SAS Content Categorization Studio is installed, typically the following: **Windows:** `C:/Program Files/SAS/SAS Content Categorization Studio` **UNIX:** `/opt/SAS_collab_server`
`Top`	The names of taxonomy nodes appear in a fixed-width font.
`www.sas.com`	The hypertext links are shown in a light blue, fixed-width font, and are underlined.
OK button	The labels for user interface controls are shown in a bold, sans-serif font.
	The Question Mark button accesses *SAS Enterprise Content Categorization Studio: User's Guide* in PDF format.

What's New in SAS Enterprise Content Categorization Studio 12.1

New and enhanced features for SAS Enterprise Content Categorization Studio enable you to do the following:

- Use predefined LITI concepts to shorten the rule-writing process.
- Use the UNLESS and NOT operators to limit rule matches.
- Specify XPath expressions in LITI rules in order to locate matching content in XML elements.
- Scroll through LITI rule types using keyboard operators.
- See the canonical forms and information fields for LITI rule matches.

Use Predefined LITI Concepts

Reference predefined LITI concepts in your rules in order to shorten the rule-writing process. These concepts are available for Arabic, Chinese, Danish, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, Swedish, and Thai languages. However, the concepts that are available for each language might vary.

Note: Some of these concepts are available for download at `http://support.sas.com/demosdownloads/setupintro.jsp`. Select the Text Analytics link. Follow the instructions on this Web page and within this document to download and use this file.

Specify the UNLESS and NOT Operators

Use the `UNLESS` and `NOT` operators to prevent matches under specific circumstances. For example, restrict a match on another Boolean operator

using the UNLESS operator. Use the NOT operator to prevent a match when a match that is specified by the AND operator also occurs.

Specify XPath Expressions in LITI Rules

Write a rule using XPath expressions for greater flexibility in choosing where to locate matching text. Specify the specific XML field, or fields, to limit matches to this text.

Scroll through LITI Rule Types

Use keyboard shortcuts to access the different LITI rule types when writing your LITI concept rules.

See the Canonical Forms and Information Fields

Click on a highlighted word in the Document pane to see a window that displays the canonical or information string for the matched term. The canonical form refers to pronoun resolution and the information field is for CLASSIFIER concepts.

Documentation Changes for the 12.1 Release

See the following table to understand the documentation for the 12.1 release:.

Documentation	12.1 Product	Tasks and 5.2 Product References
SAS Content Categorization Studio: Installation Guide	SAS Content Categorization Studio	Install the single user or the enterprise version of SAS Content Categorization Studio that you purchased. The enterprise version automatically installs support for collaborative features and LITI concepts.
	SAS Enterprise Content Categorization Studio	

Documentation	12.1 Product	Tasks and 5.2 Product References
SAS Content Categorization Studio: User's Guide	SAS Content Categorization Studio	Create a SAS Content Categorization Studio project, test, and upload the project to SAS Content Categorization Server. This guide is written for a single user and is a companion book for *SAS Enterprise Content Categorization Studio: User's Guide*.
SAS Enterprise Content Categorization Studio: Administrator's Guide	SAS Enterprise Content Categorization Studio with collaborative operations.	Configure your server for collaborative operations. (In the 5.2 release, this book was *SAS Content Categorization Collaborative Server: Administrator's Guide*.)
SAS Enterprise Content Categorization Studio: User's Guide	SAS Enterprise Content Categorization Studio with collaborative operations and LITI concepts capabilities.	See the cell above and use this guide to understand how collaborative operations work. Use the second part of this guide to write LITI rules and to upload these rules to SAS Content Categorization Server. (In the 5.2 release, LITI rules were explained in *SAS Contextual Extraction Studio: User's Guide*.)

Documentation	12.1 Product	Tasks and 5.2 Product References
SAS Enterprise Content Categorization Servers: Administrator's Guide	Download any, or all, of the following: - SAS Content Categorization Server - SAS Enterprise Content Categorization Studio - SAS Content Categorization Java API - SAS Content Categorization Python API - SAS Document Conversion Server and Java API	Install, configure, and use SAS Content Categorization Server, SAS Enterprise Content Categorization Studio, and SAS Document Conversion Server. You can also upload `.li` files using this product. In the 5.2 release, the information in this book was found in the following manuals: - *SAS Content Categorization Server: Administrator's Guide* - *SAS Content Categorization Collaborative Server: Administrator's Guide* - *SAS Document Conversion: Developer's Guide*
SAS Content Categorization Single User Servers: Administrator's Guide	Download any, or all, of the following: - SAS Content Categorization Server - SAS Content Categorization Java API - SAS Content Categorization Python API - SAS Document Conversion Server and Java API	Install, configure, and use SAS Content Categorization Server and SAS Document Conversion Server In the 5.2 release, the information in this book was found in the following manuals: - *SAS Content Categorization Server: Administrator's Guide* - *SAS Document Conversion: Developer's Guide*.

Chapter: 1

About SAS Enterprise Content Categorization Studio

- *What Is SAS Enterprise Content Categorization Studio?*
- *Benefits of Using SAS Enterprise Content Categorization Studio*
- *How Do the Collaborative Features Work in SAS Content Categorization Collaborative Server?*
- *Using Collaborative Operations*
- *Architecture*

1.1 What Is SAS Enterprise Content Categorization Studio?

In most organizations it is necessary to obtain information about, and from, data that is created internally and externally. This process is expedited when a team of subject matter experts work together to create a single SAS Content Categorization Studio project. The collaborative operations and the ability to define LITI concepts are what distinguish SAS Content Categorization Collaborative Server from SAS Content Categorization Studio.

Using the collaborative server and a Windows interface, multiple users can access a single project residing on a server. You, as an administrator specify the various levels of access for all project users, upload projects to the server, and perform other administrative operations.

Easy permission setting

The database administrator adds users to the project and sets their permission levels. Easy-to-use interfaces within the project enable the project administrator to set permissions for these users both across the taxonomy and at the individual category and concept level.

Easy distributed workflow

Processes are distributed across various areas of expertise. For example, taxonomists might develop the categories and concepts while subject matter experts write rules.

Levels of permissions

Users are assigned permissions to work on a project according to their level of expertise. For example, some users have permissions to read and write rules. All users added to the database have Read-Only permissions by default.

1.2 Benefits of Using SAS Enterprise Content Categorization Studio

SAS Enterprise Content Categorization Studio provides users with the following add-on benefits to SAS Content Categorization Studio:

Use the expertise of a wide range of subject matter experts

SAS Enterprise Content Categorization Studio adds the features that enable multiple subject matter experts to work together on one SAS Content Categorization Studio project. These benefits include permission setting, and the shared projects folders that enable two or more developers to work on a single project on one machine.

Control access to the project

The database and project administrators set permissions at the database, project, and node levels to restrict access to various components.

Automatic project updates

Use the Options window to automate the updates for your project.

Enable two or more users to build and edit separate projects on one machine

Two, or more developers, create projects on one machine using the cached version in the *Shared Projects* folders.

The following benefits are for SAS Content Categorization Studio. They are included with SAS Enterprise Content Categorization Studio:

Empower subject matter experts and taxonomists by providing a simple, visual interface where you build a taxonomy, define rules, and test

SAS Content Categorization Studio includes easy-to-use Windows interfaces that make it easy to build large, complex, and hierarchical taxonomies. Specify your own rules, test, and generate `.mco` and `.concepts` files that are applied by SAS Content Categorization Server to input documents.

Develop metadata for your information

SAS Content Categorization Studio uses advanced linguistic technologies to identify metadata in, and about, your documents.

Improve the business value of information technology and the corporate data that it manages

SAS Content Categorization Studio creates `.mco` and `.concepts` files that automate the classification and extraction of entities from input documents during real time using SAS Content Categorization Server.

Save money on information retrieval and organization costs

All of the information created by, or within, your organization can be classified and retrieved. You can find related information, whether you know the exact terms that you are seeking.

1.3 How Do the Collaborative Features Work in SAS Content Categorization Collaborative Server?

Use SAS Enterprise Content Categorization Studio to enable multiple subject matter experts to develop one project, while controlling access according to

expertise. These experts use SAS Enterprise Content Categorization Studio as explained in the following paragraphs:

SAS Content Categorization Studio is a Windows application that anyone can use to develop taxonomies that classify and extract the information found in your organization. Interactively identify the data that you need without using a programming language.

SAS Content Categorization Studio enables users to easily create taxonomies, write rules, and test these rules against a variety of testing sets. You can upload the output `.mco` and `.concepts` and `.liti` files to SAS Content Categorization Server where they are automatically applied to input documents.

1.4 Using Collaborative Operations

The ability of multiple developers to work together depends on collaborative operations:

- Create a collaborative, customized SAS Content Categorization Studio project using categories and concepts that form a taxonomy residing on a server. This project can be accessed by multiple developers using a single machine, or by users who are each working on their own local machines.
- Set project access levels for each of the individual developers and teams if you are an administrator. These permissions, or access levels, can be set at the project, and at the category and concept levels. For example, choose to restrict development permissions for specific taxonomy nodes to a single individual or group. You can simultaneously grant this person or group wider access to other categories and concepts.
- Track modifications to the project using a revision log for both a category rule and a concept definition. You can also use the information in this log file to revert to an older version of the rule or definition.
- Simplify the task of keeping a local project up-to-date with its counterpart on the server.

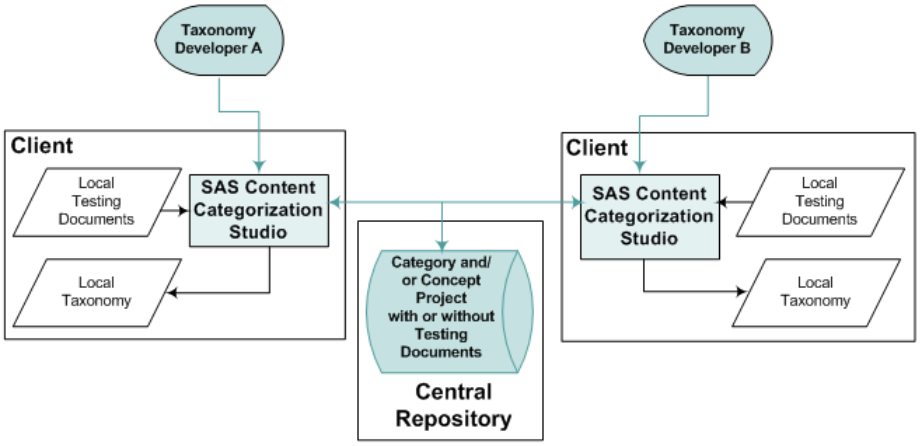
- Make changes to the taxonomy when a project is up-to-date with the server. This feature ensures that the modifications, additions, and deletions that are made by one developer are not overwritten by another.
- Automate commit operations for all of the changes to the taxonomy.
- Automate a number of server and cached project update processes and set up separate *Shared Projects* folders using the Options window. This process enables multiple developers to work on individual cached projects that are located on one machine.
- (Optional) Upload testing documents to the server to enable other developers to use these documents to reproduce your results.

In summation, these features provide the benefits of collaboration. These operations enable your team to balance necessary control with optimal flexibility in order to meet your organization's project development requirements.

1.5 Architecture

Two, or more, subject matter experts, or developers, can work together to create one SAS Content Categorization Studio project.

Figure 1-1 SAS Content Categorization Collaborative Server Architecture



A SAS Content Categorization Studio project that is used for collaborative work, resides on a remote server. The server enables multiple subject matter experts to work together on the same project. This project is cached on their local machines. A project that resides on a server is called a remote project.

Part 1: Collaborative Operations

- Chapter 2: *Setting Up Your Data Source*
- Chapter 3: *Using the Interface*
- Chapter 4: *Getting Started with Collaboration*
- Chapter 5: *Other Collaborative Operations*

2
Setting Up Your Data Source

- *Overview of Setting Up a Project on Your Server*
- *Configuring ODBC Data Sources*
- *Logging In*

2.1 Overview of Setting Up a Project on Your Server

After the database administrator sets up data sources on the server, you can create an ODBC data source on your local machine. This chapter is for those users who set up their own data sources. This data source points to the SAS Content Categorization Collaborative Server database. After you complete this process, you can access the SAS Content Categorization Studio project on the server.

The term *administrator* is used to refer to both the database and the project administrator. Database administrators are system administrators and project administrators have upload permissions and other permissions that you as a regular user do not.

For more information, see *SAS Enterprise Content Categorization Studio: User's Guide*. See the chapter used by administrators to set up data sources, if necessary.

2.2 Configuring ODBC Data Sources

2.2.1 Credentials

There are several credentials that supply the information required to create your data source:

Name

User

Password

Database: You can either specify a name or use the default entry.

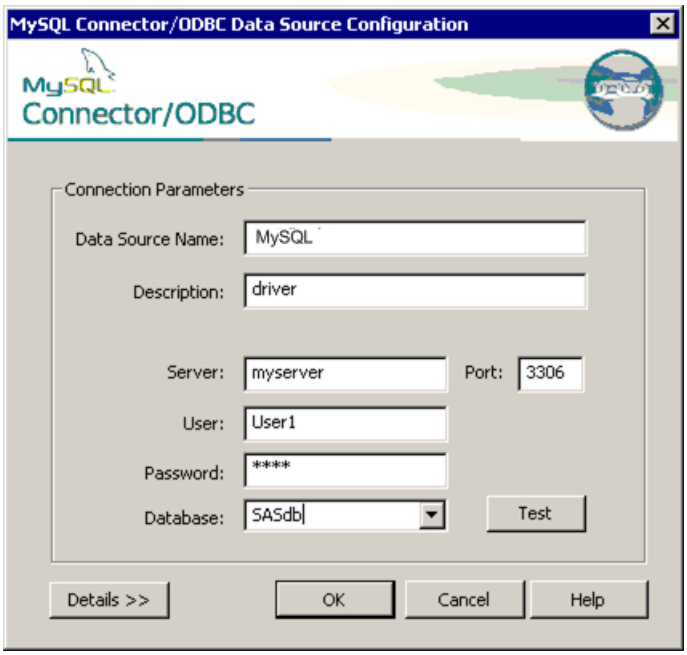
Select one of the following sections, depending on the type of ODBC data source that you are creating:

- Section 2.2.2 *Configure a MySQL Data Source* on page 25
- Section 2.2.3 *Configure a Microsoft SQL Server Data Source* on page 27
- Section 2.2.4 *Configure an Oracle Data Source* on page 33

2.2.2 Configure a MySQL Data Source

If you are working with a MySQL ODBC driver, configure this data source by completing these steps:

1. Open the MySQL Connector/ODBC page. Use this page to point the ODBC driver to the server:



2. Enter the name of the data source, selected by the administrator, into the **Data Source Name** field. For example, type `SASdb`.
3. (Optional) Enter the description into the **Description** field. For example, type `driver`.
4. Enter the server name into the **Server** field. For example, type `myserver`.
5. (Optional) The default entry, shown in the right pane of this interface is entered into the **Port** field.
6. Enter the user name, entered by the administrator to the database, into the **User** field. For example, type `User1`.

Note: You cannot use a period (.) or an at sign (@) when you create a user name for MySQL Server. For example, an e-mail address is not a valid user name.

7. Enter your password into the **Password** field. The password for the administrator is the administrator's password on the server.
8. Enter the name of the database that you are connecting to into the **Database** field. For example, type `tk240db`.
9. Click **Test**. If the connection is successful, the Connector/ODBC window appears with a message stating that the connection is successful.

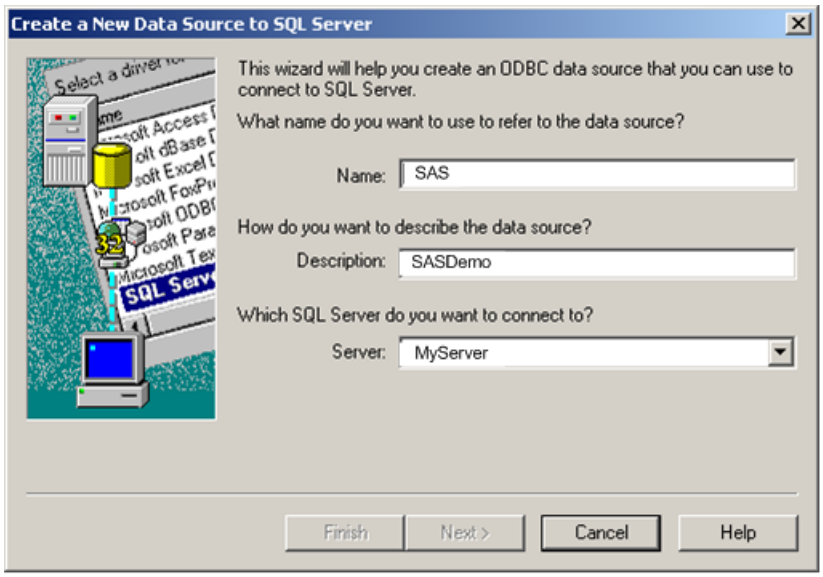
10. (Optional) Click **Details** in the MySQL Connector/ODBC window to set various types of flags.
11. Click **OK** to close this window.
8. Click **OK** in the MySQL Connector/ODBC window.

2.2.3 Configure a Microsoft SQL Server Data Source

If you are working with Microsoft SQL Server, configure this data source.

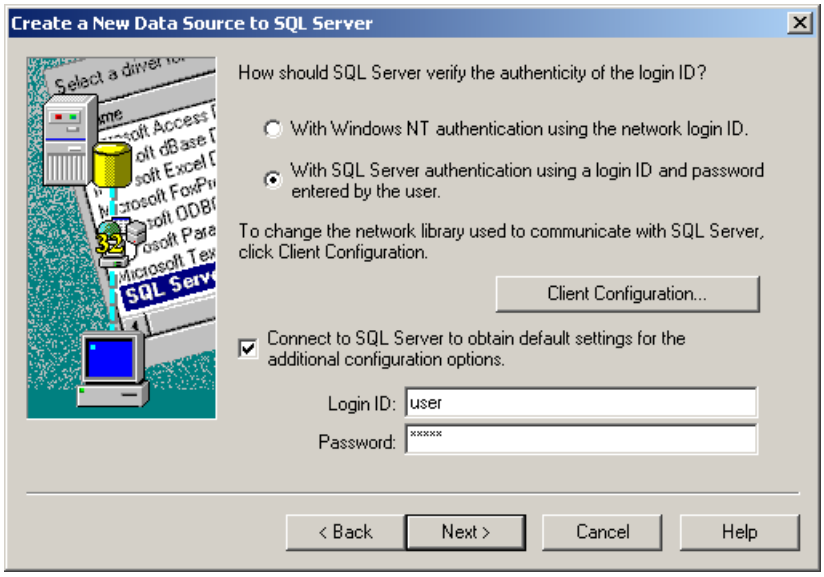
To configure your Microsoft SQL Server ODBC data sources, complete these steps:

1. Open the Create a New Data Source to SQL Server wizard.



2. Enter the name of your data source in the **Name** field. For example, type `SAS`.
3. (Optional) Enter the descriptive information for your database into the **Description** field. For example, type `SASDemo`.
4. Enter the name of the server into the **Server** field. For example, type `MyServer`.

5. Click **Next** and the next Create a New Data Source to SQL Server page appears.

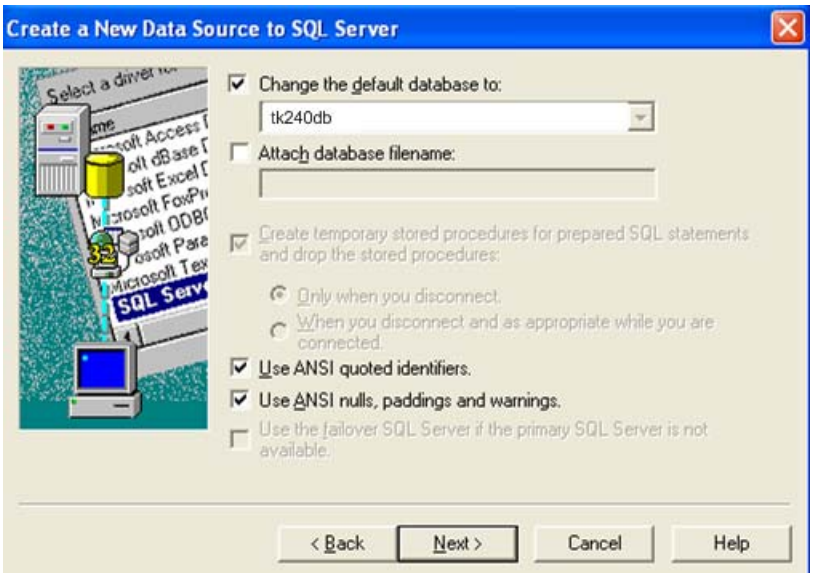


6. (Either Windows NT authentication or SQL Server authentication is OK.) Select **With SQL Server authentication using a login ID and password entered by the user.**

> **Note**: Step 7 and Step 8 apply only if SQL Server authentication is selected.

7. (Optional, but recommended) Enter your name into the **Login ID** field.
8. (Optional, but recommended) Enter your password into the **Password** field.

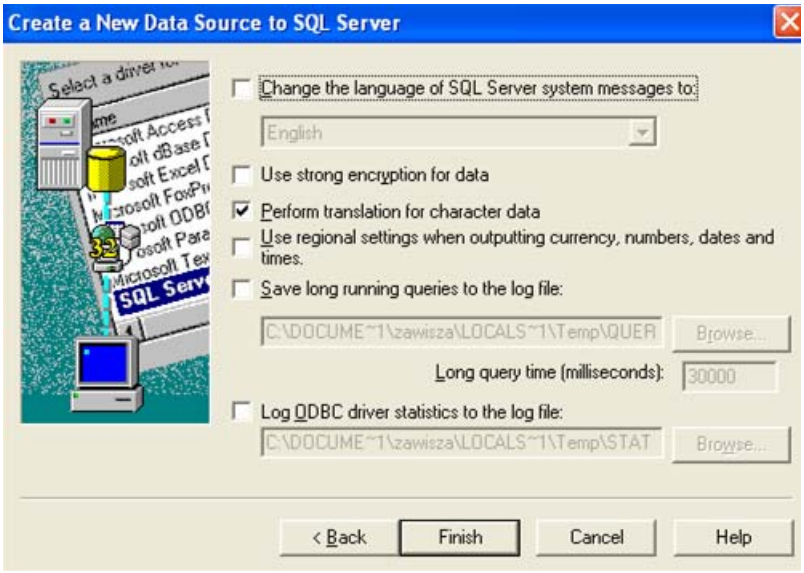
9. Click **Next** and the next Create a New Data Source to SQL Server page appears.



Any of the default settings are OK.

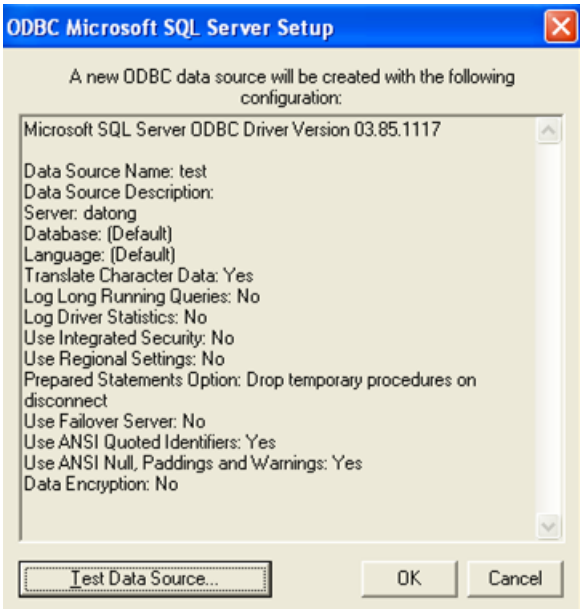
10. Click **Next** to go to the following page.

11. Click **Next** and the next **Create a New Data Source to SQL Server** page appears.

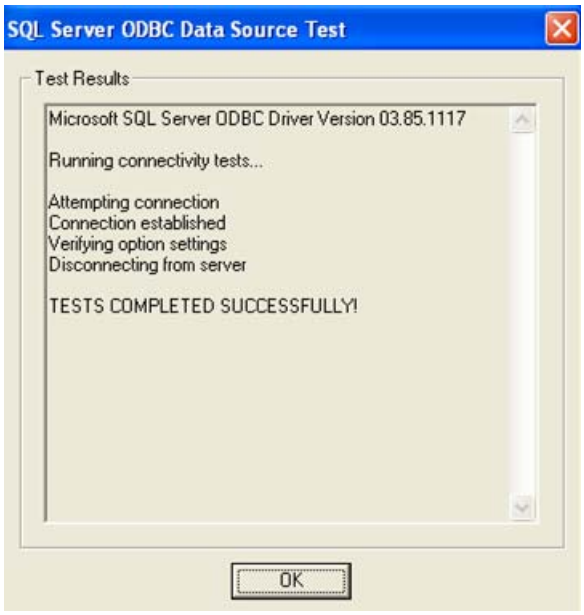


Any of the default settings are OK.

12. Click **Finish**. The ODBC Microsoft SQL Server Setup page appears where you can see the details of your ODBC data source configuration.



13. (Optional) Click **Test Data Source** and a window similar to the SQL Server ODBC Data Source Test window appears.



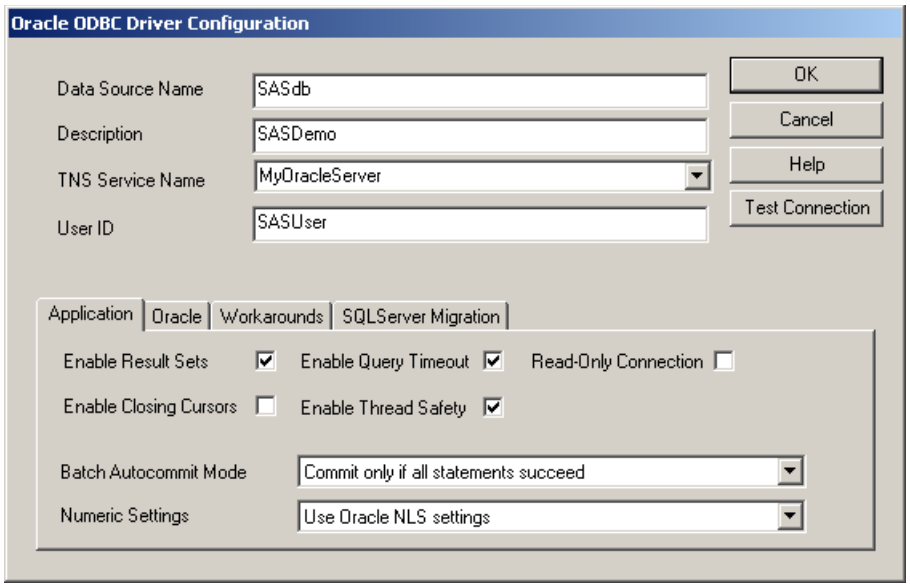
14. Click **OK**, if you see the SQL Server ODBC Data Source Test window.
15. Click **OK** in the ODBC Microsoft SQL Server Setup page.

2.2.4 Configure an Oracle Data Source

If you are working with an Oracle ODBC Driver, configure this data source.

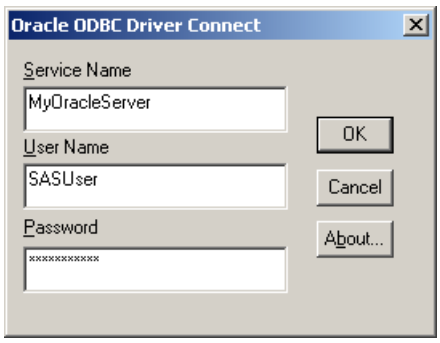
To configure your Oracle ODBC data sources, complete these steps:

1. The Oracle ODBC Driver Configuration window appears.



2. Enter the name of the data source into the **Data Source Name** field. For example, type `SASdb`.
3. (Optional) Enter the information about your database into the **Description** field. For example, type `SASDemo`.
4. Enter the server name into the **TNS Service Name** field. For example, type `MyOracleServer`.
5. Enter the user name assigned by the database administrator into the **User ID** field. For example, type `SASUser`.

6. Click **OK**. The Oracle ODBC Driver Connect window appears.



7. The fields in this window are automatically filled in for you with the exception of the **Password** field.

Hints: The **Server Name** field contains the same name as the entry in the **TNS Service Name** field.
The **User Name** field is specified by the database administrator.

---

8. Click **OK**. If the connection is successful, the Testing Connection window appears with this message.

9. Click **OK**.
10. Click **OK** to close the Oracle ODBC Driver Connect window.

2.3 Logging In

After the connections to the remote server are set up from the local host, a user can log in from any host that they are set up to use. Three pieces of information are required to log in to the remote repository:

- the data source name
- your user name
- your password

3
Using the Interface

- *How Collaborative Management Affects the User Interface*
- *The Menu Bar*
- *The Standard Toolbar*
- *The Options Window*
- *The Server Operation Windows*
- *The Miscellaneous Windows*

3.1 How Collaborative Management Affects the User Interface

Collaborative operations are enabled in order to enable multiple users to work on a single project on a server. For this reason, some of the operations that are visible in the SAS Content Categorization Studio user interface are not accessible until you set up the server. The collaborative operations enable multiple users with different permission levels to work in different capacities on the same project.

This chapter explains the interface components that regular users access. Regular users are granted various levels of permissions by the administrator. These permission levels affect your access to the project and the operations that you perform. This chapter does not describe any of the components that are limited to administrative users. The highest level of access for regular users is assumed, unless otherwise stated in this chapter.

3.2 The Menu Bar

Many of the collaborative operations for SAS Content Categorization Studio are located in the menu bar. Some of the menu commands are also available on the standard toolbar, and a few can be accessed when you right-click taxonomy nodes.

Display 3-1: Menu Bar

The table below describes the collaborative-only commands that are available in the SAS Content Categorization Studio menus for regular users. For all other operations, see *SAS Content Categorization Studio: User's Guide*.

Table 3-1: Collaborative Operations

Menu	Operation	Description
File	Open Remote Project	Access the existing projects on the server using the Select a Project window that appears.
	Upload Project to Server	Send a copy of the project on your local machine to the server.
	Remove Project From Server	Select a project to delete in the Select a Project window that appears. For more information, see Section 3.6.3 *The Select a Project Window* on page 48.
	Repository Login	Log in to the server where the collaborative projects reside using the Repository Login window that appears. For more information, see Section 3.6.1 *The Repository Login Window* on page 45.
	Change Repository Password	Change the password that you use to access the server in the Change Password window that appears. For more information, see Section 3.6.4 *The Change Password Window* on page 49.
Note: The Repository Login window appears if you are not logged in to the server before selecting one of the operations above.		
Edit	Select **Edit --> Options** and the Options window appears displaying check boxes for some collaborative operations. For more information, see Section 3.4 *The Options Window* on page 41.	

Table 3-1: Collaborative Operations (Continued)

Menu	Operation	Description
Server	Status	Learn the status of the selected node in relation to the server project using the messages that appear in the Taxonomy window. For more information, see Section 4.8.2.B *Checking the Status of the Taxonomy* on page 74.
	Update	Download the current version of the project, if you select either the `Categorizer` or `Concepts` node in the taxonomy. If you select a category or concept node, download the current version of the rule. Use this server operation to obtain changes from a project that is stored on the server. Any uncommitted rule changes are overwritten during this process. For more information, see Section 4.8.3 *Update from the Server* on page 76.
	Commit	Send your local changes to the server using this operation. During this process, SAS Content Categorization Studio places a lock on the selected node to prevent another developer from committing changes during this operation. For more information, see Section 4.8.4 *Commit Changes to the Server* on page 78.
	Revision Log	Open one of the following two RevisionLog windows in *Notepad*. The first window lists only rule or definition changes for the selected node. The second window provides information about deleted and renamed nodes and rule or definition changes for the selected branch in the taxonomy. For more information, see Section 4.8.5 *Using the Revision Logs* on page 80.
	Revert to Older Version	Use the Revert to Previous Version window that enables you to replace the rule or definition with an earlier version. Use the Revision Log window to determine the version number that meets your requirements. For more information, see Section 4.8.5.C *Revert to an Older Version* on page 83.
	Upload Test Files	Load all of the test files to the server for the project that is open on your machine. This operation makes the test files available to all of the permissioned users for this project. For more information, see Section 5.2.3 *Upload Test Files* on page 92.

Table 3-1: Collaborative Operations (Continued)

Menu	Operation	Description
	Download Test Files	Download the test files that are stored on the server to your machine. For more information, see Section 5.2.4 *Download Test Files* on page 93.
	Local Permissions	Access the Server Permissions window for the specific category that is selected. For example, see **Server Permissions for Gardening** if you selected the Gardening category and Local Permissions.
	Project Permissions	Access the Project Server Permissions window.
Help	*SAS Content Categorization Studio: User's Guide* *SAS Enterprise Content Categorization Studio: User's Guide* *SAS Enterprise Content Categorization Studio: Administrator's Guide*	Click to see a PDF version of each of these books. For more information see Appendix B: *Recommended Reading*.
	About	Open the About SAS Content Categorization Studio window that displays version, licensing, and dating information.

3.3 The Standard Toolbar

The standard toolbar is located below the menu bar. Some of the buttons are specific to server operations.

Display 3-2: Standard Toolbar Buttons

See the table below for an explanation of each of these buttons.

Table 3-2: Standard Toolbar Buttons

Icon	Button		Description
	Server Update	Project-wide update	Select either the `Categorizer` or `Concepts` node and click this button. The SAS Content Categorization Studio status window appears asking if you want to overwrite or preserve the local changes to your project. If you do *not* want to preserve your local changes, your entire project is overwritten by the project on the server. If you want to preserve your local changes, updates are downloaded for the components that you have not modified.
		Local update	Highlight a node in the taxonomy and click this button to see a message telling you whether the selected category or concept is up-to-date with its counterpart on the server.
Note: In either case, status messages appear in the taxonomy tree.			
	Server Status		Select a node in the Taxonomy window and click this button. Messages appear to the right of the nodes in the taxonomy tree providing information about local changes and whether the local copy is up-to-date with the project on the server. For more information, see Table 3-4 on page 44.
	Server Commit		Click this button to commit your local changes to the server. The Enter Comment window appears, unless this component is disabled in the Options window or the category or concept is already up-to-date with its counterpart on the server. Use the Enter Comment window that appears, unless you select **Skip comments on commit** in the Options window. Write notes to track the changes to your project.

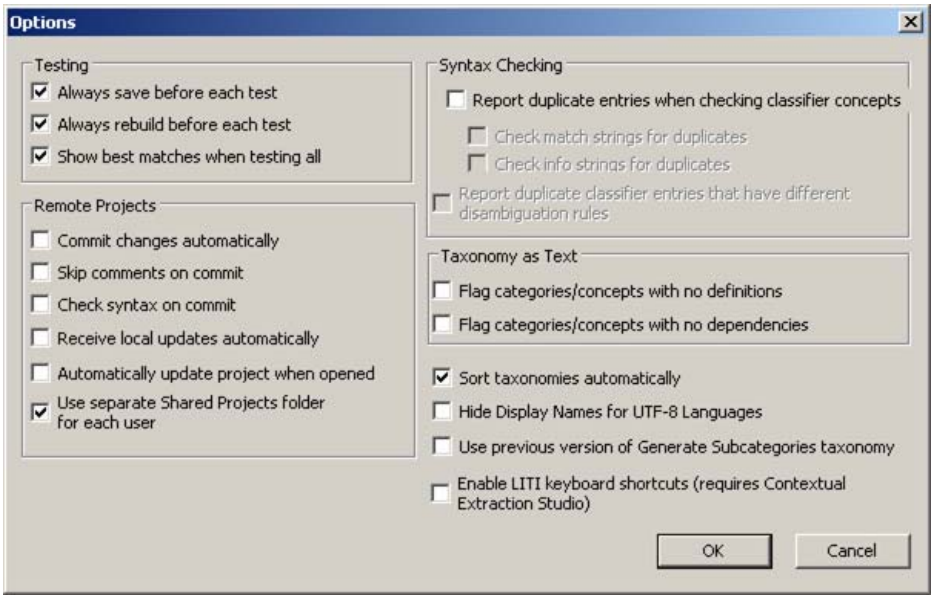
3.4 The Options Window

The Options window enables you to set installation-wide settings for collaborative work. (These specifications apply across all collaborative projects, unless you make different selections.) Use the six **Remote Projects**

selections in this interface to simplify the process of keeping your local project up-to-date with the project on the server.

When you make choices using the Options window, they remain selected as the default operations until you reset these operations. Options are installation-specific. Project Settings are project-specific.

To open the Options window, select **Edit --> Options**.



Select any of the following collaborative operations that are listed under the **Remote Projects** heading:

Table 3-3: Options Window Components

Component	Description
Commit changes automatically	Automatically commit rule and definition changes to the project on the server, after you make a change and select a different taxonomy node.
Skip comments on commit	Commit changes to the server without tracking the reasons for these revisions.

Table 3-3: Options Window Components (Continued)

Component	Description
Check syntax on commit	SAS Content Categorization Studio automatically checks the syntax of the definition or rule before committing it to the server.
Receive local updates automatically	Automatically update the category rules and concept definitions from the server whenever a change is committed by another user.
Automatically update project when opened	Automatically download any changes committed to the server for the taxonomy and its categories and concepts when you open a remote project. This operation provides an alternative to the Server Update operation. For more information, see Table 3-1 on page 38.
Use separate Shared Projects folder for each user	(Default) Enable two or more users to use the same machine to build projects in SAS Content Categorization Studio. These users can maintain separate local (cached) projects. These projects are stored in the Shared Projects folder that is automatically created on your local machine during installation. For this reason, status windows appear if you do not commit your changes before saving them. **Hint**: A SAS Content Categorization Studio window appears when you make this selection asking you to commit your changes.

3.5 The Server Operation Windows

3.5.1 Overview of the Server Operation Windows

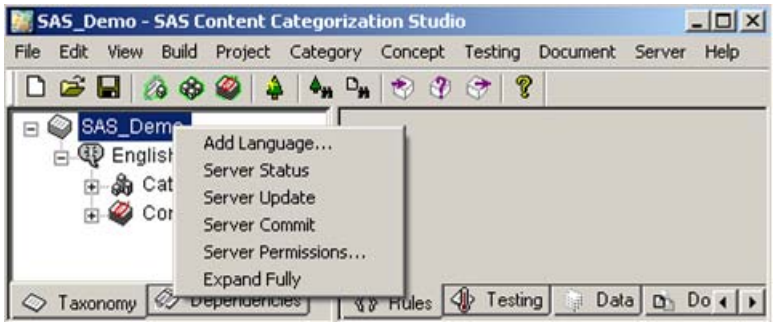
The server operations enable multiple developers to work together, collaboratively, on a project. You can access the server operations for your project through the **Server** menu or by right-clicking on a taxonomy node. For information about the server operations that are available from the project name node, see Table 3-4 on page 44. For information about the server operations that are available from the individual taxonomy nodes, see Section 3.5.2 *Server Operations Available for Individual Taxonomy Nodes* on page 44.

3.5.2 Server Operations Available for Individual Taxonomy Nodes

Some of the server operations described in Table 3-1 on page 38 and Table 3-2 on page 41 are also available when you right-click on a taxonomy node. Server operations are available for all of the nodes in the Taxonomy pane with the exception of the `Top` node.

The type of node that you select in the Taxonomy pane determines the server operations that are available. For example, if you select the project node, the available operations affect the entire taxonomy. If you select an individual taxonomy node instead, the available operations affect only the selected node.

Display 3-3: Server Operations



A drop-down menu appears. The available server operations are specific to the type of node that you select. See the table below for all of the available operations for regular users from the various nodes in the taxonomy.

Table 3-4: Server Operations for Taxonomy Nodes

Server Operation	Project Name	Language	Categorizer or Concepts Extractor	Top	Category or Concept
Server Status	X	X	X		X
Server Update	X	X	X		X
Server Commit	X	X	X		X

Table 3-4: Server Operations for Taxonomy Nodes (Continued)

Server Operation	Project Name	Language	Categorizer or Concepts Extractor	Top	Category or Concept
Revision Log			X		X
Revert to Older Version					X
Import Category from Repository				X	X
Import Concept from Repository				X	X
Note: The server operations, **Upload Test Files** and **Download Test Files** are available only in the **Server** menu.					

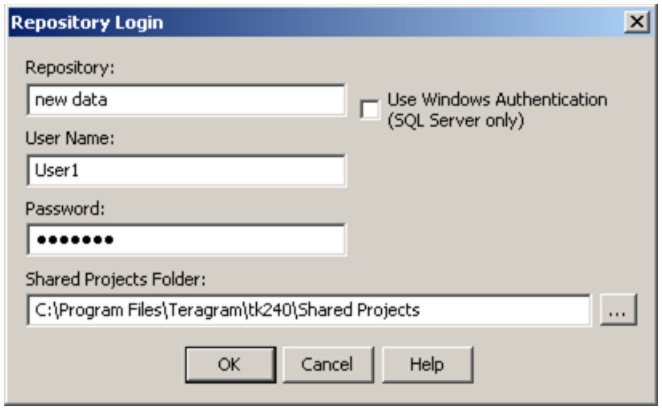
3.6 The Miscellaneous Windows

3.6.1 The Repository Login Window

Log in to the server before you work on a collaborative project. If you do not log in and you select any of the following operations from the **File** drop-down menu, the Repository Login window automatically appears:

- **Open Remote Project**
- **Remove Project from Server**
- **Repository Login**
- **Change Repository Password**

Display 3-4: Repository Login Window



To use the Repository Login window, complete these steps:

1. Enter the name of the ODBC data source into the **Repository** field.
2. (Optional) Select **User Windows Authentication (SQL Server only)** to use the Windows user account principal token to connect. If you select this check box, the **User Name** and **Password** fields are grayed.
3. Enter your user name into the **User Name** field.
4. Enter your password into the **Password** field.

Hints: If you previously used SAS Content Categorization Collaborative Server, the **Repository**, **User Name**, and **Shared Projects Folder** fields are all filled in.
In Windows Vista/2008 Server/7 a preference is set so that data is not stored in the Program Files directory hierarchy.

---

5. (Optional) Click ... under **Shared Projects Folder** and the Select a Directory window appears. Use this window to select a location for the Shared Projects folder. For more information, see Section 3.6.2 *The Select a Directory Window* below.
6. Click **OK** to access the repository.

3.6.2 The Select a Directory Window

Use this section when you want to store the locally cached copy of your remote project in a directory that is not the default location. For example, if you are running Windows XP and access to the Program Files directory is restricted to administrative users, use the this window.

Use the Select a Directory window to change the location of the cached copy of your project.

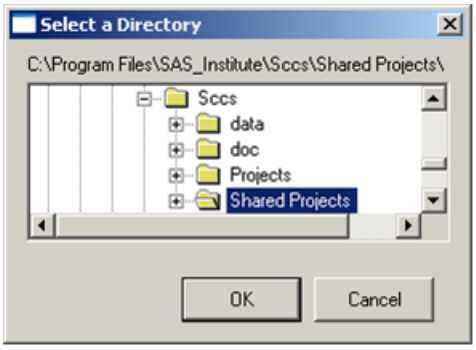
To open and use the Select a Directory window, complete these steps:

1. Select **File --> Open Remote Project**. The **Repository Login** window appears. See Display 3-4 on page 46.
2. Use Step 1 to Step 4 on page 46.

Hint: These fields are automatically filled in for you, unless this is the first time you are using this window.

---

3. (Optional) Click  under the **Shared Projects Folder** heading and the Select a Directory window appears.



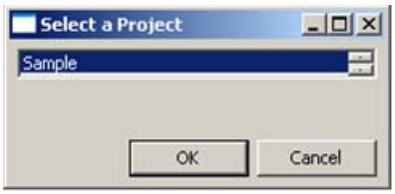
4. Select the `Shared Projects` folder or a cached project in another location on your machine.
5. Click **OK**.

3.6.3 The Select a Project Window

Use the Select a Project window to locate a project on the server. You can open the project on your local machine, or delete the project.

To open and use the Select a Project window to open a project, complete these steps:

1. Select **File --> Open Remote Project**. The Select a Project window appears.



2. Highlight the project in the Select a Project window that you want to open. For example, select `Sample`.
3. Click **OK**.

To remove a project, complete these steps:

1. Select **File --> Remove Project From Server**.
2. Use Step 2 through Step 3 above. A SAS Content Categorization Studio confirmation window appears.

> **Note**: The project that you want to remove from the server cannot be open and running on your machine during the removal process.

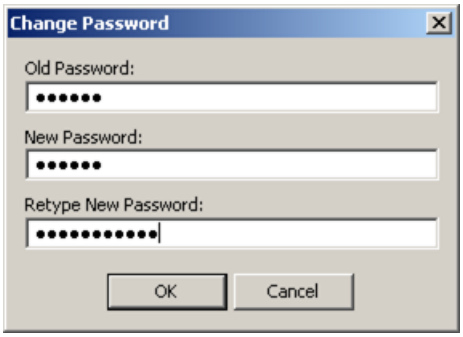
3. Click **Yes**.

3.6.4 The Change Password Window

Use the Change Password window to specify a new password.

To open and use the Change Password window, complete these steps:

1. Select **File --> Change Repository Password** and the Change Password window appears.



2. Enter the password assigned to you by the database administrator into the **Old Password** field.
3. Enter your new password into the **New Password** field.
4. Re-enter the new password into the **Retype New Password** field.
5. Click **OK**.

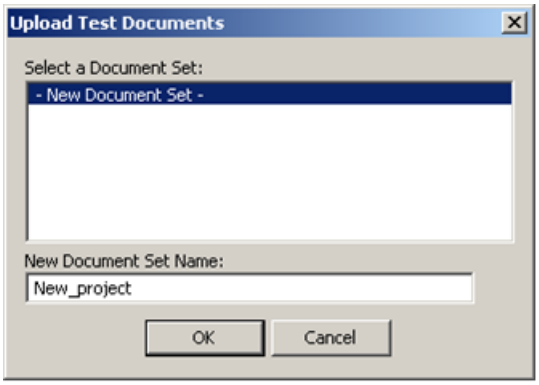
3.6.5 The Upload Test Docs Window

Testing documents that are stored on the server enable other subject matter experts to test their rules and definitions against the same set of texts that you use. Other users can also validate the results that you see.

A user who has, at a minimum, `Read, Write, and Change Taxonomy` permissions can use the Upload Test Documents window. This operation enables the user to transfer a set of testing documents from a local machine to the server. The SAS Content Categorization Studio project is accessible to other permissioned users here. This subject matter expert can also use the Download Test Documents window to access test files uploaded to the server by another user.

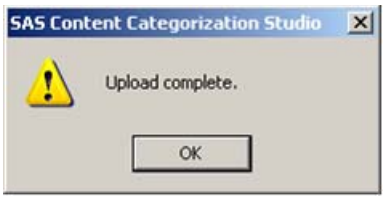
To open and use the Upload Test Documents window, complete these steps:

1. Open the Data window and set the path to the testing documents using the **Testing Path** field, unless it is already entered.
2. Open the Testing window and select **Server --> Upload Test Files**. The Upload Test Documents window appears.



   Any uploaded sets of documents are listed under the **Select a Document Set** heading. If no testing documents have been uploaded to the server, you see `- New Document Set -`.

3. Type the name of the new set of texts that you are uploading into the **New Document Set Name** field.
4. Click **OK** and a SAS Content Categorization Collaborative Server confirmation window appears.



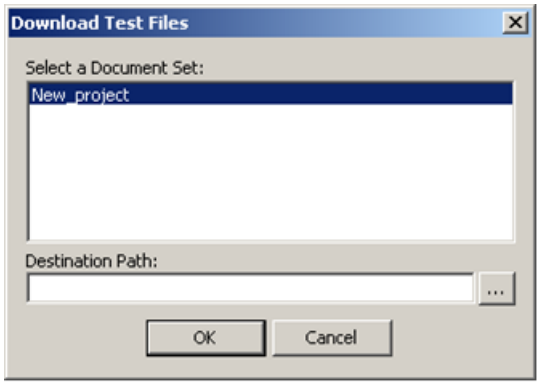
5. Click **OK**.

3.6.6 The Download Test Docs Window

If you have at a minimum, `Read, Write, and Change Taxonomy` permissions, you can download testing documents from the server to your machine. Use these files to test your taxonomy. When you use shared test documents, you can validate the testing results of another user.

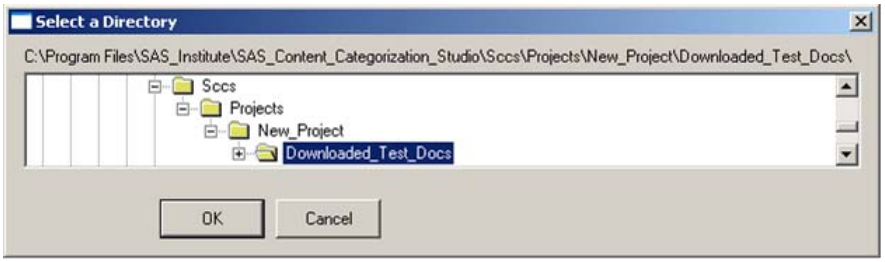
To open and use the Download Test Files window, complete these steps:

1. Open the Testing window and select **Server --> Download Test Files**. The Download Test Files window appears.



2. Select a file under the **Select a Document Set** heading. For example, choose `New_project`.

3. Click  to the right of the **Destination Path** field and the Select a Directory window appears.



4. Select the folder location for the uploaded testing documents.

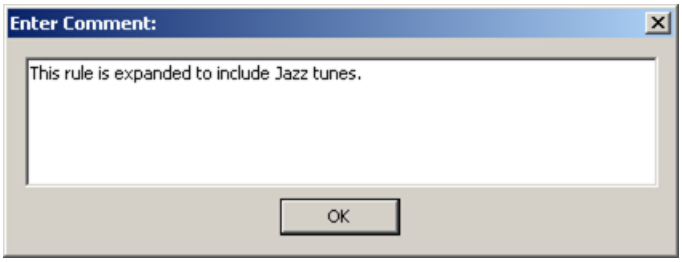
5. Click **OK**. This path appears in the **Destination Path** field.
6. Click **OK** in the Download Test Files window.

3.6.7 The Enter Comment Window

Use the Enter Comment window to write notes about changes to a category or a concept, for either a taxonomy change or a **Server Commit** operation. These notes appear in the RevisionLog windows.

To open and type notes into the Enter Comment window, complete these steps:

1. Make a change to your project at either the taxonomy, or individual node, level.
2. Select **Server --> Commit** and the Enter Comment window appears.



3. Type your comments in the blank field. For example, type T*he rule is expanded to include Jazz tunes*.
4. Click **OK**.

Your notes on the changes appear in the RevisionLog windows. For more information, see Section 3.6.8 *The RevisionLog Screens* on page 52.

3.6.8 The RevisionLog Screens

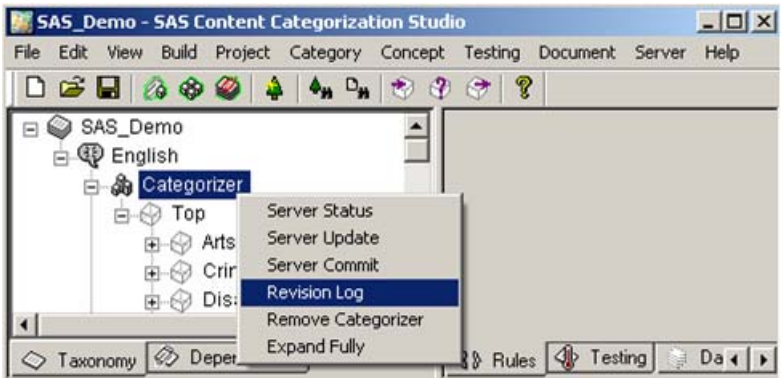
Use both of the RevisionLog screens for reference purposes or to revert to an earlier version of a category or concept. There are two types of RevisionLog screens that make it possible for you to track changes after they are committed to the server:

- Select a taxonomy node and use the revision log operation to see only the rule or definition changes for the selected category and concept.

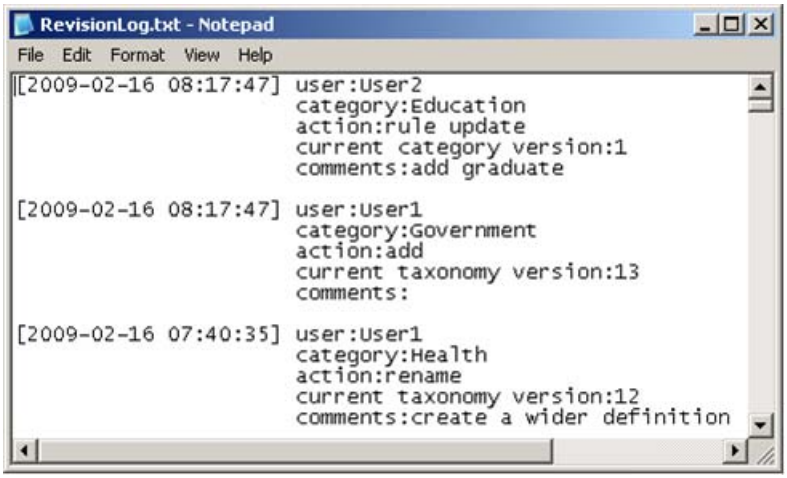
- Select either the `Categorizer` or `Concepts` node. Use the **Revision Log** operation to see the changes to all of the nodes that comprise one branch of the taxonomy. In other words, see the changes for either the categorizer or the concepts branch and only for one language.

In both cases the revision logs are comprehensive. In other words, they track all of the changes, whether for a rule, a definition, or a selected branch of the taxonomy. An example is provided for the `Categorizer` node. Similar results are displayed for the `Concepts` node. If you instead, select an individual category or concept node, only the results for the selected node are displayed.

To open the RevisionLog window for a taxonomy branch, right-click on, the `Categorizer` node in the Taxonomy window, and select the **Revision Log** operation.



The RevisionLog screen appears displaying the changes made to all of the nodes in the selected taxonomy branch.



Use the components of the RevisionLog screen to learn about the changes made to this branch of the taxonomy:

Table 3-5: RevisionLog Window Components

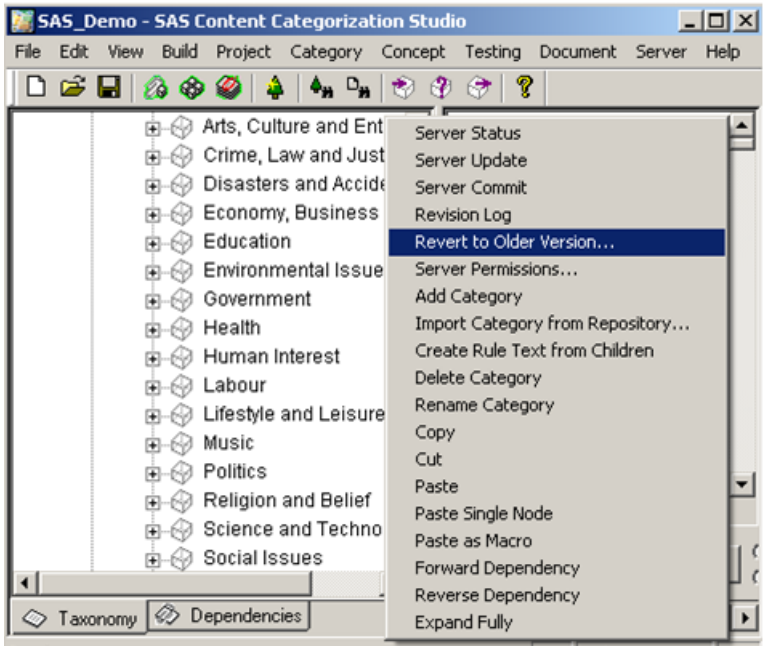
Component	Description
date and time	Date in `YYYY-MM-DD` and `hh-mm-ss` formats.
`user`	Name of the user making the changes appears here.
`category`	Name of the category appears here.
`action`	Type of change that was made appears here.
`current taxonomy version`	Count of the saved changes made for the referenced node. For example, the Education category rule was updated making the `current category version: 1`.
`comments`	Comments entered in the Enter Comment window, if any.

3.6.9 The Revert to Older Version Window

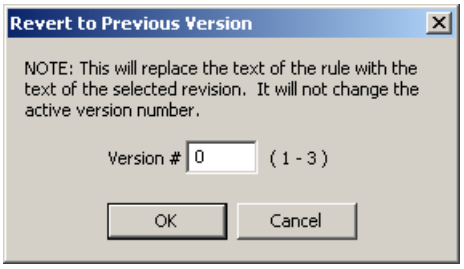
Use the Revert to Older Version window to replace the selected rule or definition with an older version of this rule or definition.

To replace a rule or definition, complete these steps:

1. Right-click on a category or a concept node and select **Revert to Older Version** from the drop-down menu that appears.



The Revert to Previous Version window appears.



2. Enter the version number in the **Version #** field. The numbers in parentheses (**()**) represent the range of changes made to the selected node. For example, theses numbers indicate the range of versions `1-3`.
3. Click **OK**.

3.6.10 The Import Category (or Concept) Window

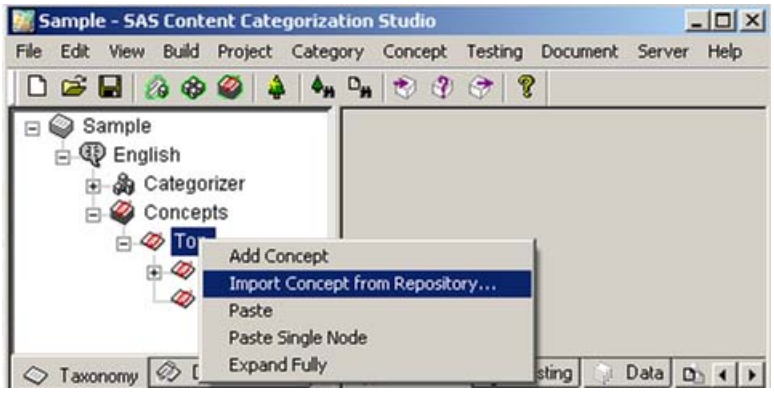
A project, or database, administrator can use the Import Concept or Import Category windows to add children to the selected node. Choose to download concepts and categories from a project on the server into the current project using these import operations. For example, if you created a project with a concept that you also want to use in your current project, use this import operation.

When you import a concept or a category, this node appears as a child of the selected node in the taxonomy. For example, if you want your concept to be a top level concept, select `Top`. If you select a child of the `Top` node, the imported node becomes a child of that node.

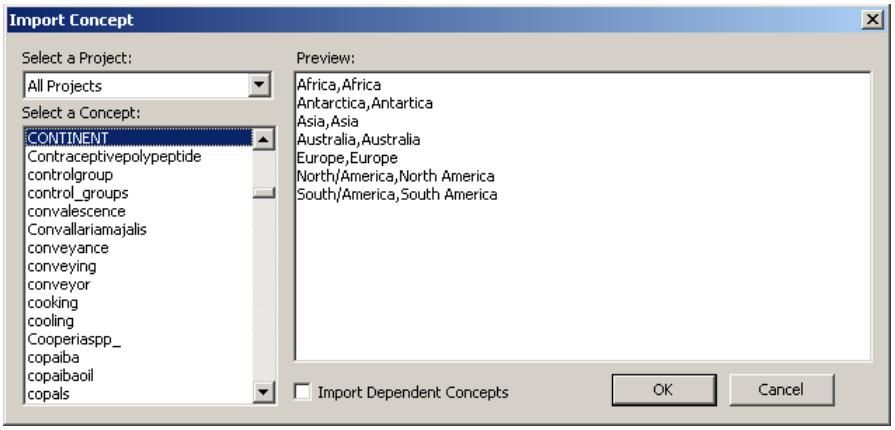
The example provided is for a concept. Adapt these steps to import a category.

To open and use the Import Concept window, complete these steps:

1. Right-click on a node in the concepts branch of the taxonomy. This node becomes the parent of the concept that you import. For example, if you want your concept to be a top level concept, select `Top`.



2. Select **Import Concept from Repository** in the drop-down menu that appears. The Import Concept window appears:



3. (Optional) Click ▼ to the right of the **Select a Project** field and a drop-down menu displaying a list of all of the SAS Content Categorization Collaborative Server projects residing on the server appears. If you select one of these projects, the display of concepts that you can import is limited to this project.

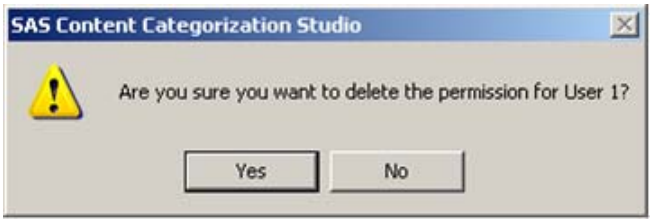
Hint: By default, select `All Projects` in the **Select a Project** field is selected. This is true although `All Projects` does not appear in the **Select a Project** field. This operation enables you to see all of the concepts for all of the SAS Content Categorization Studio projects on the server.

---

4. See the definition for the selected concept in the **Preview** screen. If there is an error with the database see `Error receiving preview for concept`.
5. Click the **Import Dependent Concept** check box, located below the **Preview** pane, to import any dependent concepts for the selected concept.
6. Click **OK** to close this window.

3.6.11 An Example of Status Windows

When you perform an operation that has significant consequences, a SAS Content Categorization Studio confirmation window might appear. For example, after you delete the permission level for a user, a SAS Content Categorization Studio confirmation window appears.



To confirm this operation, click **Yes**.

4
Getting Started with Collaboration

- *How to Begin Collaborative Work*
- *Cached Project*
- *Open a Project*
- *Keeping Projects Up-to-Date*
- *Understanding Your User Permission Level*
- *Working with a Cached Project*
- *Collaborative Changes*
- *Server Operations*
- *Save a Project*

4.1 How to Begin Collaborative Work

This chapter explains how to set up, and begin working within, a SAS Content Categorization Collaborative Server project. Use this chapter after you read Chapter 3: *Using the Interface* and before Chapter 5: *Other Collaborative Operations*.

The process of two or more developers working together on one project, whether their permission settings are equivalent or different, is defined as *collaboration*. A SAS Content Categorization Studio project that is used for collaborative work, resides on a remote server. This server enables multiple subject matter experts to work together on the same project that is cached on their local machines. The project that resides on the server is called a *remote project*.

Multiple developers can specify the categories and concepts that comprise a taxonomy, write, and modify category rules and concept definitions, while revising a project. By committing changes from local machines to the server, and updating the local project from the server, the project is kept up-to-date.

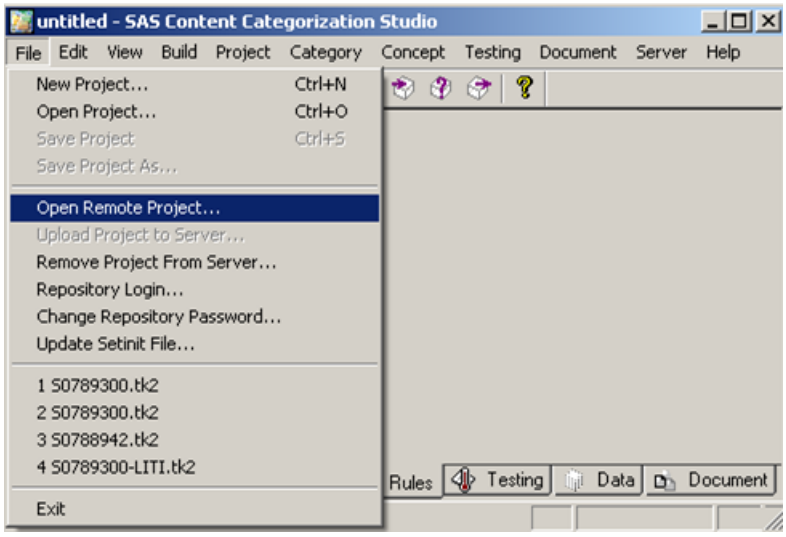
4.2 Cached Project

A copy of the project is automatically saved to your local machine. This project is stored in the *Shared Projects* folder that is automatically created by SAS Content Categorization Studio. When you log in to the repository, you open your local version of the selected project. You can simplify the process of synchronizing the local project with the version stored on the server by using the installation-specific settings that are available in the Options window. For more information, see Section 3.4 *The Options Window* and Section 4.6 *Working with a Cached Project*.

4.3 Open a Project

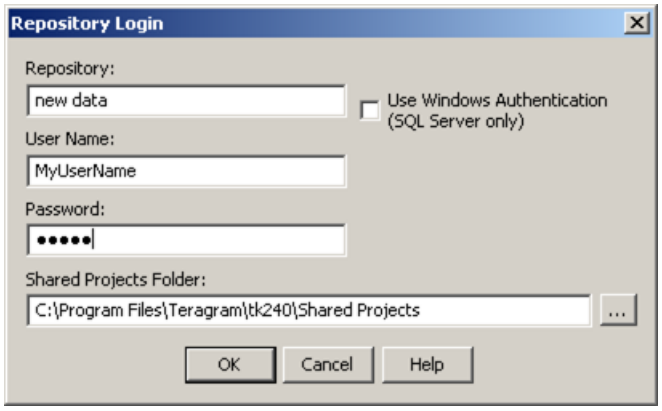
To begin working in your collaborative projects, complete these steps:

1. Select **Start --> Programs --> SAS --> SAS Content Categorization Studio --> SAS Content Categorization Studio**.



The untitled SAS Content Categorization Studio user interface appears.

2. Select **File --> Open Remote Project**. The Repository Login window appears.



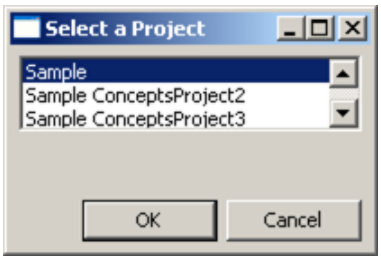
1. Enter the name of the ODBC data source into the **Repository** field.
2. (Optional) Select **Use Windows Authentication (SQL Server only)** to use the Windows user account principal token to connect. If you select this check box, the **User Name** and **Password** fields are grayed.
3. Enter your user name into the **User Name** field.
4. Enter your password into the **Password** field.

Hints: If you previously used SAS Content Categorization Collaborative Server, the **Repository**, **User Name**, and **Shared Projects Folder** fields are all filled in.
In Windows Vista/2008 Server/7 a preference is set so that data is not stored in the Program Files directory hierarchy.

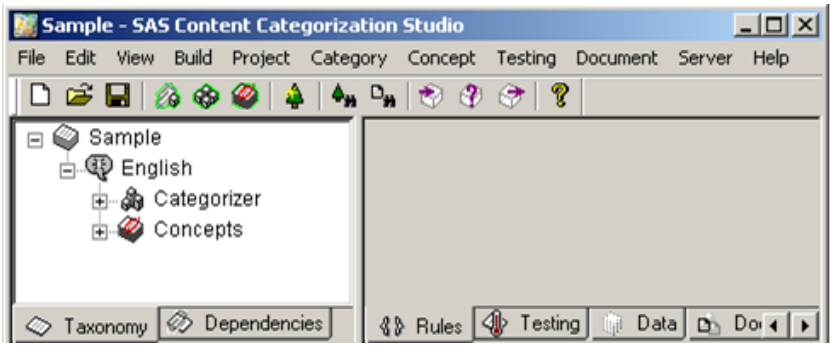
---

5. (Optional) Click  under the **Shared Projects Folder** heading and the Select a Directory window appears. Use this window to select a location for the Shared Projects folder. For more information, see Section 3.6.2 *The Select a Directory Window* below.

6. Click **OK**. The Select a Project window appears.



7. Select a project. For example, choose `Sample`.
8. Click **OK**. The selected project appears in the interface.



4.4 Keeping Projects Up-to-Date

4.4.1 Overview of Keeping Projects Up-to-Date

You can keep your local project up-to-date with the project on the server. You can either automate this process, or choose to perform these operations manually.

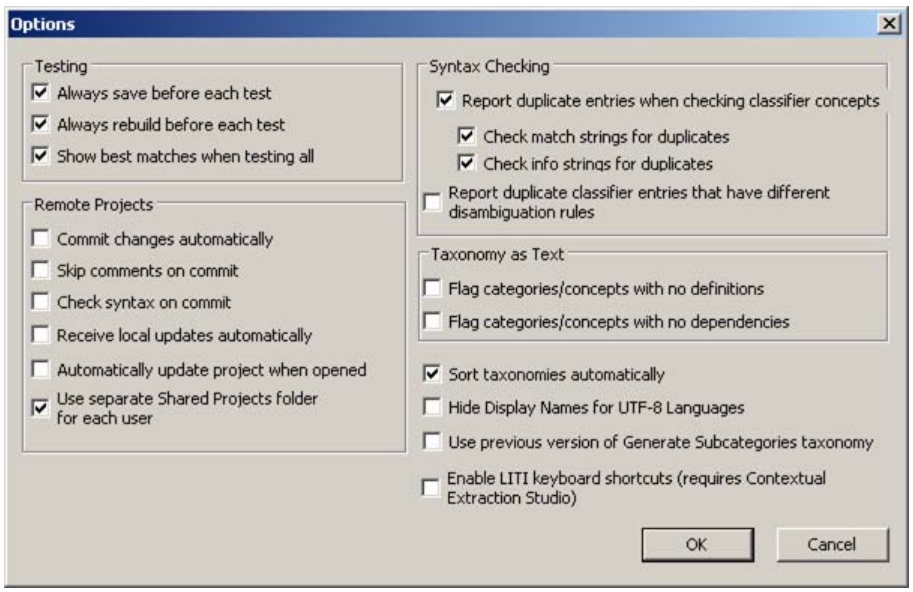
4.4.2 Specify Options

The Options window enables you to set installation-wide settings for collaborative work. Use any of the **Remote Projects** selections in this window to simplify the process of keeping your local project up-to-date with the project on the server. These operations are automatically performed by SAS Content Categorization Studio to save you time.

When you make choices using the Options window, they remain selected as the default operations until you uninstall and reinstall SAS Content Categorization Studio or make new selections. Options are installation-specific.

To specify your installation-specific options, complete these steps:

1. Select **Edit --> Options**. The Options window appears.



2. Select any of the following collaborative operations that are listed under the **Remote Projects** heading:

Commit changes automatically

Automatically commit rule and definition changes to the project on the server. This process occurs after you edit a rule or definition and select a different taxonomy node.

Skip comments on commit

Commit changes to the server without tracking the reasons for these revisions.

Check syntax on commit

Automatically verify the syntax of the definition or rule before the syntax is committed to the server.

Receive local updates automatically

Automatically update the category rules and concept definitions from the server whenever a change is committed by another user.

Automatically update project when opened

Ensure that any changes made to the taxonomy, categories, and concepts are automatically downloaded to your project when it is opened.

Use separate Shared Projects folder for each user

(Default) Enable two or more users to use the same machine to build projects in SAS Content Categorization Studio. These users can maintain separate local (cached) projects. These projects are stored in the Shared Projects folder that is created on your local machine.

3. Click **OK** to save your changes.

You can also perform some of these operations manually. For more information, see Section 4.4.3 *Manually Accessing Server Operations* below.

4.4.3 Manually Accessing Server Operations

Manually use the server operations that are available when you right-click on the nodes in the Taxonomy window. Similar operations are also available in the **Server** drop-down menu, but because they are located in the **Server** menu, the word *server* is omitted:

Server Status

See the status of your taxonomy.

Server Update

Obtain any changes committed by other users and overwrite any local changes made to a single category, or concept, by using its counterpart on the server.

Server Commit

Update the category rule or concept definition on the server with the local rule or definition that you develop.

Revert to Older Version

Select an earlier version of the rule or definition and restore this statement as the current category rule or concept definition.

4.5 Understanding Your User Permission Level

The key to successful collaboration is keeping each user's cached project current with the changes that are made according to the expertise of other users.

Permissions ensure that all users have access to the project that are consistent with their level of expertise. By default, all users are assigned `ReadOnly` permissions. If you are assigned a higher permission level by an administrator, you can work on the project.

The levels of permissions are ordered from the highest levels of access:

`Read Only`

Users can see the SAS Content Categorization Collaborative Server project. These users are unable to make any changes to the project.

`Read and Write Rules`

Subject matter experts can read the project and develop category rules and concept definitions for the taxonomy.

`Read, Write, and Change Taxonomy`

Users see the project, develop rules and definitions, and modify the taxonomy.

For more information, see Section 4.5 *Setting Permission Levels in a Project*.

4.6 Working with a Cached Project

After a project is uploaded to the server and opened on your machine, a local copy is automatically stored in the Shared Projects folder on your hard drive. This cached project is the project that appears in the user interface. For example, see the Sample project shown in Step 8 on page 63.

The cached version of the SAS Content Categorization Collaborative Server project contains the changes that you make to the project, unless you select another operation or cache location.

As you modify the project, it is important to keep your local project up-to-date with the project on the server. For more information, see Section 4.8 *Server Operations*.

However, there are reasons for saving changes only to the local project. To see an overview of these purposes for the Save As operation, see Section 4.9 *Save a Project*.

4.7 Collaborative Changes

4.7.1 Overview of Collaborative Changes

When you are working on a collaborative project, each developer ensures that all of their changes are stored in the project repository. This makes it possible for other developers, working on the same project, to access the updated project. The two major types of changes stored in the repository are taxonomy and syntax changes:

Taxonomy changes are defined as the addition, deletion, and renaming of categories and concepts. These changes are automatically committed to the server. However, they are permitted only if your cached project has a taxonomy structure that is identical to the project on the server.

Syntax changes are defined as the changes that are made to a category rule or to a concept definition. These changes are committed to the server either singularly, meaning one rule at a time, or collectively where all of the changed rules are committed at the same time.

4.7.2 Modifying the Taxonomy

The taxonomy structure is modified whenever you choose to add, delete, or rename a category or a concept. Each of these taxonomy changes requires the server to be up-to-date. Only users with the `Read, Write, and Change Taxonomy` permission level can perform these operations. If the taxonomy is not up-to-date, update the local project that is cached on your machine. For more information, see Section 4.8.3 *Update from the Server*.

> **Note**: The symbolic links component is limited to SAS Content Categorization Studio projects that are not collaborative. You can, however, create dependencies (and dependent concepts can be uploaded with the imported concept) for SAS Content Categorization Studio collaborative projects. For more information, see the *SAS Content Categorization Studio: User's Guide*.

When you make taxonomy changes, these node revisions are automatically committed to the taxonomy.

4.7.3 Changing the Rules or Definitions

Modify rules or definitions for categories and concepts, respectively, if you are a user with `Read and Write Rules` permissions. Unlike taxonomy changes, you manually commit these modifications to the server. For more information, see Section 4.8.4 *Commit Changes to the Server*. However, you cannot change concept types. For example, you cannot make a classifier into a grammar concept.

When you select **Receive local updates automatically** in the **Remote Project** section of the Options window, the server automatically updates your category rules and concept definitions. This automatic operation is performed whenever a change is committed to the server by another user.

4.8 Server Operations

4.8.1 Understanding the Server Operations

As you begin developing and making changes to your project, you use the server operations to keep your local project up-to-date with the project residing on the server.

These commands check the status of your local copy against the project residing on the server. You can update your project, commit your changes to the server, see detailed revision logs, and return to an earlier version of the project using these operations. You can also use server operations to prevent conflicting changes that might corrupt the collaborative project. Like permissions, server operations can affect one node, or the entire project.

When you save your project, a cached version of the current project is saved only to your local machine. It is stored in the Shared Projects folder on the local machine. Multiple Shared Projects folders allow several developers to create separate projects on this machine.

The operations for server operations are listed in the table below:.

Table 4-1: Server Operations

Menu Operation	Description for Category and Concept Level Operations	Description for Taxonomy or Branch Level Operations
Server Status	See the relationship between the local and server nodes: - `Up to date`: The local version is identical to the project on the server. - `Local Changes`: The cached, or local, copy has changes that are not committed to the server. - `Out of date`: Another user made syntax changes to the server project. - `Conflict`: The syntax for the selected node has been changed at both the local and server levels. - `Deleted`: This node is not in the server project. - `OK{}`: The parent node is up-to-date with the project on the server, but one or more (the number in the curly braces (`{}`) of the child nodes might not be up-to-date.	See the correlation between your cached project, or the selected branch of your taxonomy, and the project on the server. This operation also checks to see whether a language was added to the project on the server. The nodes in the taxonomy, or selected branch, display one of the following messages: - `OK`: This message appears for either of the top two nodes when these nodes are synchronized with the project residing on the server. - `Up to date`: The cached and server versions are identical. - `Local Changes`: Your local project has changes that are not committed to the server.
Server Update	Update a single category or concept with its counterpart on the server. This operation deletes any local changes. A SAS Content Categorization Studio status window gives you the opportunity to reject the download. When the update is complete, a confirmation message appears in the Taxonomy window.	Download the current project in its entirety, or as one branch, from the server, deleting any local changes. This operation checks to see whether a language was added to the project. If not, the new language is added to the local project. When the update process is complete, an `Up to date` message appears in the Taxonomy window.
Note: For more information, see Section 4.8.2 *Using the Server Operations*.		

Table 4-1: Server Operations (Continued)

Menu Operation	Description for Category and Concept Level Operations	Description for Taxonomy or Branch Level Operations
Server Commit	Commit a changed rule or definition to the server. The `Commit Successful` string appears when this operation is successful. If the local rule or definition is the same as the one on the server, the `Already Up to Date` message appears. Unless this operation is deselected in the Options window, the Enter Comment window appears when you commit a change to the server. For more information, see Section 3.6.7 *The Enter Comment Window*.	Automatically commit changes to the server, if your project is up-to-date with the project residing on the server. The `Commit Complete` string appears to the right of the selected `Categorizer` or `Concepts` node. The `Commit Successful` string appears to the right of each changed taxonomy node when this operation is successful.
Revision Log	Open a RevisionLog screen displaying the following information for each revision that is relevant to the selected node: - date - time - user - category or concept - action taken - current category or concept version number - comments entered when the change was committed For more information, see Section 4.8.5 *Using the Revision Logs.*	

Table 4-1: Server Operations (Continued)

Menu Operation	Description for Category and Concept Level Operations	Description for Taxonomy or Branch Level Operations
Revert to Older Version	Open the Revert to Previous Version window that enables you to select an older version of the project. When you select this operation, the rule text for the selected version of the changes replaces the rule for the selected category or concept. For more information, see Section 4.8.5.C *Revert to an Older Version*. **Note**: See the RevisionLog window (see the table row above) to determine the revision number that you want to use to replace your current rule or definition.	not applicable
Import Concept (Category) from Repository	Open the Import Concept, or Import Category, window to copy a concept, or category, making the imported node a child of the selected node. For more information, see Section 4.8.6 *Import Categories or Concepts from a Repository*.	Select the `Top` node and choose this operation to open the Import Concept, or Import Category, window. Select a concept or category and make it a child of the `Top` node. For more information, see Section 4.8.6 *Import Categories or Concepts from a Repository*.

4.8.2 Using the Server Operations

4.8.2.A Check the Server Status for Single Nodes

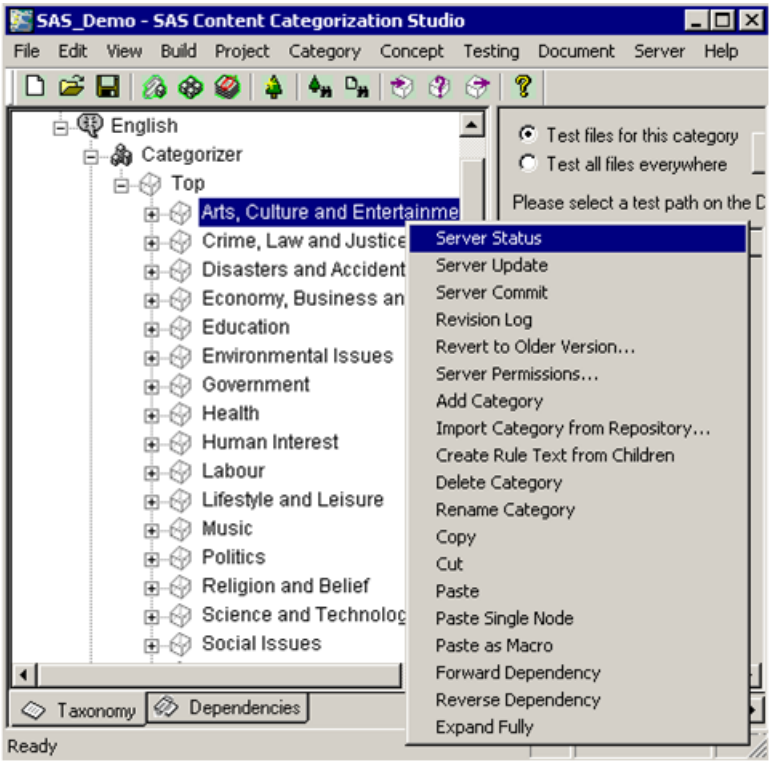
Right-click on a node, and use the **Server Status** operation in the menu that appears, to see how the currently selected rule corresponds to the matching rule on the server.

To check the server status, complete these steps:

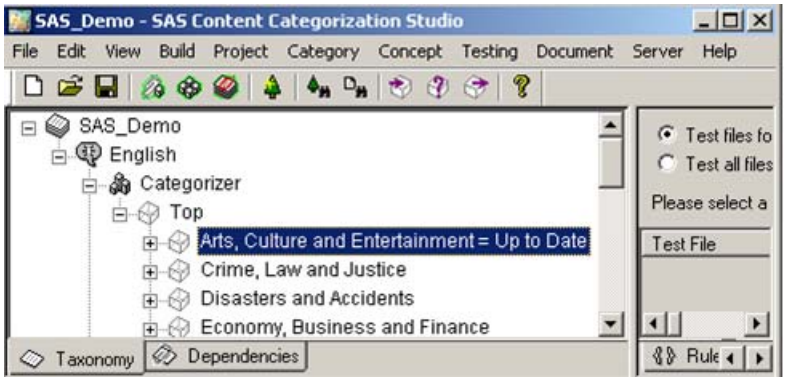
1. Right-click on an individual category or concept node.

> **Note**: If you click the `Categorizer` or `Concepts` node, you are checking the server status for the selected taxonomy branch.

2. Select **Server Status**.



A message appears in the Taxonomy window, to the right of the selected node, displaying the server status for this node.



3. (Optional) If a displayed message reads, `Out of Date`, `Local Changes`, or `Conflict`, there is a difference between the rule in the local project and the server project. When you see one of these messages, you can use either of the following two operations to correct the difference:

   - Overwrite your local changes by updating the rule with the rule on the server. For more information, see Section 4.8.3 *Update from the Server*.
   - Overwrite the rule on the server by committing your local changes to the server. For more information, see Section 4.8.4 *Commit Changes to the Server*

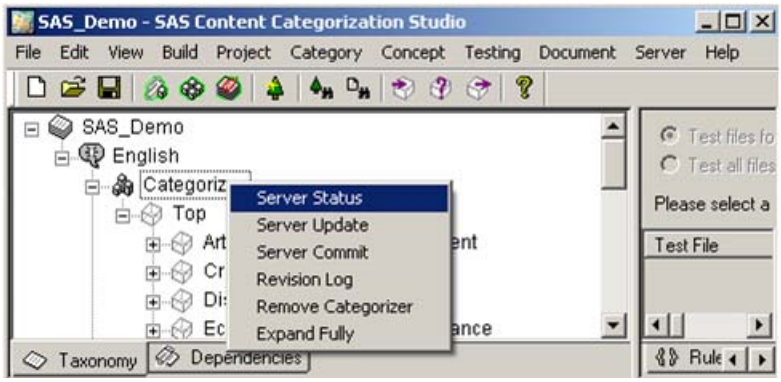
For information about removing messages from the Taxonomy window, see Section 4.8.2.C *Removing Taxonomy Tree Messages*.

4.8.2.B Checking the Status of the Taxonomy

Check the status of all of the nodes in a specific branch of the taxonomy against the project on the server at any point during the process of project development.

To check the server status for a node, complete these steps:

1. Right-click the Categorizer or the `Concepts`, node and select **`Server Status`** from the drop-down menu that appears.



2. Check the messages that appear to the right of the `Categorizer` or `Concepts` node. If you see the **`Deleted`** message, the node is removed from the project on the server, but still appears in the taxonomy of the local project.
3. (Optional) To remove these messages from the Taxonomy window, click  .

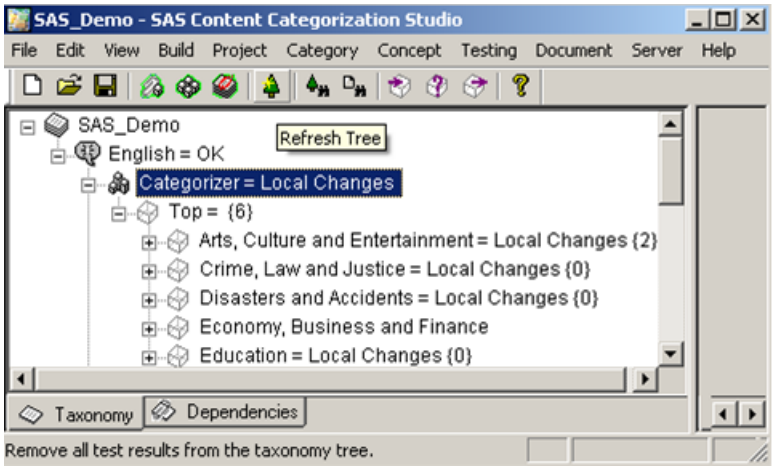
For more information and examples, see Section 4.8.2.A *Check the Server Status for Single Nodes*.

4.8.2.C Removing Taxonomy Tree Messages

You can remove the messages that appear in the taxonomy to refresh the tree.

To remove the messages, complete this step:

Click  .



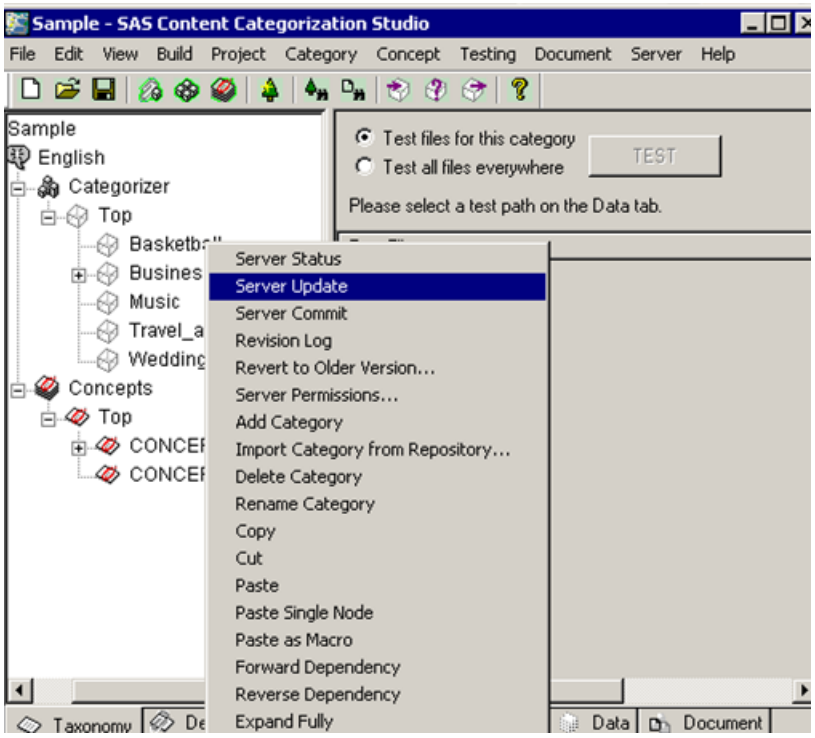
Note: The Revert to Previous Version operation does not change the rule on the server. The current rule text is replaced with the selected rule. To change the server version, a Server Commit operation is performed after the Revert to Previous Version operation.

4.8.3 Update from the Server

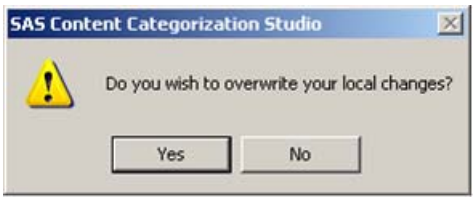
The Server Update operation modifies your local project to synchronize it with the version on the server. Use the Server Update operation when **Automatically update project when opened** is not selected in the Options window.

To download individual node changes from the server to your local project, complete these steps:

1. Right-click on a category or concept node and select **Server Update**.



A SAS Content Categorization Studio status window appears.



2. Click **Yes** to overwrite these changes.

You can receive automatic updates from the server for individual categories and concepts. For more information, see Section 3.4 *The Options Window*.

4.8.4 Commit Changes to the Server

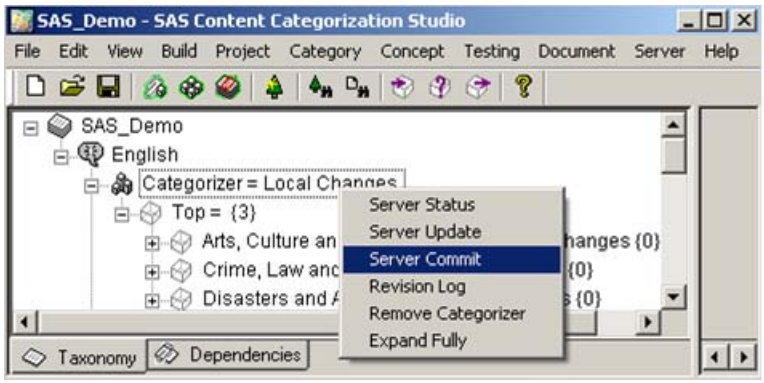
Commit each individual rule change to the server, update all of the changes to a single branch, or commit all of the changes to the project at one time. To perform any of these operations, right-click on a taxonomy node and select **Server Commit** from the menu that appears.

If you right-click on the `Categorizer` or `Concepts` node, you commit all of the changes for the selected branch. Right-click on the language node to commit all of the changes for the project. If you are building a project with more than one language, the changes are committed to the selected language branch. For more information and examples of the message that appear, see Table 3-4 on page 44.

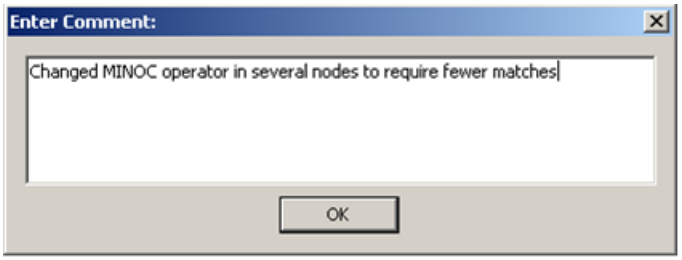
Use this operation if you did not select **Commit changes automatically** in the Options window. For more information, see Section 4.8 *Server Operations*.

To commit changes to the server, complete these steps:

1. Right-click on a taxonomy node and select **Server Commit**.



The Enter Comment window appears, unless you select **Skip comments on commit** in the Options window.



2. Enter an explanation for the changes. This comment appears the RevisionLog window enabling all of the project developers to see the reasons for any changes made to the project.
3. Click **OK**. `Commit Complete` and `Commit Successful` messages appear to the right of the relevant nodes in the Taxonomy window.

4.8.5 Using the Revision Logs

4.8.5.A Overview of Using the Revision Logs

Use the RevisionLog window to access a history of the changes that were made to a specific taxonomy branch, or to an individual node. The type of RevisionLog window that appears depends on the node that you select in the taxonomy.

Taxonomy branch

Select the **Revision Log** operation using the `Categorizer` or `Concepts` node. The RevisionLog screen that appears enables you to see the changes that were made to all of the categories or concepts in this branch of the taxonomy.

Individual category or concept node

Select an individual category or concept node that enables you to see only the history of the rule or definition changes for the selected node. For more information, see Section 4.8.5.B *Revision Logs*.

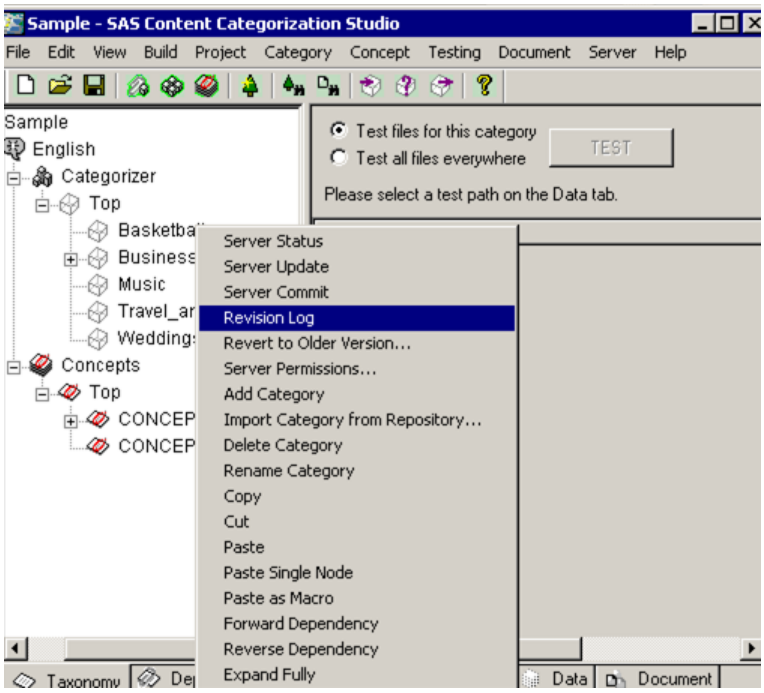
You can also use the revision logs to determine the rule or definition version that you want to use to replace the current rule or definition. For more information, see Section 4.8.5.C *Revert to an Older Version*.

4.8.5.B Revision Logs

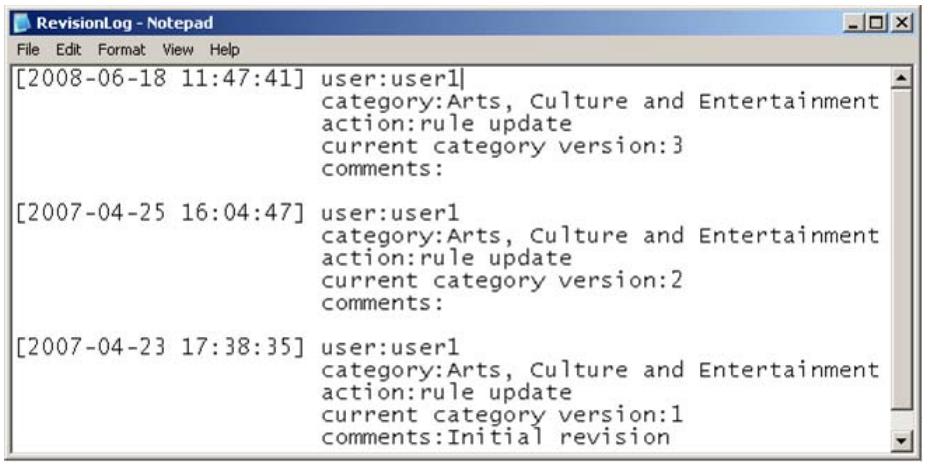
Use revision logs to see the changes made to your taxonomy, or to a selected rule or definition. These screens also enable you to select a version that you can revert to.

To open and use Revision Log screens for either an individual category or concept, or an entire taxonomy, complete these steps:

1. Right-click on a category or concept node to see only the log file for this node. Select **Revision Log**.

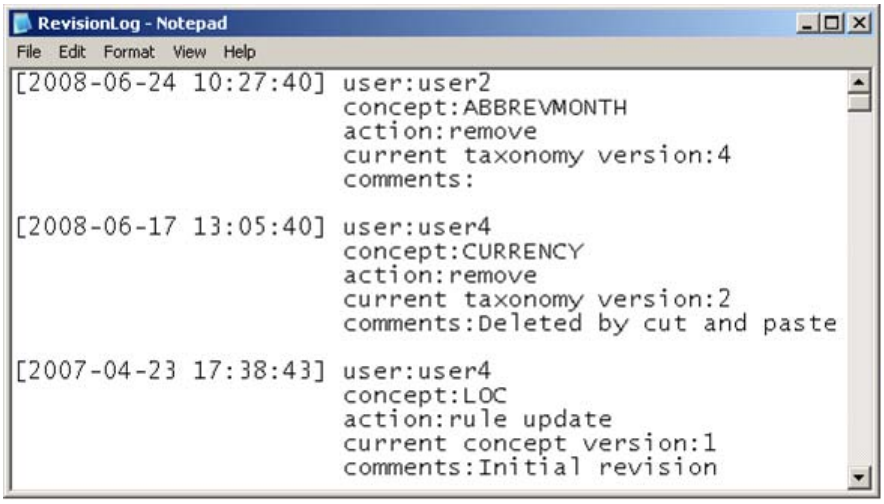


A RevisionLog screen appears.



2. Use Table 3-5 on page 54 to understand the information in this screen.
3. (Optional) If you want to revert to an older version of the rule, note the version number in this window.
4. `Right-click the Categorizer` or `Concepts` node in the taxonomy to see the RevisionLog screen for all of the concepts or categories in the

selected branch. These modifications include the addition and deletion of the nodes in this branch and any changes to the rules or definitions.



5. Use Table 3-5 on page 54 to analyze the information in this screen.
6. (Optional) If you choose to revert to an older version of the taxonomy or the node, note the version number using this window.

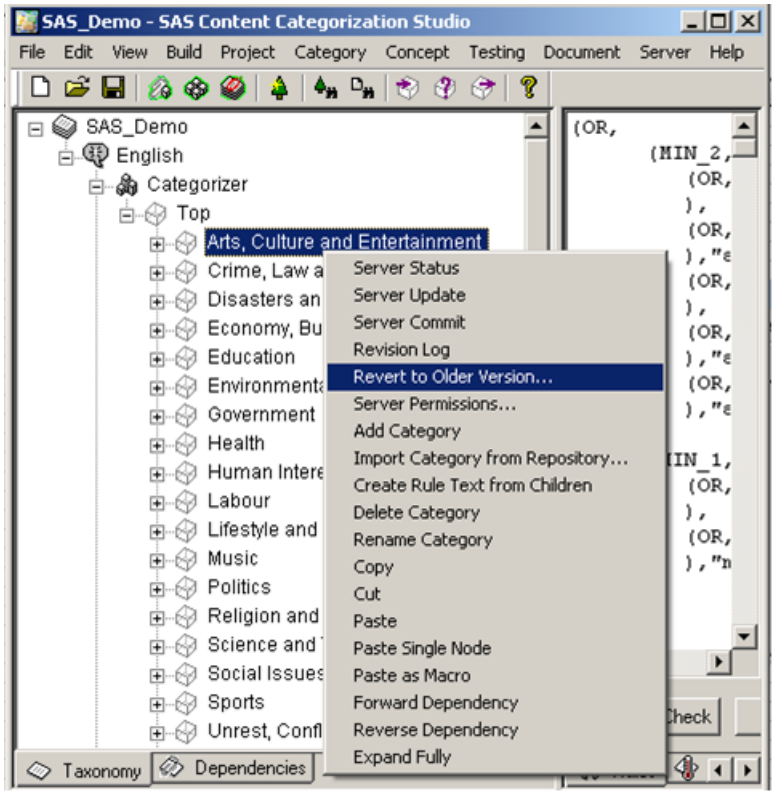
4.8.5.C Revert to an Older Version

Return to an older version of the selected category rule or concept definition after you choose a version using the RevisionLog screen. If you want to change several nodes within one branch, work through the changes node-by-node.

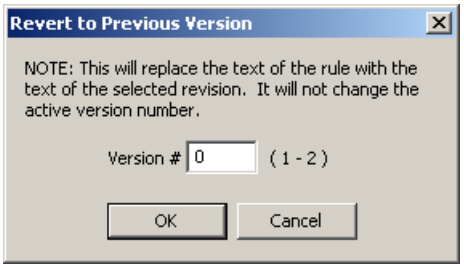
To revert to an older version of a rule or definition, complete these steps:

1. Complete Step 1 on page 81 through Step 3 on page 82.
2. Use the RevisionLog window to determine the revision number.

3. Right-click on the node. Select **Revert to Older Version**.



The Revert to Previous Version window appears.



4. Enter the version number into the **Version #** field.
5. Click **OK** to save your change.

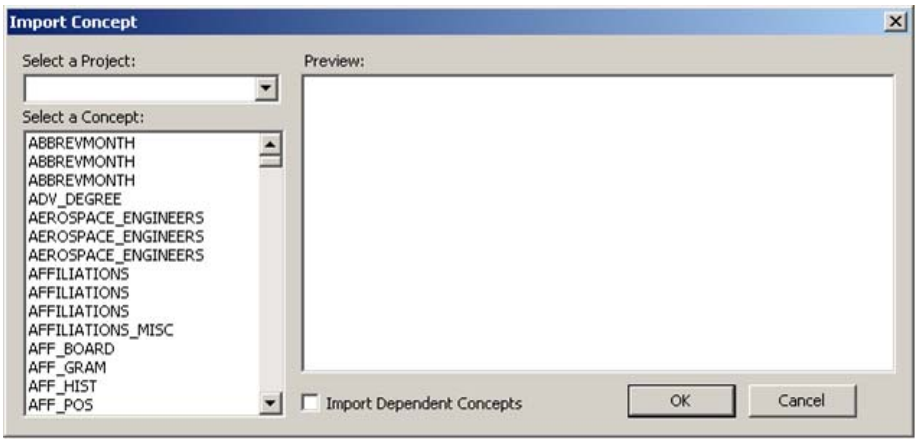
4.8.6 Import Categories or Concepts from a Repository

Add existing categories and concepts, with their definitions, to your taxonomy using the **Import Concept from Repository** operation, individually, or as a group. Each concept is imported with its definition and placed in your taxonomy as a child of the selected node. This section applies to both categories and concepts. If you are importing categories, modify the following directions as necessary.

To import a single concept, or a group of concepts, complete these steps:

1. `Right-click the Top` node, or select another concept node. This node is the parent of the imported concept. Select **Import Concept from Repository**.

The Import Concept window appears.

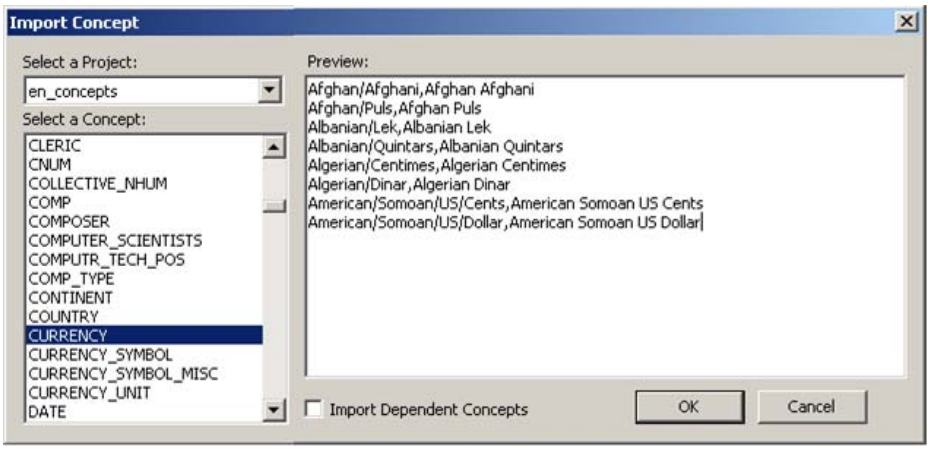


2. (Optional) Click ▼ to the right of the **Select a Project** field. Select a project to limit the display of concepts to those contained in that project. Select a project in the drop-down list that appears in the **Select a Project** field.

Hint: By default, all of the concepts for all of the projects on the server are displayed.

---

3. Select a concept in the **Select a Concept** pane.



4. Check the appropriateness of the definition for the selected concept in the **Preview** screen.
5. (Optional) Select **Import Dependent Concept** if dependent concepts appear in this definition and you also want to import these concepts.

> **Note**: The Import Dependent Concept operation does not apply to categories.

6. Click **OK** to import your selection.
7. (Optional) Use these steps reiteratively to add multiple concepts to your project.

4.9 Save a Project

Save a project to your local machine by selecting **File --> Save Project**. When you make this selection, SAS Content Categorization Studio saves a copy of this project to the Shared Projects folder on your hard drive.

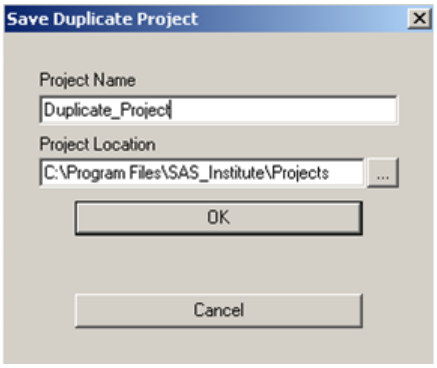
When making changes to a collaborative project, you should commit your changes to the server. However, in some circumstances, it might be better to save your project locally:

- Make changes offline
- Test syntax changes for rules and definitions
- Make project demonstrations
- Work off-site
- Build a project using a synonym list
- Using the auto rule generation tool without exporting all of the generated rules

> **Note**: When you choose to save your project locally, your changes are also committed to the server.
> To save your changes to the local machine only, use the Save As operation.

To save a project to your local machine, complete these steps:

1. Go to **File --> Save Project As** and the Save Duplicate Project window appears.



2. Enter the name of the duplicate project into the **Project Name** field.
3. Enter the path to the location of your duplicate project into the **Project Location** field.
4. Click **OK** to save this project.

When you select **File --> Save As**, the collaborative features for your SAS Content Categorization Studio project disappear.

5
Other Collaborative Operations

- *Overview of Other Collaborative Operations*
- *Sharing Test Files*
- *Change Your Server Password*

5.1 Overview of Other Collaborative Operations

This chapter provides information about collaborative operations that facilitate project development. These operations are not required to build a successful project, but enable you to perform certain operations such as sharing test files. Use these operations to facilitate your work for a collaborative project.

5.2 Sharing Test Files

5.2.1 Overview of Shared Test Files

Shared test files enable multiple users who are working on different, or the same, areas of a taxonomy to see how category rules perform. Use the same set of test files to test your category rules as another developer. You can also use this testing set to test the performance of a different set of rules for the purposes of maximizing rule matching.

You can also create a shared repository of documents that should fail to match a specific rule, but which might not fail. For example, a document with the topic *rose bushes* should not match a classifier definition rule specifying the definition for the *Rose Kennedy* category.

The operations to upload and download test files to a server are restricted to users who have, at a minimum, *Read, Write and Change Taxonomy*

permissions. With this access level, you can upload a set of test files, within their taxonomy structure, to the server where they can be downloaded by users with appropriate permissions.

5.2.2 Before Uploading or Downloading Test Files

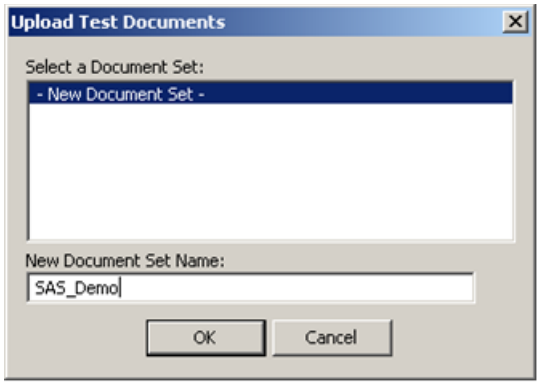
Before you can use the upload and download operations, initialize the document repository where sets of testing documents can be associated with a project and stored. For more information, see *SAS Content Categorization Collaborative Server: Administrator's Guide*.

5.2.3 Upload Test Files

Before other users can access the test files that you want to share, upload these files to the server. Although any user can create a set of test files to share, only an administrator can upload these files.

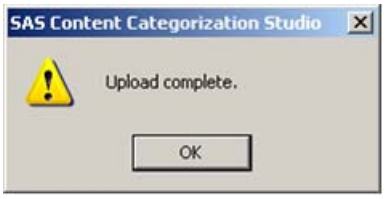
To upload a set of test files to the server, complete these steps:

1. Select **Server --> Upload Test Files** to access the Upload Test Documents window.



2. Select the testing directory to upload in the **Select a Document Set** field. Alternatively, enter the name of the testing directory to upload into the **New Document Set Name** field.

3. Click **OK** and a SAS Content Categorization Studio status window appears.



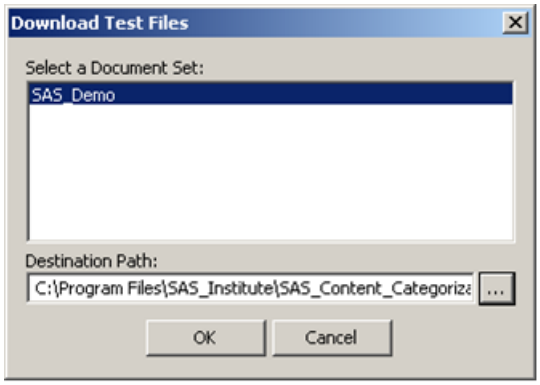
4. Click **OK** to close this window.

5.2.4 Download Test Files

After a set of test files has been uploaded to the server, a user with a minimum of *Read, Write, and Change Taxonomy* permissions can download this set.

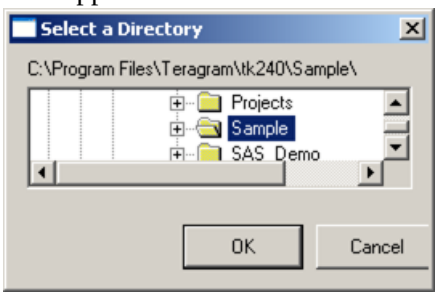
To download a set of testing documents, complete these steps:

1. Select **Server --> Download Test Files**. The Download Test Files window appears.



2. Select a testing folder in the **Select a Document Set** pane.

3. Click ... to the right of the **Destination Path** field. The Select a Directory window appears.



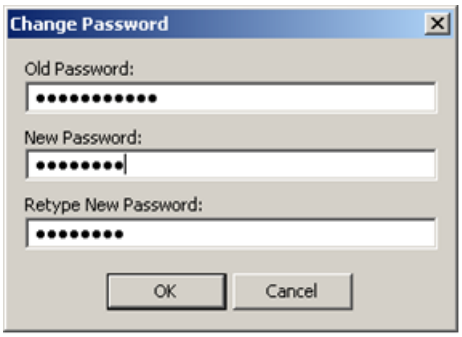
4. Select the folder to store these documents on your local machine.
5. Click **OK**. The path to the selected directory appears in the **Destination Path** field.
6. Click **OK** in the download Test Files window to complete this operation.

5.3 Change Your Server Password

You can change your server password during at any time during project development.

To change your password, complete these steps:

1. Select **File --> Change Repository Password** and the Change Password window appears.



2. Enter your existing password into the **Old Password** field.
3. Enter your new password into the **New Password** field.
4. Enter your new password into the **Retype New Password** field.
5. Click **OK** top save your changes.

Part 2: LITI Concepts

- Chapter 6: *Interface Components*
- Chapter 7: *Writing Contextual Extraction Concept Definitions*

6
Interface Components

- *Your First Look at the LITI Interface Components*
- *Start Using LITI*
- *The LITI Check Box in the Options Window*
- *The LITI Radio Button in the Definition Tab*
- *The Priority Setting in the Data Window*
- *The Predefined LITI Concepts Window*
- *The Concept Priorities Window*
- *The Compile Concepts Window*
- *The Project Settings Interface*
- *The Matched Concepts Information Windows*
- *The Export Results Wizard*
- *The Upload LITI Operation in the Build Menu*
- *Using the <language>.li File*

6.1 Your First Look at the LITI Interface Components

The interface components specific to LITI concepts appear in the SAS Content Categorization Studio interface with classifier and grammar concepts. This chapter presumes that you have a working knowledge of SAS Content Categorization Studio and have read *SAS Content Categorization Studio: User's Guide*. For this reason, the information in this chapter is specific to LITI components.

6.2 Start Using LITI

To access LITI definitions in SAS Content Categorization Studio, select **Start —> Programs —> SAS Content Categorization Studio —> SAS Content Categorization Studio**. The SAS Content Categorization Studio user interface appears.

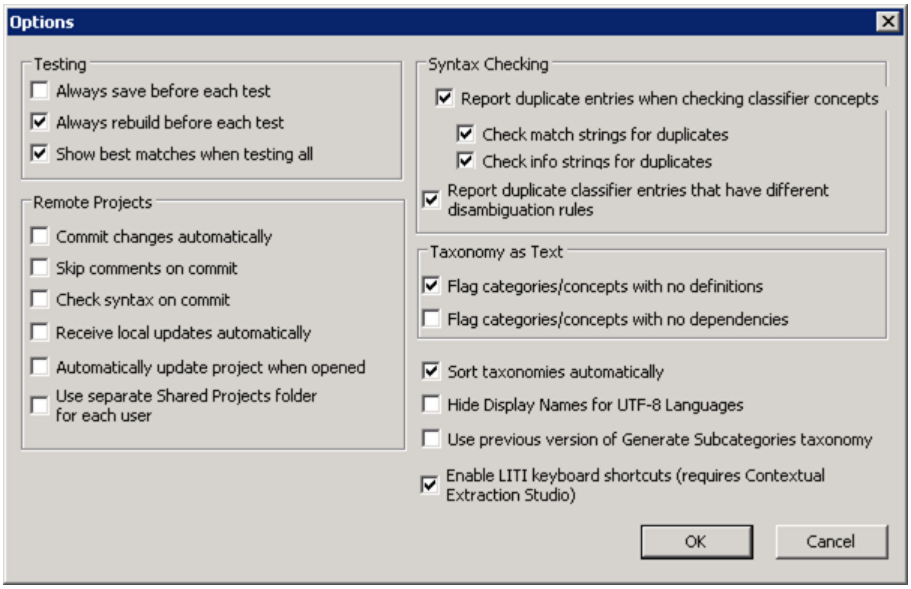
6.3 The LITI Check Box in the Options Window

By default, the **Enable LITI keyboard shortcuts** check box is enabled in the Options window. This operation enables you to enter a definition type. Scroll through a list of these types when you press Ctrl and click the up or down arrow button in the Definition pane. For more information, see Section 7.4.2 *Adding a Rule Type to the Definition Pane* on page 129.

> **Note**: When you try to remove a rule type, use the Backspace or Delete buttons with care to avoid entering a second instance of the rule type.

Deselect this check box to enter the rule types manually.

Display 6-1 Using the Options Window

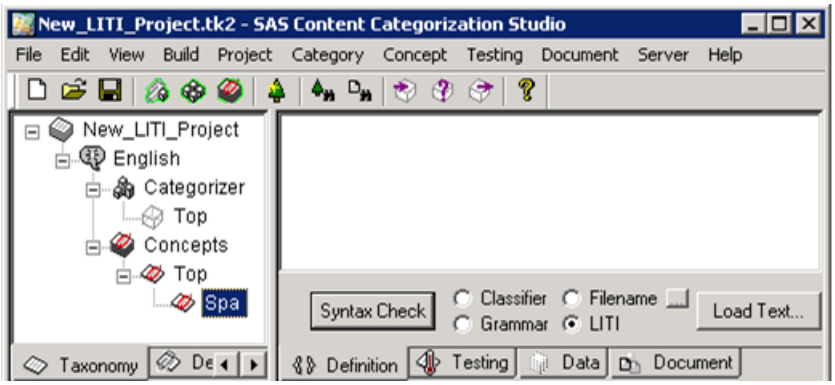


6.4 The LITI Radio Button in the Definition Tab

The **LITI** button appears in the Definition window after you add one or more concepts to your taxonomy. This button enables you to specify the LITI definition type for your concepts.

To specify a LITI concept, complete these steps:

1. Right-click the `Top` node and select **Add Concept** from the drop-down menu that appears.
2. Name the concept.



3. Select **LITI**. This radio button becomes available only after you name the concept.

> **Note**: Select **LITI** *before* you write the concept definition.

4. Write the concept definition. For more information, see Chapter 7: *Writing Contextual Extraction Concept Definitions*.

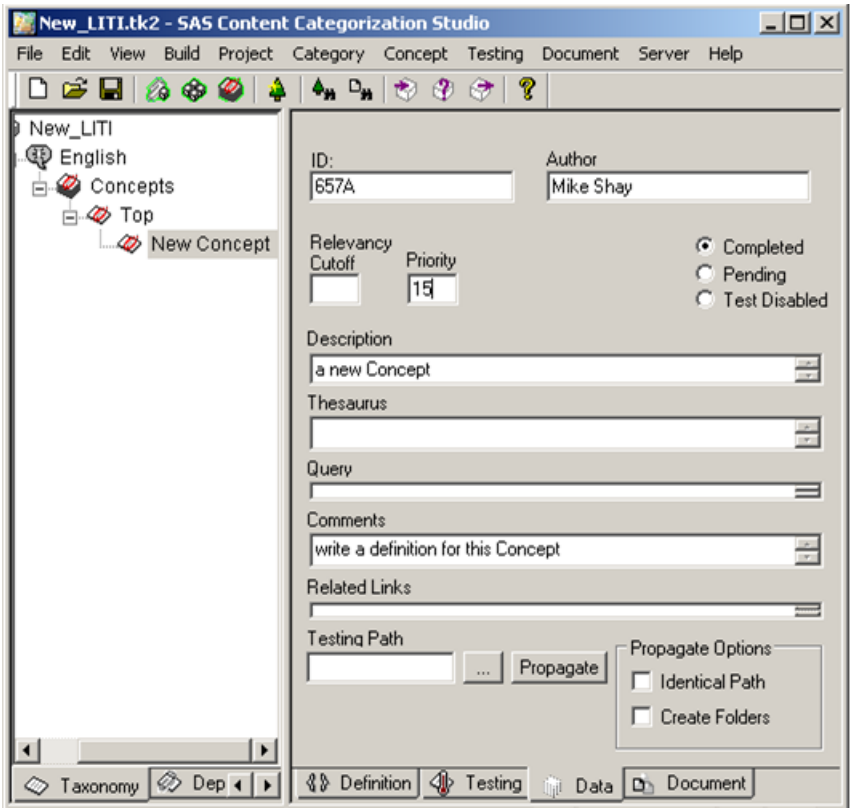
6.5 The Priority Setting in the Data Window

The **Priority** field in the **Data** tab enables you to set the ranking order of LITI concepts higher, or lower, than classifier or grammar concepts. The default setting, `10`, ranks all of the LITI concepts higher than the default settings for classifier and grammar concepts.

You can change the default setting, one concept at a time. If you do not want to prioritize the selected LITI concept, specify `0`. To increase the priority of the selected LITI concept enter a number that is higher than `10`.

To reset the **Priority** setting for one LITI concept, complete these steps:

1. Select the **Data** tab.
2. Enter a new number into the **Priority** field.



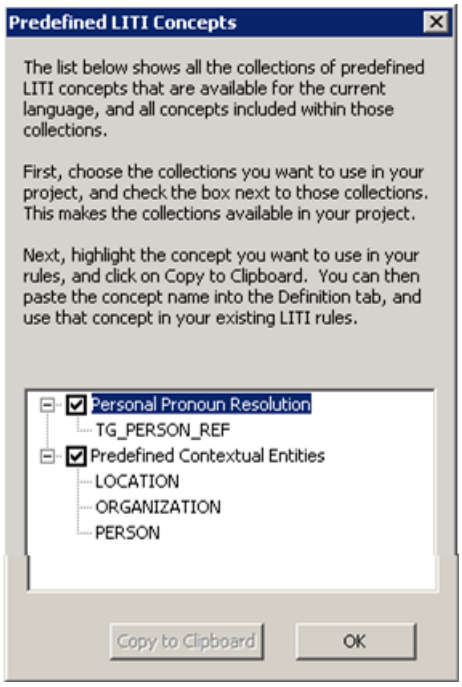
3. Select **Build —> Compile Concepts**.
4. Select **File —> Save**.

6.6 The Predefined LITI Concepts Window

Use the Predefined LITI Concepts window to select a predefined LITI concept definition. The predefined LITI definitions that you select can be added into a LITI rule, or comprise an entire rule.

To access and use the Predefined LITI Concept window, complete these steps:

1. Go to **Concept --> Show Predefined LITI Concept List**.
2. The Predefined LITI Concepts window appears.



3. Select a concept from the list and the **Copy to Clipboard** button is activated. These are the concept types for English and might differ for the languages for which this feature is available:

Personal Pronoun Resolution

Specify with a `CLASSIFIER` rule that specifies the name, or names, that are returned with the pronouns that reference the appropriate noun.

Predefined Contextual Entities

Matches entities in your documents based on the contextual information in your documents.

4. Click **OK** to close this window.
5. Paste the selected concept into a LITI rule. If you select a concept under `Personal Pronoun Resolution`, the `ACTIVATE` term is automatically entered into the definition.

Note: You can paste a Predefined Contextual Entities concept into a Classifier or a Grammar concept. However, predefined concepts do not work as expected in these regular SAS Content Categorization Studio concepts.

6.7 The Concept Priorities Window

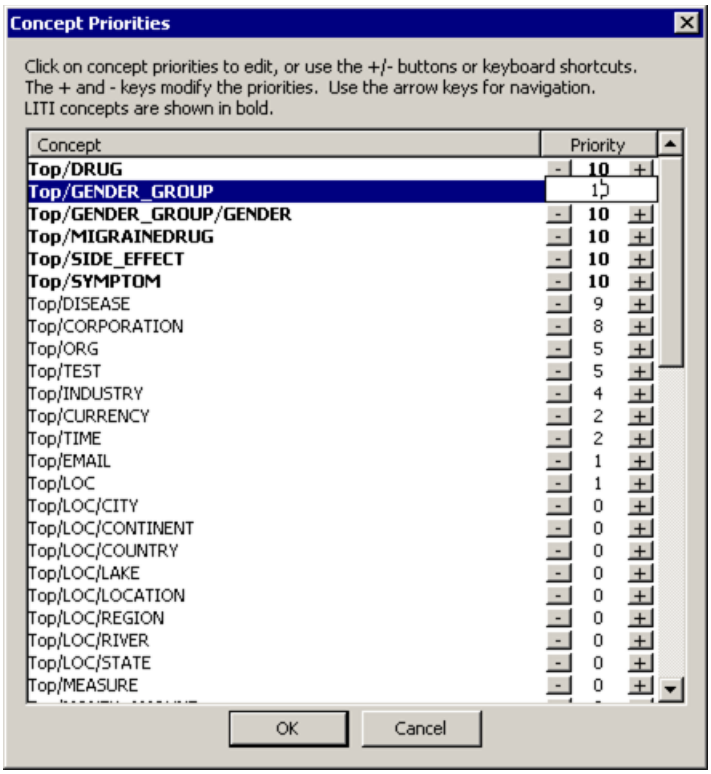
The Concepts Priorities window displays the priority settings for all of the concepts in your project. The LITI concepts appear in bold type. By default, classifier and grammar concepts have a default specification of `0` and LITI concepts have a default setting of `10`. Priority determines the matching concept when one input document matches two or more concepts and no other determiner makes one concept a better match than another.

The Concepts Priorities window displays the priorities setting from each Data window in the concepts taxonomy and ranks these concepts. You can also use the Concept Priorities window to reset the priorities and to sort from A-Z, or from highest priority setting to the lowest.

Note: By default the **Priority** setting is set to `10` in the Data window for LITI concepts.

To access the Concept Priorities window, complete these steps:

1. Select **Concept --> Priorities**. The Concept Priorities window appears.



2. See a ranked list of concepts according to the priorities specified by default, or set by you.
3. (Optional) Select a concept priority setting and enter a new number to change the priority for the selected concept. For example, change the number `10` to `8`.

4. (Optional) Click **Concept** to list the concepts from A - Z.
5. (Optional) Click **Priority** to prioritize the concepts from highest to the lowest number.

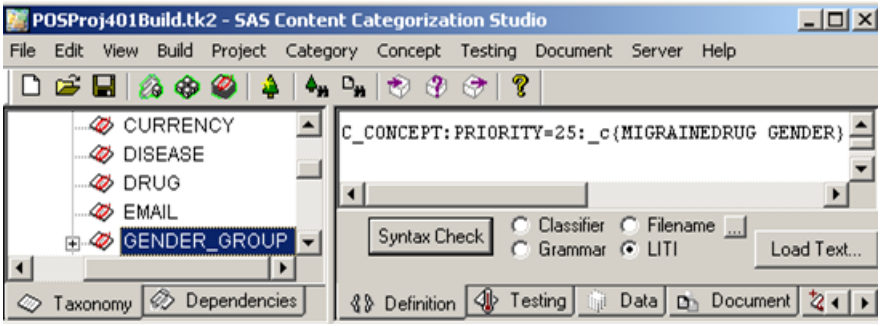
Notes: At this time, reverse sorting is not available for priorities.
The Concept Priorities window does not list concept matches.

---

6. Click **OK** to save your changes.

If you specify `PRIORITY=` in a rule and you reset the priority in the Concept Priorities window, SAS Content Categorization Studio uses the highest specification when a rule matches text. See the following rule example:

Display 6-2 Specifying a PRIORITY Setting in a Rule



In this example, if the priority setting in the Concept Priorities window is reset to `30`, a match is assigned a priority of `30`. However, if instead the priority setting in the Concept Priorities window is reset to `12`, the match is assigned a priority of `25`.

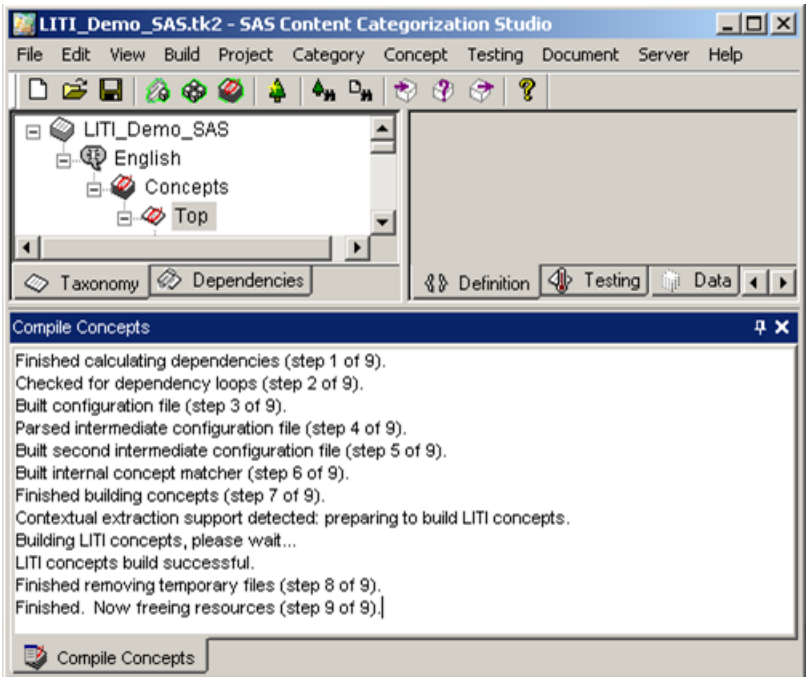
6.8 The Compile Concepts Window

Compile your concepts to ensure accuracy and to integrate all of the changes into the SAS Content Categorization Studio project. The Compile Concepts

pane that appears at the bottom of the SAS Content Categorization Studio interface displays information about the LITI concepts. This data confirms a successful build, or it points to errors in the definitions.

To compile your concepts, complete these steps:

1. Select **Build —> Compile Concepts** and the Compile Concepts window appears at the bottom of the SAS Content Categorization Studio interface.



2. Locate the lines explaining the LITI concepts.

   These self-explanatory messages include the following:

   - `Contextual extraction support detected: preparing to build LITI concepts`
   - `Building LITI concepts, please wait...`
   - `LITI concepts build successful`. If unsuccessful, the build fails and an explanation appears.

3. Click **X** in the upper right-hand corner of the Compile Concepts window to close this pane.

6.9 The Project Settings Interface

Set project-wide settings for your LITI concepts by using the Project Settings - LITI window. These settings determine how matches in input documents are returned. For this reason, the settings that you specify in the Project Settings - LITI window affect the testing results that you see in the Document window and those returned by SAS Content Categorization Server.

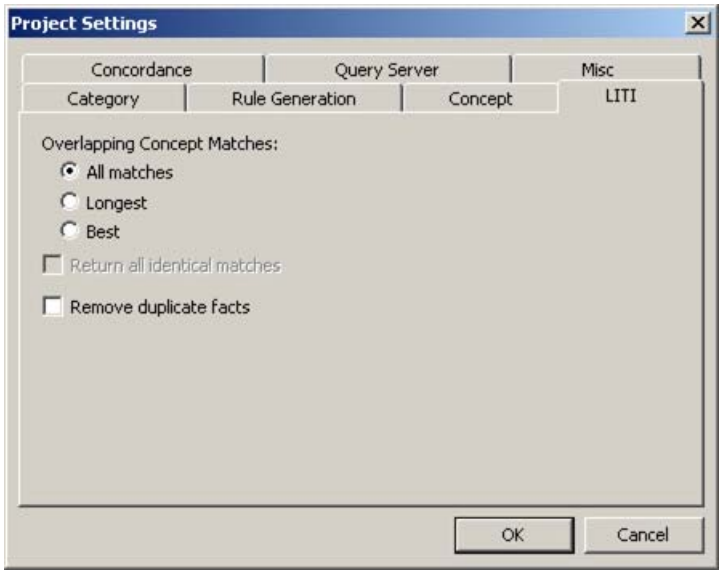
Use the Project Settings Concordance window to specify the surrounding text that is returned with a match, if you choose to use the Concordance selection in the Document window. For more information, see the *SAS Content Categorization Studio: User's Guide*.

These radio buttons and check boxes enable you to make decisions concerning overlapping and identical matches and to remove all duplicate facts. The term *fact* is used to refer to two or more concepts or tokens. These concepts, or terms, are specified in one definition in order to define a relationship between otherwise isolated instances of information. For more information, see Section 7.9 *Locating Facts* on page 187.

To specify settings in the Project Settings - LITI window, complete these steps:

1. Select **Project —> Settings** and the Project Settings window appears.

2. Select LITI and the Project Settings - LITI pane appears.



3. (Optional) Select a radio button under the **Overlapping Concept Matches** heading that determines how SAS Content Categorization Studio treats overlapping matches. Overlapping matches are strings where part, or all, of the string matches more than one concept.

   - Leave the default selection, **All matches**, selected and SAS Content Categorization Studio returns all of the terms that match any of the LITI concept definitions in this project.
   - Select **Longest** to return the longest match for the concept definition.
   - Select **Best** to return only the match with the highest priority setting.

Note: If all of the tested concepts have the same priority setting, only the longest matches are returned. For more information, see Section 7.6.19 *The Priorities and Project Settings* on page 138.

---

4. If you select either the **Longest** or **Best Matches** radio button, **Return all identical matches** becomes available. Select this check box and SAS Content Categorization Studio returns all of the identical longest or best matches.
5. Select **Remove duplicate facts** when you want to limit matches on a fact to one occurrence. This selection applies only to a predicate sequence, or to predicate, definitions. For more information, see Section 7.9 *Locating Facts* on page 187.

Note: These settings do not affect the returns specified by the `REMOVE_ITEM` rule that excludes matches on a concept for disambiguation purposes. For more information, see Section 7.8.6 *Disambiguating Matches* on page 158.

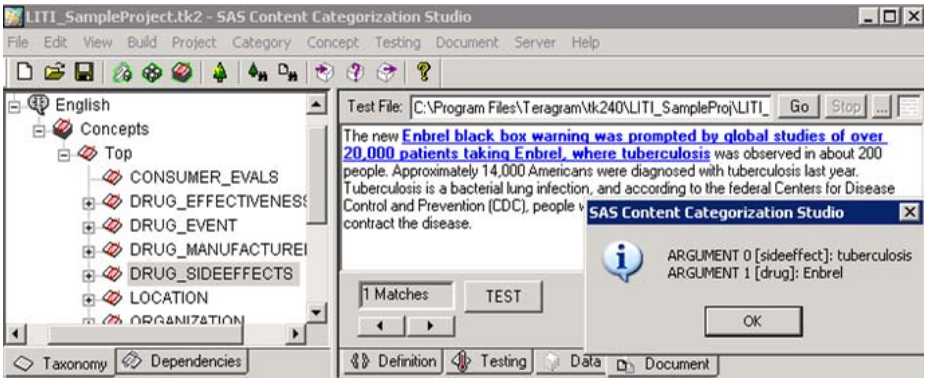
For information about the XML fields in the **Misc** tab, see the *SAS Content Categorization Studio: User's Guide*.

6.10 The Matched Concepts Information Windows

You can see more information about the matched arguments in `PREDICATE_RULE`s, coreference rules, and the information strings of `CLASSIFIER` rules. To see this information in the SAS Content Categorization Studio pop-up window, click on the blue highlighted match in the document pane.

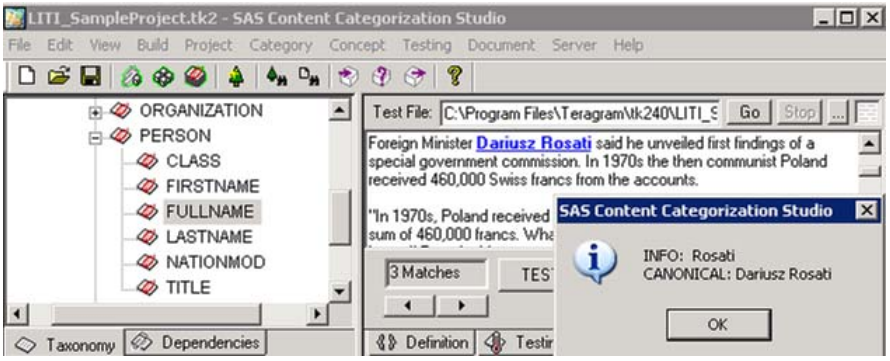
An example of the window that appears for matches on the arguments in a `PREDICATE_RULE` is shown below:

Figure 6-3 Matched Arguments



Matches on `CLASSIFIER` rules with information strings return the matched string in the **INFO** section of the SAS Content Categorization Studio window. Rules that specify coreference return information about the canonical form of the word in the **CANONICAL** section. See the following example where both strings are returned for one match.

Figure 6-4 Matched Information and Canonical Strings



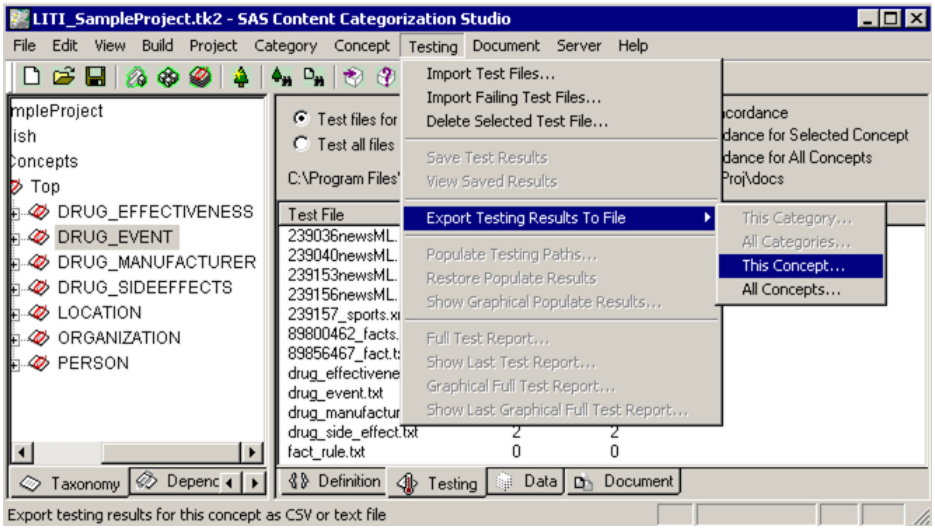
If you click on another type of rule match and do not see the expected results, see Appendix A: *on page 229*.

6.11 The Export Results Wizard

Use the Export Results Wizard to place testing results into a `.csv`, or a `.txt`, file that can be turned into SAS data sets or used with *Microsoft Excel*. You can select a check box to make this *Notepad* file automatically appear after the **Testing --> Export Testing Results To File** operation is complete.

To access this *Notepad* file and to use the Export Results Wizard, complete these steps:

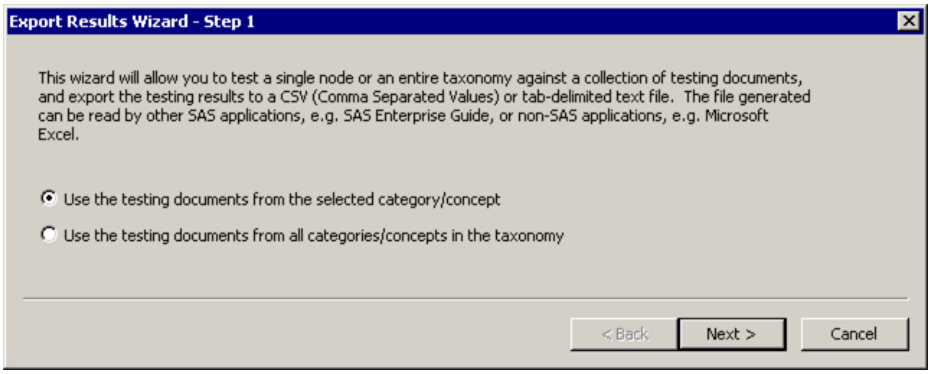
1. Select an LITI concept in the Taxonomy pane.



2. Select **Testing --> Export Testing Results To File**.
3. Choose one of the following selections:
   - **This Concept**: Export only the testing results for the selected concept.
   - **All Concepts**: Export the testing results for all of the concepts in your taxonomy.

Notes: This example uses **This concept**. The wizard pages are identical regardless of the selections that you make in this wizard.

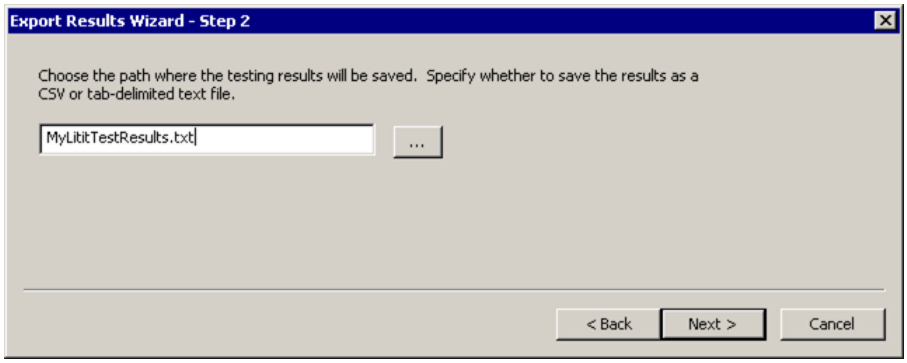
The Export Results Wizard - Step 1 appears:



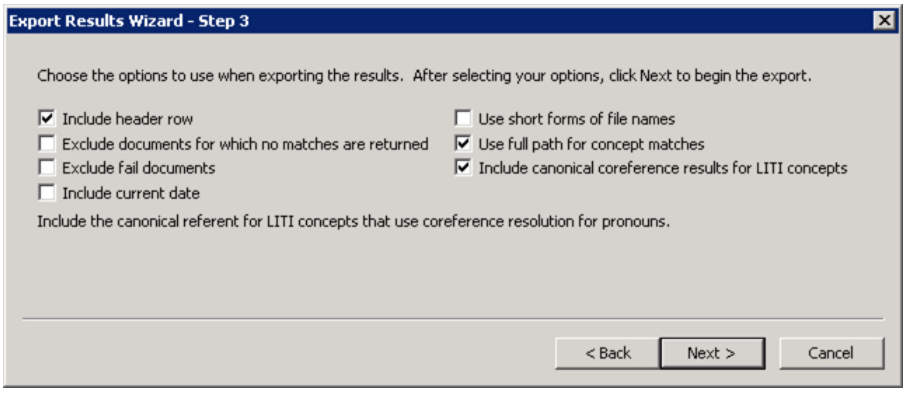
4. Select one of the following operations:

   - **Use the testing results from the selected category/concept**: (Default) Export only the testing results for one LITI concept.
   - **Use the testing results from all categories/concepts in the taxonomy**: Export all of the testing results for all of the concepts. This statement is true whether you selected a LITI concept or another concept type.

5. Click **Next** and the Export Results Wizard - Step 2 appears.



6. Click the ellipsis button (...) to choose an existing `.csv` or a `.txt` file. (You can also enter a new filename here.) For example, type `savedresults.csv`. By default, this file is stored in the `Doc` folder inside the program directory.
7. Click **Next** and the Export Results Wizard - Step 3 appears.



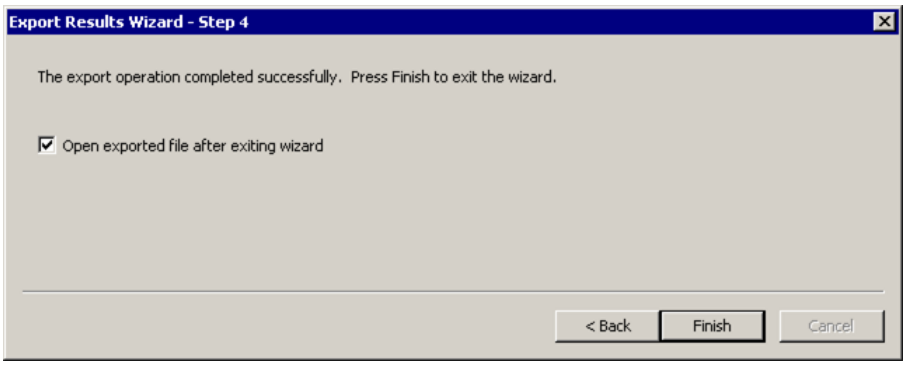
8. To use operations that apply to concepts other than LITI concepts see *SAS Content Categorization Studio: User's Guide*:
   - **Use full paths for the concept matches**: Display the concept name with its full path.

- **Include canonical coreference results for LITI concepts**: Display the canonical forms for matches generated by LITI co-reference rules.

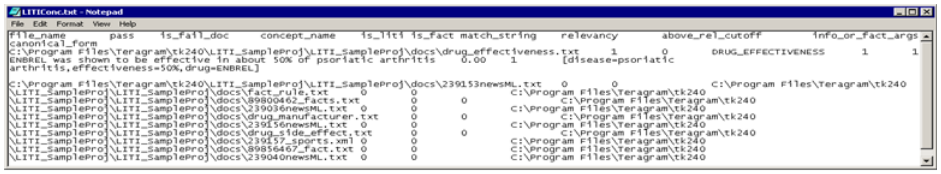
Hint: In this example, **Include header row** is also selected for clarity in reporting purposes.

---

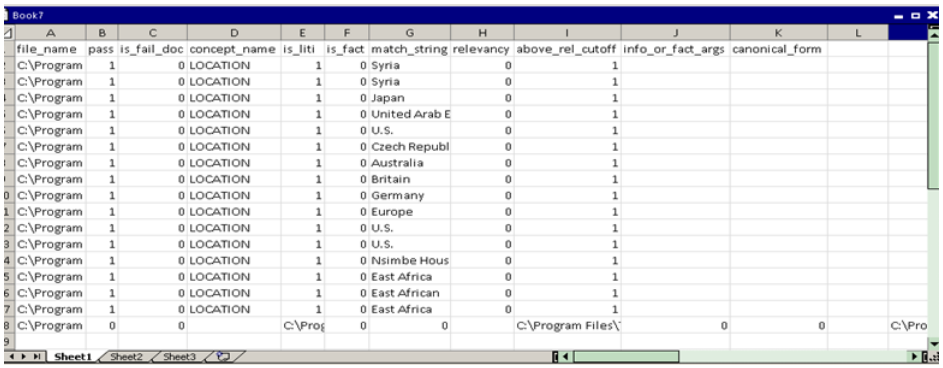
9. Click **Next** and the Export Results Wizard - Step 4 appears:



10. (Optional) Select **Open exported file after exiting wizard** to see the output immediately.



11. Open the file into a *Microsoft Excel* spreadsheet to see the results displayed in columns.



	A	B	C	D	E	F	G	H	I	J	K	L
	file_name	pass	is_fail_doc	concept_name	is_liti	is_fact	match_string	relevancy	above_rel_cutoff	info_or_fact_args	canonical_form	
	C:\Program	1	0	LOCATION	1	0	Syria	0	1			
	C:\Program	1	0	LOCATION	1	0	Syria	0	1			
	C:\Program	1	0	LOCATION	1	0	Japan	0	1			
	C:\Program	1	0	LOCATION	1	0	United Arab E	0	1			
	C:\Program	1	0	LOCATION	1	0	U.S.	0	1			
	C:\Program	1	0	LOCATION	1	0	Czech Republ	0	1			
	C:\Program	1	0	LOCATION	1	0	Australia	0	1			
	C:\Program	1	0	LOCATION	1	0	Britain	0	1			
0	C:\Program	1	0	LOCATION	1	0	Germany	0	1			
1	C:\Program	1	0	LOCATION	1	0	Europe	0	1			
2	C:\Program	1	0	LOCATION	1	0	U.S.	0	1			
3	C:\Program	1	0	LOCATION	1	0	U.S.	0	1			
4	C:\Program	1	0	LOCATION	1	0	Nsimbe Hous	0	1			
5	C:\Program	1	0	LOCATION	1	0	East Africa	0	1			
6	C:\Program	1	0	LOCATION	1	0	East African	0	1			
7	C:\Program	1	0	LOCATION	1	0	East Africa	0	1			
8	C:\Program	0	0		C:\Prog	0	0		C:\Program Files\	0	0	C:\Pro
9												

12. (Optional, if you selected **Include header row**.) See the results for the following headings:

Table 6-1: Column Headings for Exported Results

Heading	Description
file_name	The name of the file is listed here. (The full path to the concept is also displayed here, whether you select **Use full path for concept matches**.)
pass	1: if the file matched. 0: if the file did not match.
is_fail_doc	1: if the file is a document located in a Fail directory. 0: if the file is not located in the Fail directory. For more information, see *SAS Content Categorization Studio: User's Guide*.
concept_name	The name of the matched concept is listed here.
is_liti	1: if the matched concept is a LITI concept. 0: for concepts that are not LITI. For more information, see *SAS Content Categorization Studio: User's Guide*.

Table 6-1: Column Headings for Exported Results (Continued)

Heading	Description
is_fact	1: if a fact is located. 0: if there is no fact match. For more information, see Section 7.9 *Locating Facts* on page 187.
match_string	The string in the document that matches the concept definition is listed here.
relevancy	**Note**: At this time, relevancy is not computed.
above_rel_ cutoff	1: above relevancy cutoff. 0: otherwise. (The result shown under the **Above Relevancy Cutoff** heading in the Testing pane is not displayed here.) **Note**: At this time, relevancy is not computed.
date	(Optional): The date and time that the operation was performed is listed for each tested document. (For this reason, the date is the same for each displayed document and reflects the date and time that the export operation is performed.)
info_or_fact _args	**fact_args** specify the matching arguments for facts. For more information about facts, see Section 7.9 *Locating Facts* on page 187. (For **info_args** information, see SAS Content Categorization Studio: User's Guide.)
canonical_ form	See the referenced word. For more information, see Section 7.7.3 *The Operators for Coreference Resolution* on page 146.

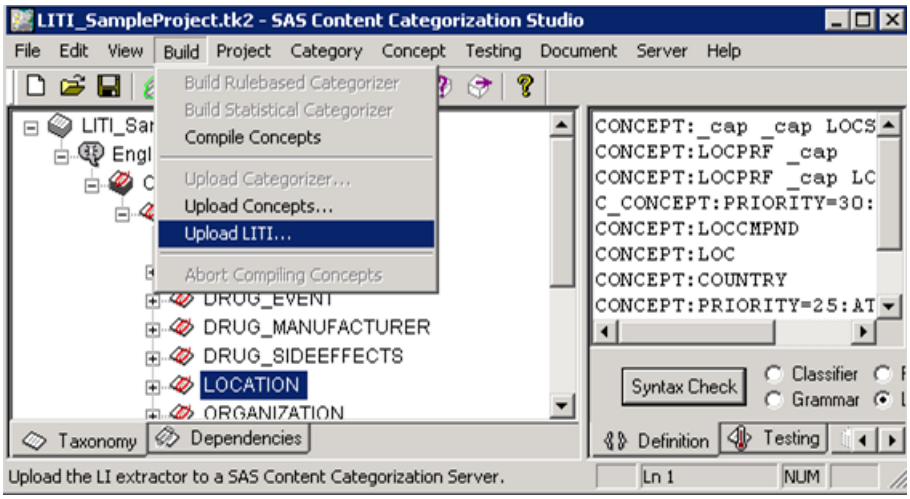
13. Click **x** to close *Notepad*.

6.12 The Upload LITI Operation in the Build Menu

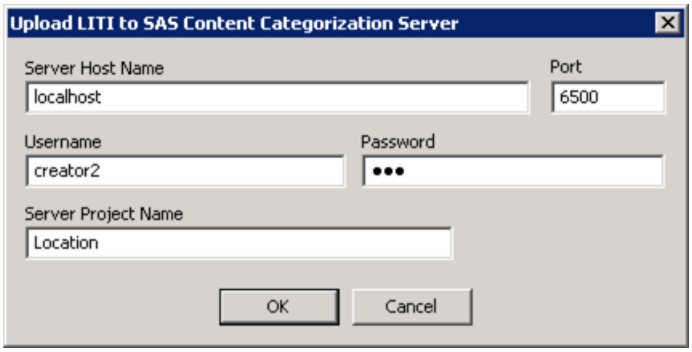
Use the Upload LITI operation to transfer a selected LITI definition to SAS Content Categorization Server where the rules are automatically applied to input documents.

To upload a LITI concept definition to SAS Content Categorization Server, complete these steps:

1. After you write a definition for a LITI concept, compile the concepts and save your project.
2. Highlight a LITI concept node in the Taxonomy pane.



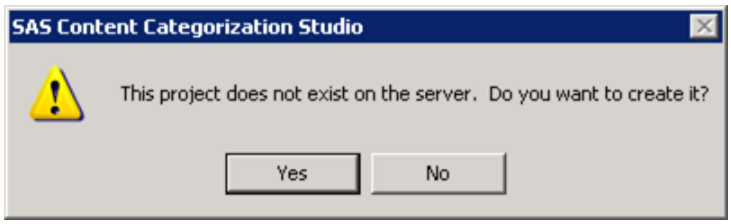
3. Go to **Build --> Upload LITI**. The Upload LITI to SAS Content Categorization Server window appears:



4. Enter the following information:

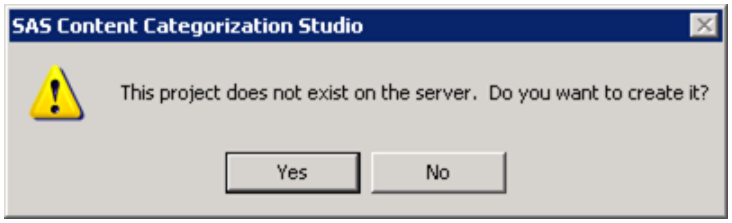
- **Server Host Name**: By default, this setting is specified. Change this specification if necessary. For example, type `localhost`.
- **Port**: By default, the query port that is specified during installation is specified. Change this specification if necessary. For example, type `6500`.
- **Username**: Enter your name as specified in the configuration file for SAS Content Categorization Server. For example, type `creator2`.
- **Password**: Enter your password as specified in the configuration file for SAS Content Categorization Server. For example, type `pw2`.
- **Server Project Name**: Enter the name of the project that you are uploading to the server. For example, type `Location`.
- Click **OK** to upload the selected concept.

If you are an administrator, creating a new project, the following SAS Content Categorization Studio window appears:

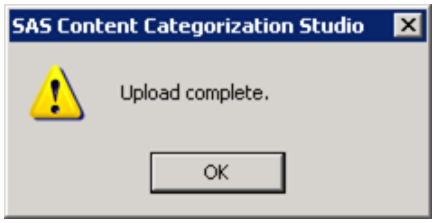


Click **OK** to create the project and close the SAS Content Categorization Studio window.

If you are not an administrator and you try to upload a new project, a SAS Content Categorization Studio window appears, stating `Invalid username or password`.



5. A SAS Content Categorization Studio window appears, stating `Upload Complete`, when this operation is complete.



6. Click **OK** to create the project and close the SAS Content Categorization Studio window.

6.13 Using the <language>.li File

When you upload your LITI concepts, SAS Content Categorization Studio creates a `<language>.li` file. This binary file includes the `<language>.concepts` file that is created for the classifier and grammar concepts in SAS Content Categorization Studio.

This feature enables you to reference SAS Content Categorization Studio classifier and grammar concepts when you write your definitions for LITI concepts. However, when grammar and classifier concepts are built, the resulting `<language>.concepts` file does not include a `<language>.li` file.

7
Writing Contextual Extraction Concept Definitions

7.1 Overview of Definitions

SAS Enterprise Content Categorization Studio uses linguistic technologies to expand the concept development capabilities in SAS Content Categorization Studio:

- Write a simple rule that matches one term specified in a list of entries.
- Locate a match for a unique concept where each individually specified term in the concept appears in one of the rules that together define this concept.

- Match a concept if the match appears within the specified context, only.
- Locate multiple partial matches and return them as full concept matches. These matches can occur only if there is a match on the fully defined concept within the input document.
- Write restrictive rules to prevent matches from occurring within specified contexts.
- Use predefined LITI concepts to simplify rule-writing. These concepts are available for Arabic, Chinese, Danish, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, Swedish, and Thai languages. However, these concepts might vary according to the language of your project.
- Disambiguate matches. Avoid possible matches on concepts that are specified using identical terms with different meanings.
- Specify part-of-speech tags to locate concepts.
- Use Boolean operators and various types of operators to increase the matching precision of your rule.
- Specify case-insensitive rule matches in the Data window in order to change the case-sensitive default setting.
- Use the stemming operator to return all of the forms of a word. Alternatively, choose to return only the noun or verb forms of the word.
- Specify coreference operators for pronoun resolution. In other words, when a pronoun or another word refers to the canonical form for a term, return the canonical form.
- Specify the canonical form that is used for pronoun resolution.
- Use the `PRIORITY` setting to specify that one rule is weighted more than another and to prevent the return of false positives for coreference matches. (In other words, you can rank one rule higher than another rule within the same definition.) You can also use the Concept Priorities window to reset the priorities.

Note: By default the **Priority** setting is set to `10` in the Data window for contextual extraction concepts. (However, this setting is applied to any SAS Content Categorization Studio concepts with a **Priority** setting

of less than `10`. This change is made when the project is uploaded as a binary file.)

- Match predicates by specifying multiple arguments to extract a fact.
- Identify the semantic relations between concepts by using predicate rules with logical operators.
- Specify XML fields to limit matches to these fields.
- Write XPath expressions into rules for greater flexibility in locating matches in valid `.xml` documents.
- You can write comments into your rules.

This chapter describes how to write the definitions that specify your concepts and provides examples of these rules. Before you write your rules, see the following sections. These sections provide information about prerequisite knowledge.

7.2 Create a Project

To create your concepts project, complete these steps:

1. Select **Start —> Programs —> SAS —> SAS Content Categorization Studio —> SAS Content Categorization Studio**.
2. Name the new project, specify a language, and enable concepts.
3. Right-click the `Top` node under the `Concepts` node in the Taxonomy window and select **Add Concept** from the drop-down list that appears.
4. After the new concept is added, enter the name of this concept into the box that appears to the right of the new node. Click the **LITI** radio button in order to specify its type.
5. Write the rule using the Definition window.
6. Click **Syntax Check** to check the rule.
7. Select **Build --> Compile Concepts** before you apply the rule to any input documents.

7.3 Before You Write Your Contextual Extraction Definitions

Consider the following information before you write your contextual extraction concept definitions:

- The terms *rule* and *definition* are used interchangeably. Properly speaking, definitions apply to all of the rules for one contextual extraction concept.
- Rule types are written using uppercase letters. For example, see `CLASSIFIER` and `C_CONCEPT`.
- Choose to use only uppercase letters to specify concept names. This practice makes it easy to identify concepts that are referenced within rules.
- Concept names can consist only of alphanumeric characters and underscores (_) and form a single word. Use all uppercase letters in a concept name to distinguish the concept from a term that you want to match. If you do not follow these requirements and refer to a concept from within a `PREDICATE_RULE` or a `CONCEPT_RULE`, matching might not occur on the referenced rule.
- By default, SAS Enterprise Content Categorization Studio performs case-sensitive matching.
- Matches are returned for contextual extraction concepts when matches also occur on classifier or grammar SAS Content Categorization Studio concepts. By default, the **Priority** setting in the Data window is set to `10` for all contextual extraction concepts. You can also specify a `PRIORITY` setting that overrides this setting within some contextual extraction rules, or choose to use the Concept Priorities window.

Note: When you define classifier and grammar SAS Content Categorization Studio concepts as well as LITI concepts, the priority for the classifier and grammar concepts changes to `10`. This behavior occurs when both types of concepts are extracted to the binary file.

- By default, matches can occur in any part of an input document. When the PARA or various SENT operators are specified, a match is returned if the matches occur in one paragraph, sentence, or the specified number of sentences.
- The settings in the Project Settings - LITI dialog box can affect match returns.
- The rules for predefined LITI concepts are not accessible. For this reason, test your rule to ensure that the rule matches are expected.

7.4 The Rule Types

7.4.1 Selecting a Rule Type

There are many types of contextual extraction concept rules. Unlike the classifier and grammar concepts in SAS Content Categorization Studio, you can specify more than one rule for each of your contextual extraction concept definitions. A match on the concept occurs if there is a match on any one of these rules.

`CLASSIFIER`

Specify lists of terms, like you do for classifier concepts in SAS Content Categorization Studio. However, in SAS Enterprise Content Categorization Studio, each Classifier rule consists of the word `CLASSIFIER` followed by a string. For more information, see Section 7.8.1 *The Classifier Rules* on page 147.

`CONCEPT`

Reference one or more concepts and use the `_cap` term to specify that a match only occurs on a word that begins with an uppercase letter. When more than one concept is referenced, a relationship is specified between the matching terms. You can also use `CONCEPT` rules to locate, or to discover, related information. For more information, see Section 7.8.2 *Specifying a Sequence of Classifier Entries* on page 149.

`C_CONCEPT`

Specify the order for the match components in an input document using these contextual extraction concept rules. For more information, see Section 7.8.3 *Context Matching* on page 152.

`NO_BREAK`

Prevent partial matches on a term that is specified within this definition. Use this rule to determine that an entire phrase is treated as a single word. For more information, see Section 7.8.5 *Eliminating Partial Matches* on page 156.

`REMOVE_ITEM`

Eliminate a false match in input documents where one word is a unique identifier for two concepts. This rule ensures that the correct context for the match is considered. For more information, see Section 7.8.6 *Disambiguating Matches* on page 158.

`REGEX`

Match information that follows a preset pattern. This rule uses the same syntax as the regular expression classifier concept definitions in SAS Content Categorization Studio. For more information, see Appendix B: *Using the Directive and Regex Syntax* and Section 7.8.10 *Specifying Regular Expressions* on page 168.

`CONCEPT_RULE`

Specify Boolean operators to increase precision (relevancy of the matches) and recall (return all matching texts). For more information, see Section 7.8.11 *Specifying a SENT Operator* on page 170.

`SEQUENCE`

Extract facts from input documents if the facts appear in the order specified. For more information, see Section 7.9.2 *A Predicate Sequence Example* on page 188.

`PREDICATE_RULE`

Specify the arguments that define your facts. Facts are related pieces of information in a text that are often located and matched as phrases. For more information, see Section 7.9.3 *The Predicate Examples* on page 191.

7.4.2 Adding a Rule Type to the Definition Pane

You can specify a rule type in the Definition window by placing your cursor at on the upper left corner. Hold Ctrl while you click the up arrow button on your keyboard to scroll through all of the rule types. See the following example:

Display 7-1 Adding a Rule Type to the Definition Window

Note: To remove a rule type, highlight a rule type and click either the Backspace or Delete key. At this time, perform this operation twice to remove the rule type.

7.5 The Rule Modifiers Table

The following table provides an overview of the various operators and modifiers that are available for the rules that you write. Use the links for each capability in order to see a more detailed explanation of each modifier. This table is provided as a quick reference.82

Table 7-1: Rule Modifiers

Modifier	CLASSIFIER	CONCEPT	C_CONCEPT	CONCEPT _RULE	REMOVE _ITEM	NO_ BREAK	SEQUENCE	PREDICATE _RULE	REGEX
Match specified strings	X	X	X	X	X	X	X	X	X
Comments	X	X	X	X	X	X	X	X	X

Table 7-1: Rule Modifiers (Continued)

Modifier	CLASSIFIER	CONCEPT	C_CONCEPT	CONCEPT_RULE	REMOVE_ITEM	NO_BREAK	SEQUENCE	PREDICATE_RULE	REGEX
arguments							required	required	
_c marker			required	required	required	required		X	
_w		X	X	X	X	X	X	X	
_cap		X	X	X	X	X	X	X	
> symbol			X	X					
@, @N, and @V		X	X	X	X	X	X	X	
Boolean Operators				X				X	
Part-of-speech tags		X	X	X	X	X	X	X	
export feature	X								
Coreference		X	X	X					
XML fields and XPath expressions	X	X	X	X	X	X	X	X	X
Regular expressions									required
intermediate concepts		X	X	X	X	X	X	X	
Predefined concepts		X	X	X	X	X	X	X	X
	Note: Pronoun resolution does not use any rule type.								

7.6 The Building Blocks

7.6.1 Overview of the Building Blocks

SAS provides `n-gram` sequence features that are often used in natural Language Processing (NLP). These sequences specify the context that is necessary for the specified concept to match. Before you write your contextual extraction concept rules, consider the building blocks that are explained in this section.

7.6.2 Case-Insensitive Matching

By default, SAS Enterprise Content Categorization Studio applies rules to input documents in a case-sensitive manner. You can specify case-insensitive matching when you click **Case Insensitive Matching** in the **Data** tab. This setting applies to the entire definition of the selected concept, only.

7.6.3 Entering Comments into Rules

Any character, or characters, following the pound sign (#) are considered to be comments. For a literal # to match, # should be escaped as \#. You can use comments with any rule types.

7.6.4 The Tokens

Add tokens to your definitions:

- words, including noise words such as *and*, *the*, and *a*
- numbers including date and time
- newline mark
- URLs

Specify an undetermined token using the `_w` term. When you specify this term, SAS Enterprise Content Categorization Studio returns a match on any word that occurs in this position in the document. If, on the other hand, there is an exact token that you want this contextual extraction concept to match, you can

specify this word in any concept rule. When tokens are specified in CONCEPT_RULEs and PREDICATE_RULEs, these tokens are set off with quotation marks (""). For more information, see Section 7.6.6 *The _w Term* below.

7.6.5 The _c Marker

Use the context marker (_c) to specify that a match is returned if the keyword is located within the specified context. For example, you can match any COMPANY concept that is immediately followed by the term *New York:*

```
C_CONCEPT:_c{COMPANY} New York
```

The context marker displays only the match on the term in the curly braces ({}) that follows the context marker when this term in found in the specified context. For example, see a match on COMPANY, but not on New York.

You can also use this marker to locate and return known and unknown words. See the following examples:

```
C_CONCEPT:COMPANY _c{New York}
C_CONCEPT:COMPANY _c{_cap}
```

Use only one context marker with the C_CONCEPT, CONCEPT_RULE, and the REMOVE_ITEM rule. This statement is true unless you specify an OR operator, which enables you to specify a context marker for each concept or term that you specify. Do not specify an argument such as _a or _b in place of a context marker.

Caution: **Use only one _c marker with each REMOVE_ITEM rule to prevent the application from closing.**

7.6.6 The _w Term

Use the word term (_w) to specify that a match can occur on a word. For example, you can match any type of business. This is true if _w immediately follows a reference to the COMPANYTYPE concept*:*

```
C_CONCEPT:_c{COMPANYTYPE} _w
```

This example could also return a match on law *firm*.

Hint: The _w term matches any single term. A term can consist of alphabetic or non-alphabetic characters. For example, see *today*, <, *Web*, *1.0*, and so on.

7.6.7 The _cap Term

Use the _cap term in ways that are similar to the _w term. However, _cap only returns matches on words that begin with an uppercase letter. Use _cap to locate an unknown term that begins with an uppercase letter, or to match a single upper case letter. Alternatively, specify this term multiple times. When you repeatedly specify _cap, you can locate all of the unknown, consecutive occurrences of words that begin with an uppercase letter. This term can be used with all of the contextual extraction rule types except for the CLASSIFIER and REGEX rules. You can also replace _w with _cap in the example provided for Section 7.6.6 *The _w Term* above. In this case, the word *Firm*, or another word beginning with an uppercase letter, is a match.

7.6.8 The > Symbol

Documents often reference a unique, full string only once. After this match these references might be made by one word from the original string. Use the greater than (>) symbol with either the C_CONCEPT, or CONCEPT_RULE, or a coreference operator (_ref). For more information about coreference, see Section 7.11.3 *How to Use the _ref Operator with the > Symbol* on page 206. Every occurrence of the bracketed term is a match if the entire rule is matched at least once in the input text.

Specify the greater than symbol within the C_CONCEPT rule using the _c{ }> syntax. For example, use this symbol to specify that every instance of the last name *Pelosi* should be returned as a match after the entire term *Ms. Nancy Pelosi* is located. See the following example where TITLE and FIRST refer to classifier concepts with a list of titles and first names, respectively:

```
C_CONCEPT:TITLE FIRST _c{_cap}>
```

7.6.9 The Quotation Marks

Use quotation marks (“”) to enclose tokens and concepts when writing a CONCEPT_RULE, REMOVE_ITEM, or PREDICATE_RULE. This example returns a match on *Mount Washington* if the term *Mount*, and a match on the concept NAME, appear within seven words of a match on the STATE concept:

```
CONCEPT_RULE:(DIST_7, “_c{Mount NAME}”, “STATE”)
```

7.6.10 The Parentheses, Square Braces, and Curly Braces

Use parentheses (()), square braces ([]), and curly braces ({}) as appropriate. These symbols qualify the matches for all of the contextual extraction definitions except the CLASSIFIER and CONCEPT types.

Use parentheses (()) to group the elements that comprise CONCEPT_RULE, REMOVE_ITEM, SEQUENCE, and PREDICATE_RULE definitions. For example, use parentheses with arguments and logical operators. Parentheses are also used with the AND, OR, SENT, DIST_n, ORDDIST_n, and ALIGNED Boolean operators. These operators are followed by a comma (,) and a space.

Use square braces ([]) to group REGEX rule elements with the Export operation. For more information, see Section 7.6.17 *The Export Feature* on page 137.

Use curly braces to delimit the information that is returned as a match. Curly braces ({}) are used with or without parentheses (()), according to the type of definition that is specified. For more information, see Section 7.9.2 *A Predicate Sequence Example* on page 188 and the following example:

```
CONCEPT_RULE:(SENT,  “_c{FIRST, _cap}”,
              “TITLE”, “COMPANY”)
```

7.6.11 The Commas

Commas (,) always follow definition elements:

- Boolean operators are enclosed in parentheses (()) and a space follows the comma (,) after this string.

- Quotation marks (“”) enclose concept names and a comma follows the second quotation mark.
- Separate the arguments used to construct facts with commas.
- Commas follow logical operators in a `PREDICATE_RULE`.

7.6.12 The Colon

Use a colon (`:`) in the following cases:

- Enter a colon after specifying the concept rule type. For example, use a colon with these rules `CONCEPT`, `CLASSIFIER`, and `CONCEPT_RULE`.
- Use a colon when specifying terms to export to `CLASSIFIER` rules. For more information, see Section 7.8.7 *Exporting Classifiers* on page 160.
- Use colons between arguments for a `SEQUENCE` or `PREDICATE_RULE` concept. For more information, see Section 7.9.2 *A Predicate Sequence Example* on page 188 and Section 7.9.3 *The Predicate Examples* on page 191.
- Enter a colon before a part-of-speech tag. For example, type `:Prep` and `:sep`. For more information, see Section 7.8.9 *Specifying Part-of-Speech Tags* on page 166.

7.6.13 The Spaces

When you write `CONCEPT`, `CONCEPT_RULE`, or `C_CONCEPT` definitions, type at least one space before each of the following items, tokens, concepts, part-of-speech tags, `_w` terms, and `_cap` terms. Also enter a space before the `_c` marker if it is preceded by a token, comma (`,`), or the name of a concept. See the following example:

```
CONCEPT_RULE:(ORDDIST_9, “_c{_cap} :sep _cap :sep and
              _cap”, “ORGTYPE”)
```

7.6.14 The Part-of-Speech Tags

Specify part-of-speech tags when you do not know the exact word that you are seeking. For example, `:Prep` to represent preposition and `:sep` to specify a separator character. A separator character is any punctuation mark. These part-

of-speech tags are preceded by a colon (:) and a space. See the following example:

```
CONCEPT_RULE:(SENT, "_c{VACATION :Prep _cap :sep
                LOCATION}", "vacation")
```

For a complete list of part-of-speech tags, see Appendix C: *Part-of-Speech Tags*.

7.6.15 The Predefined Concepts

Use predefined LITI concepts to quickly build a rule that references one, or more, concept rules. These predefined LITI concepts are shipped with the product and for this reason, the rules are unavailable for viewing or editing.

See the following types of predefined concepts:

Personal Pronoun Resolution

Specify along with a `CLASSIFIER` rule that specifies the name, or names, that are referenced by the pronouns that reference the appropriate noun.

Predefined Contextual Entities

Match entities in your documents based on the surrounding information.

Notes: Predefined Contextual Entities are not available for `CLASSIFIER` rules.
Test your documents frequently when adding predefined concepts to make sure that the match results are expected.

Some of these concepts are available for Arabic, Chinese, Danish, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, Swedish, and Thai languages.

7.6.16 The Intermediate Concepts

Intermediate concepts are concepts that have rules that are referenced in another rule. For example, see the VACATION and LOCATION concepts in the `CONCEPT_RULE` in Section 7.6.14 *The Part-of-Speech Tags* above.

7.6.17 The Export Feature

Export a matched term to one or more concepts. Use the `Export=` operation to define a term that matches a classifier concept. Also use the coreference operator (`_ref`) with the export symbol to eliminate false positives. You can specify this operation within the definition. Alternatively, declare an acronym as part of the definition for the concept where the selected term is exported. See the following example:

```
FULLNAME: CLASSIFIER: [export=eLN:Clinton}: Bill Clinton
LASTNAME: eLN
```

The term `Clinton` is exported to the LASTNAME concept. When you write the export operation into a Classifier rule, all instances of partial matches such as *Clinton*, are returned. For this reason, the export feature functions in ways that are similar to the effects of placing the greater than symbol (>) at the end of a rule. For more information, see Section 7.8.7 *Exporting Classifiers* on page 160.

7.6.18 The Regular Expressions

Match known patterns by using regular expressions to specify a range of letters or numbers inside square braces (`[]`). For example, type `a-z` or `0-9` within the square braces. This specification matches any word beginning with an ASCII character whose value is between a and z, or numbers between 0 and 9 inclusive. You can also add a plus (+) sign after the last square brace. See the following example:

```
REGEX:[a-z]+
```

When you add the plus sign, all instances of terms beginning with a lowercase letter match any and all occurrences of a word that appears in the input document. You can continue to build this definition by specifying a context for the word occurrence.

You can also add either the `%` symbol or write out `percent`, after these bracketed numbers. This feature enables you to locate percentage matches in your documents. See the following example:

```
REGEX:[0-9]+%
REGEX:[0-9]+ percent
```

This regular expression specifies that only numbers followed by the percentage sign match. For example, 99%, or 50 percent, are both matches.

For more information, see Appendix B: *Using the Directive and Regex Syntax*.

7.6.19 The Priorities and Project Settings

7.6.19.A Overview of Priorities

Priorities determine the concept that is matched when a match on more than one concept is located in an input document by SAS Content Categorization Server. You can see these matches in the text of a document using the Document pane. After the concept rules are uploaded to SAS Content Categorization Server as binary files, they are automatically applied to input documents.

For example, you might have a document that contains matches for both concept A and concept B. To prioritize a match on concept A, set the **Priority** setting in the **Data** tab for concept A to a higher number than that of concept B. Alternatively, you could specify `PRIORITY=n` in one or more rules in your definitions.

The `PRIORITY` rule specification that is set higher than `10`, overrides the **Priority** setting in the **Data** tab. (By default, the **Priority** setting in the **Data** tab is set to `10`.) For this reason, a `PRIORITY` setting in a rule also ranks overlapping rule matches in one concept definition as well as matches on different concept definitions. For more information, see Section 7.8.8 *Setting Priorities for Overlapping Matches* on page 163 and Section 7.11.7 *Rank Coreference Definitions and Eliminate False Positives* on page 211.

See the following example where `35` overrides the default **Priority** setting of `10` in the Data window:

```
C_CONCEPT:PRIORITY=35: _c{CITY COUNTRY}
```

You can also choose to use the Concept Priorities window to reset the priorities for your concepts. When you reset a value in this window, this number is applied to any matches on the rule. However, if you specify a priority within a rule, the larger of the specifications is applied to the match.

Note: When you upload SAS Content Categorization Studio concepts to SAS Content Categorization Server, the default priority settings are also set to `10` for the SAS Content Categorization Studio concepts.

7.6.19.B Choose Project Settings

Use the **LITI** tab in the Project Settings window to choose the types of matches that you want to return. These settings are particularly important when you specify priorities and when multiple matches occur within one input document.

To specify Project Settings, complete these steps:

1. Select **Project --> Settings** and the Project Settings window appears.

2. Leave the default setting **All matches** to return matches on all of the matching rules in an input document. Alternatively,
   - Select **Longest** to return only the match with the most characters.
   - Select **Best** to return only the best match.
3. Select **Return all identical matches** when you want to locate each instance of a rule match.
4. If you specified either a `PREDICATE` or a `SEQUENCE` rule, you can select **Remove duplicate facts** to return the first instance of a match, only.

7.6.19.C Choosing Priorities and Project Settings

Selecting project settings is an important consideration when you specify priorities. For more information and an example, see Section 7.8.8 *Setting Priorities for Overlapping Matches* on page 163.

7.7 The Operators

7.7.1 The Boolean Operators

To locate related information with greater precision, specify Boolean, or logical operators, with some types of contextual extraction rules. These operators are required for `CONCEPT_RULE`s and `PREDICATE_RULE`s.

Table 7-2: Boolean Operators

Operator	Description
`ALIGNED`	Disambiguate between matches on two contextual extraction concept rules. Disambiguation enables SAS Enterprise Content Categorization Studio to determine the correct match based on context. When terms are disambiguated, only the match is returned.
`AND`	Specify that a match can occur only when both arguments are present, somewhere within the entire document.
`NOT`	Specify a `NOT` operator to preclude a match when the match that is specified by the `AND` operator also occurs in the input document.
`OR`	Specify that a match is returned if one, but not both, of the concepts or tokens is located.
`ORD`	Specify the order for a match. If matched instances are located out of order, a match is not returned for the document.
`DIST_n`	Specify the number of words between matches on rule terms. The first match takes the starting position 1. The last match falls at or before the specified number of words.
`ORDDIST_n`	Specify the maximum word count between arguments. Otherwise, this operator functions like the `DIST` operator above.

Table 7-2: Boolean Operators (Continued)

Operator	Description
SENT	Specify a sentence delimiter. For example, type ., ?, or !. A match is returned when all of the specified components are located in the sentence where the first match occurs.
SENT_n	Specify a sentence delimiter that returns matches on multiple sentences.
SENTSTART_n	Specify that matches are returned within n words from the start of the sentence. **Note**: Start the count from zero (0).
SENTEND_n	Specify that matches are returned within n words from the end of the sentence **Note**: Start the count from zero (0).
PARA	Match only within a paragraph.
UNLESS	Restrict a match on another Boolean operator.

Specify a comma (,) and a space after a Boolean operator and enclose it in parentheses (()). For example, write (SENT, “NAME”).

7.7.1.A The ALIGNED Operator

Specify the ALIGNED operator to refer to a term that matches two concepts within one rule. The presence of this operator enables SAS Enterprise Content Categorization Studio to determine what concept is an exact match for this term.

For example, the following rule specifies that if a term matches both the LOC and PERSON concepts, only a match for the PERSON concept is returned. Matches for the LOC concept, such as Washington, are returned as a match on the PERSON concept:

```
REMOVE_ITEM:ALIGNED, (“_c{LOC}”, “PERSON”)
```

7.7.1.B The AND Operator

Specify the AND operator for two or more arguments. A match occurs only if both arguments are present. For example, the following rule limits matches to Bills in documents where the word football also occurs:

```
CONCEPT_RULE:(AND, “_c({Bills}”, “football”)
```

7.7.1.C The NOT Operator

The NOT operator, unlike other LITI operators, is applied to the entire document. For this reason, specify the NOT operator with the required AND operator. Use the NOT operator with a CONCEPT_RULE or with a PREDICATE_RULE. When a match is located for both the NOT and the AND operators, as specified within a rule, a match is returned for the AND operator.

The NOT rule helps ensure contextual accuracy for a match. For example, the following rule limits matches on Amazon to documents where the word river does not occur:

```
CONCEPT_RULE:(AND, “_c{Amazon}”, (NOT, “river”))
```

Hint: To prevent matches on both *Amazon river* and *Amazon rainforest*, write two rules for the selected definition. You can also specify intermediate concepts and use the intermediate concepts to reference all of the terms.

For more information, see Section 7.8.15 *Specifying the NOT Operator with the AND Operator* on page 181.

7.7.1.D The OR Operator

When you specify the OR operator, a match occurs for an input document if at least one of these arguments is located. For example, the following rule matches if either the token Barack or Obama is present in the text:

```
CONCEPT_RULE:(OR, “_c{Barack}”, “_c{Obama}”)
```

Hint: When you specify the OR operator in a rule, with the exception of the REMOVE_TIME rule, you can specify more than one instance of _c.

7.7.1.E The ORD Operator

Specify the `ORD` operator to specify the matching order for two or more matched rule components. A match is returned only when the specified rule components are matched in the specified order. For example, the following rule matches if the tokens `dump` and `truck` are present in the text in this order:

```
CONCEPT_RULE:(ORD, “_c{dump}”, “truck”)
```

7.7.1.F The DIST_n Operator

Specify the maximum distance, in words, between located terms in order for a match to be returned for the selected concept. For example, if you want to specify that a match on the FULLNAME concept that appears within eight words of *Harvard University* is a match for the concept, write:

```
CONCEPT_RULE:(DIST_8, “_c{FULLNAME}”,
              “Harvard University”)
```

7.7.1.G The ORDDIST_n Operator

Specify the order and distance between the terms or concepts that you want the selected concept to match. This operation locates and returns a match even when the usual contextual clues provided by adjacent matches are missing. For example, a match can be located when a name and position do not follow one another. The following example returns a match on the POSITION concept when it is followed by the word `Obama`. This is true only if the term `Obama` is located within `12` words from a match on the POSITION concept.

```
CONCEPT_RULE:(ORDIST_12, “_c{POSITION}”, “Obama”)
```

7.7.1.H The SENT Operator

Locate matches in the same sentence. For example, write a concept definition that locates a match for the term *Amazon* when the token *river* also occurs within the same sentence:

```
CONCEPT_RULE:(SENT, “_c{Amazon}”, “river”)
```

7.7.1.I The SENT_n Operator

Locate matches that occur in the specified number of sentences. For example, write a concept definition that locates matches for the PER concept and the term *he* within two sentences:

```
PER concept:  CLASSIFIER:Obama

CONCEPT_RULE:(SENT_2, “_c{PER}”, “he”})
```

7.7.1.J The SENTSTART_n Operator

Locate matches that occur within the specified number of words from the beginning of the sentence. For example, write a concept definition that locates matches for the term *Democratic* that occur within five words from the start of the sentence:

```
CONCEPT_RULE:(SENTSTART_5, “Democratic”)
```

Note: Start the count from zero (0).

7.7.1.K The SENTEND_n Operator

Locate matches that occur within the specified number of words from the end of the sentence. For example, write a definition that locates matches on a term in the PER concept if these matches occur within five words from the end of a sentence. The following example shows how the SENT_n, SENTSTART_n, and SENTEND_n qualifiers work together with a contextual operator and a classifier concept:

```
PER concept:  CLASSIFIER:Obama

CONCEPT_RULE:(SENT_2, (SENTSTART_5, “Democratic”),
            (SENTEND_5, “_c{PER}”))
```

Note: Start the count from zero (0).

7.7.1.L The PARA Operator

When you add the paragraph (PARA) operator, you specify that matches are located only within one paragraph. Determine the paragraph boundaries by

typing one or more separator characters into the **Paragraph Separator** field in the **Project Settings - Misc** tab. When you specify more than one type of paragraph separator, use a comma (,) to identify each string as a paragraph separator. For example, you can enter the following strings to specify the paragraph separators `\n\n,\t\t,<P>`.

You can then write one of the following rules that specify that matches can be located only in the text bounded by one or more of these separators:

```
CONCEPT_RULE: (PARA, "_c{SAS}", (OR, "statistics", "TM"))
CONCEPT_RULE: (PARA, "_c{TM}", (OR, "Enterprise Miner"))
```

7.7.1.M The UNLESS Operator

A match is returned for the `UNLESS` operator when the match that it specifies is not located within a specified window, or context. Use the `UNLESS` operator in a `CONCEPT_RULE` or in a `PREDICATE_RULE`, only. The `UNLESS` operator requires two parameters that set the boundaries for the match. These parameters, which are preceded by a second Boolean operator, establish the window in which the match on the `UNLESS` operator can be located. The second Boolean operator takes two arguments.

Note: When concepts are specified in a rule that uses the `UNLESS` operator, specify concepts that contain only `CLASSIFIER` or `REGEX` rules.

The `UNLESS` operator matches only when the quoted argument for the `UNLESS` operator is not found between these two arguments. If the match is not located within this window, a match is returned for this rule. For example, specify the `UNLESS` operator in order to return matches for the Mississippi entities that do not include the word *river* between the terms *Mississippi* and *United States*.

```
CONCEPT_RULE:(UNLESS, “river”, (SENT,
              “_c{Mississippi}”, “United States”))
```

Specify only the following Boolean operators with the `UNLESS` operator: `AND`, `SENT`, `DIST`, `ORD`, and `ORDDIST`.

If you want to specify that a match cannot occur when another match occurs anywhere in the entire document use the `NOT` operator. For more information, see Section 7.7.1.C *The NOT Operator* on page 142.)

Example 7-1: A CONCEPT_RULE with an UNLESS Operator

Concept Name	Entry
TELEVISION_SERIES	CONCEPT_RULE:(UNLESS, “SHOW”, (SENT, “_c{SHOW}”, “SEASON”))
SHOW	CLASSIFIER:The Office
SEASON	CLASSIFIER:fall
	CLASSIFIER:spring

For more information, see Section 7.8.16 *Specifying the UNLESS Operator* on page 185.

7.7.2 The Stemming Operator

When you add an @ symbol as a suffix to a word, you enable the expansion of the word into all of its forms. For example, if you append an @ sign to the word *book*, matches on books, booking, bookings, and so on, could be returned:

```
CONCEPT:book@
```

You can also append the @ sign followed by the letter N or the letter V to stem the word into all of its noun or verb forms, respectively. See this example:

```
CONCEPT_RULE:(SENT, "_c{book@N}", "train@V")
```

Note: The @ symbol cannot be used in CLASSIFIER and REGEX definitions.

7.7.3 The Operators for Coreference Resolution

Coreference resolution enables you to match pronouns and other words to the canonical forms that these terms reference. (This is also known as *anaphora resolution*.) When you use coreference resolution, you can specify the canonical form of the referencing word. For example, specify *Barack*, *Obama*, and *President* as referring terms for the canonical form *Barack Obama*. Alternatively, choose to make *President Barack Obama* the canonical form for these terms.

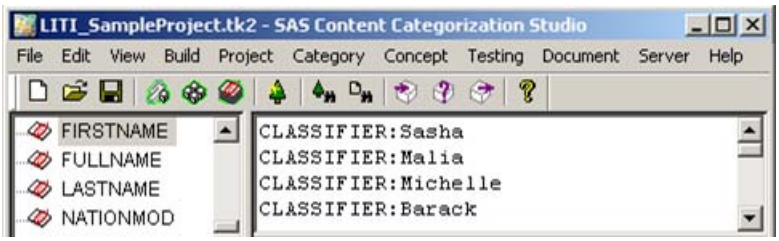
For more information about coreference operators, see Section 7.11 *The Coreference Operators* on page 204.

7.8 Contextual Extraction Concept Rule Examples

7.8.1 The Classifier Rules

Specify a `CLASSIFIER` rule to match one string, or dictionary entry. Like classifier definitions in SAS Content Categorization Studio, these rules specify a string to match in an incoming document. Unlike classifier concepts, each `CLASSIFIER` line is one `CLASSIFIER` rule.

Display 7-1 Classifier Rules



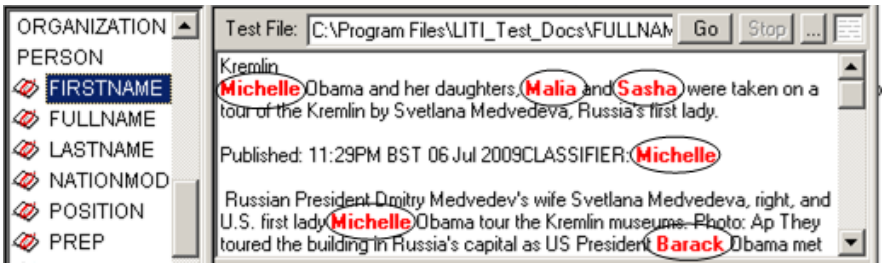
This FIRSTNAME concept consists of several `CLASSIFIER` rules.

Example 7-2: Matching a Sequence of Dictionary Entries

Concept Name	Entry
FIRSTNAME	CLASSIFIER:Sasha
	CLASSIFIER:Malia
	CLASSIFIER:Michelle
	CLASSIFIER:Barack

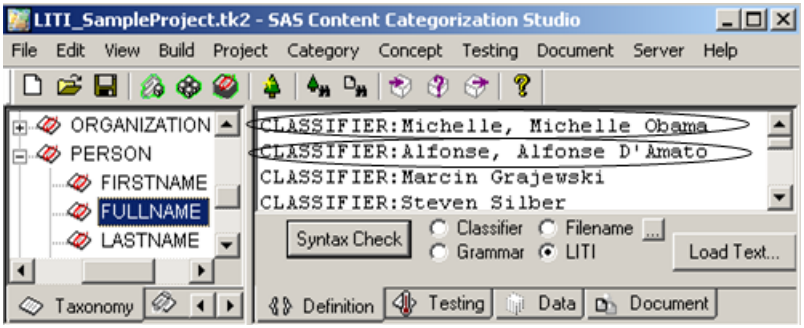
The FIRSTNAME concept matches any of the names to the right of the CLASSIFIER specifications in incoming texts. For example, a match occurs on any instance of Sasha, Malia. Michelle, or Barack.

Figure 7-1 FIRSTNAME matches in an Input Document



You can also specify a returned information string after a comma (,). In this case the returned information is the value for the matched concept. See the following example:

Figure 7-2 CLASSIFIER Rules with INFO Strings



This FULLNAME concept consists of the following CLASSIFIER rules.

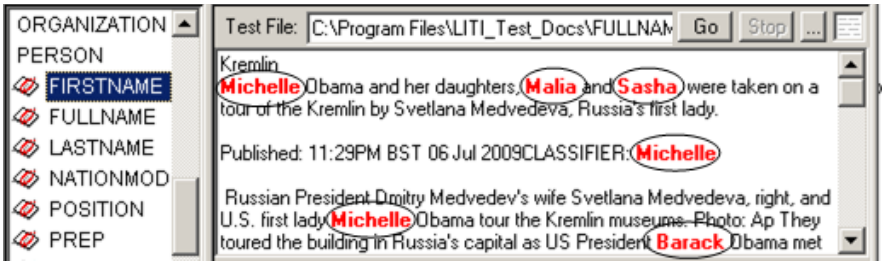
Example 7-3: Matching CLASSIFIER Rules

Concept Name	Entry
FULLNAME	CLASSIFIER:Michelle, Michelle Obama
	CLASSIFIER:Alfonse, Alfonse D'Amato

The comma character (,) is reserved for use as a separator character between the match_key entry and returned_information entry. The comma follows the match_key entry regardless of whether returned_information is specified. You can choose to match a comma using allowed characters for either the match_key or the returned_information string.

The FULLNAME concept matches any of the names in the match_key entry of the CLASSIFIER definition in incoming texts. The INFO string that is returned is the returned_information entry. For example, Alfonse D'Amato is returned for a match on Alfonse.

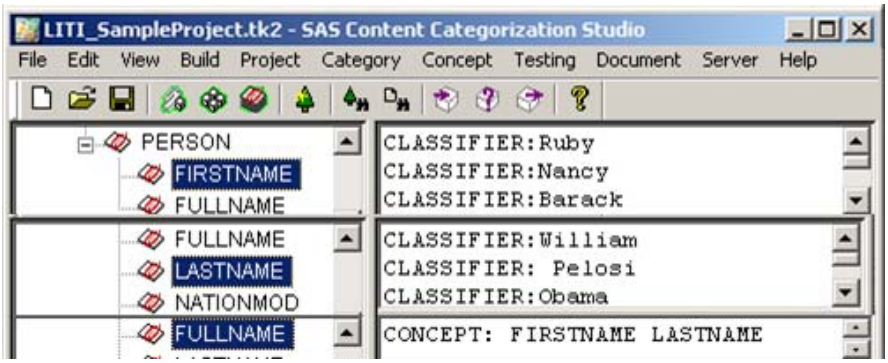
Figure 7-1 FULLNAME matches in an Input Document



7.8.2 Specifying a Sequence of Classifier Entries

Write a CONCEPT rule to identify related information, whether these relationships are known beforehand. For example, you might want to identify all of the lakes in the state of Michigan, but not know the names of these lakes when you write the rule. The CONCEPT definition specifies the ordering of CLASSIFIER concepts. A match occurs when matching CLASSIFIER strings are located in the specified order in an input document.

Figure 7-2 CONCEPT Rule



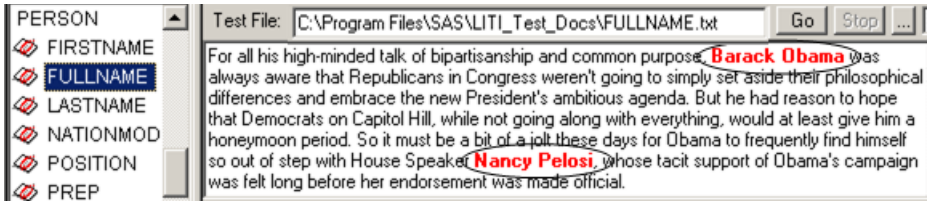
This FULLNAME concept defines a relationship between the FIRSTNAME and LASTNAME concepts.

Example 7-4: Matching a Sequence of Dictionary Entries

Concept Name	**Entry**
FIRSTNAME	CLASSIFIER:Ruby
	CLASSIFIER:Nancy
	CLASSIFIER:Barack
LASTNAME	CLASSIFIER:William
	CLASSIFIER:Pelosi
	CLASSIFIER:Obama
FULLNAME	CONCEPT: FIRSTNAME LASTNAME

The FULLNAME concept uses the lists of terms that are specified by the `CLASSIFIER` definitions in the FIRSTNAME and LASTNAME concepts. A relationship between matches on the FIRSTNAME and LASTNAME concepts is specified by the FULLNAME concept. For example, the terms *Nancy Pelosi* and *Barack Obama* match both the FIRSTNAME and the LASTNAME concepts. These matches are also a match for the FULLNAME concept rule.

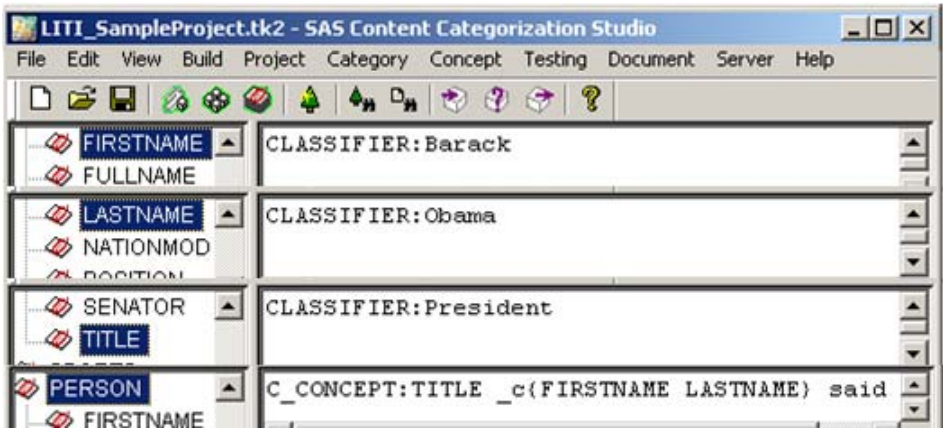
Figure 7-3 FULLNAME Concept Matches in an Input Document



7.8.3 Context Matching

Write a C_CONCEPT rule to match text in an input document based on the context of the matches. You can also use tokens with C_CONCEPT rules.

Figure 7-4 C_CONCEPT Rule



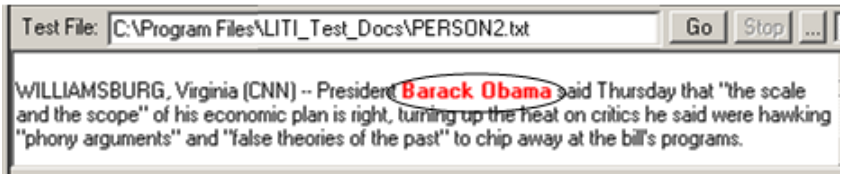
This C_CONCEPT rule specifies a relationship between the CLASSIFIER concept rules and the token said.

Example 7-5: Matching within Context

Concept Name	Entry
FIRSTNAME	CLASSIFIER:Barack
LASTNAME	CLASSIFIER:Obama
TITLE	CLASSIFIER:President
PERSON	C_CONCEPT:TITLE _c{FIRSTNAME LASTNAME} said

The PERSON concept locates matches for the FIRSTNAME and LASTNAME concepts. These matches occur in the context (_c) specified by the curly braces ({}) preceded by a match on the TITLE concept and followed by the token *said*. In this example, *Barack Obama* matches on the PERSON concept.

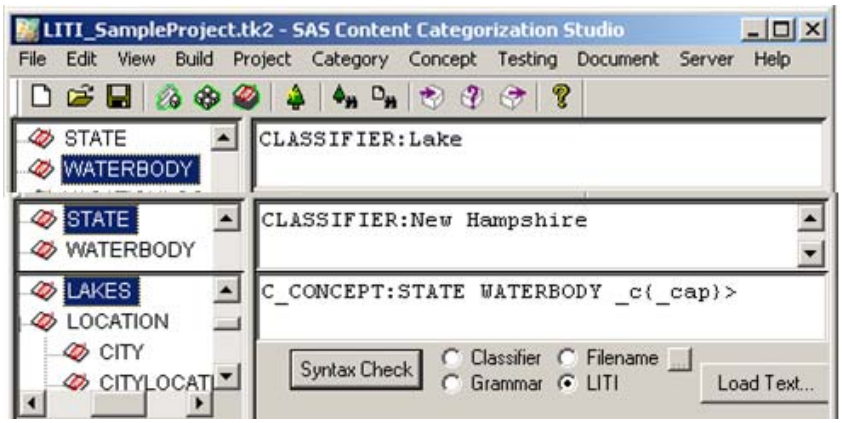
Figure 7-5 A C_CONCEPT Match in an Input Document



7.8.4 Matching within Context

Write a C_CONCEPT definition to locate and match a word that you do not know until a match on this definition is located. However, you should know the context for this match. For example, you might want to locate, and return each duplicate instance of *New Hampshire lake* in an input text.

Figure 7-6 C_CONCEPT Rule



This C_CONCEPT definition specifies a relationship between matching concepts and a word beginning with an uppercase letter.

Example 7-6: Using a Reference for a Match

Concept Name	Entry
WATERBODY	CLASSIFIER:Lake
STATE	CLASSIFIER:New Hampshire
LAKES	C_CONCEPT:STATE WATERBODY _c{_cap}>

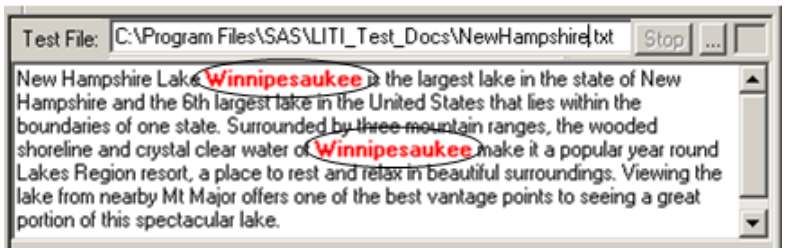
The LAKES concept specifies the context for the matched terms:

- When a match on the STATE concept is followed by a match on the WATERBODY concept, a partial match is located. For example, *New Hampshire Lake* is a partial match for this rule.
- _c(_cap) specifies that the matches above also appear in the context of a word that begins with an uppercase letter. In this example, a match occurs on the word *Winnipesaukee*.

- By default, all of the matches in an input document are returned. When the greater than (>) symbol is specified, every instance of the matched term in the document is returned as a match regardless of the context.

In this example, all instances of *Winnipesaukee* are matched. The second match occurs because the greater than (>) symbol is specified.

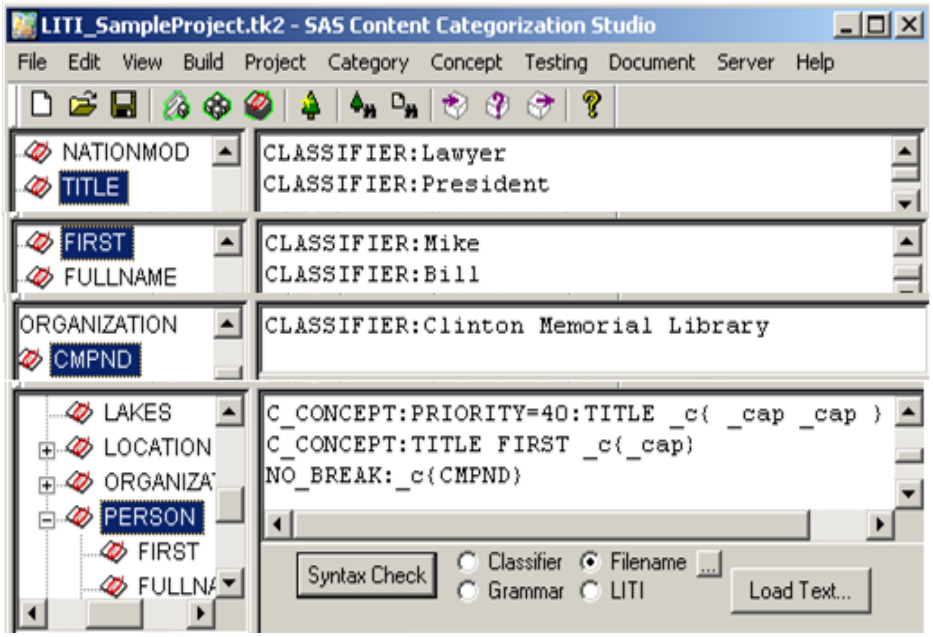
Figure 7-7 C_CONCEPT Matches in an Input Document



7.8.5 Eliminating Partial Matches

Specify a NO_BREAK rule to prevent partial matches on terms. This rule stipulates that a match can occur only if the entire string is located in an input document. This statement is true for any rules that might otherwise locate a partial match.

Figure 7-8 NO_BREAK Rule



The PERSON concept specifies the NO_BREAK rule.

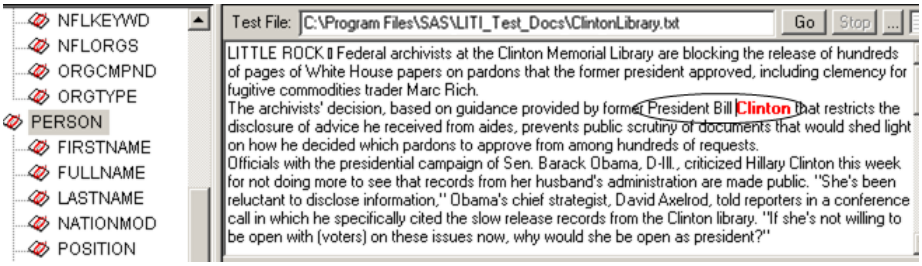
Example 7-7: Excluding Spaces

Concept Name	Entry
TITLE	CLASSIFIER:President
FIRST	CLASSIFIER:Bill
CMPND	CLASSIFIER:Clinton Memorial Library

```
PERSON                  C_CONCEPT:TITLE FIRST _c{_cap}
                        NO_BREAK:_c{CMPND}
```

When you add the `NO_BREAK` rule to the PERSON concept definition, the token `Clinton` is not matched when it occurs in the phrase *Clinton Memorial Library*. For this reason, matches are not returned for any definition that matches part, but not all, of this term.

Figure 7-9 NO_BREAK Rule Match in an Input Document

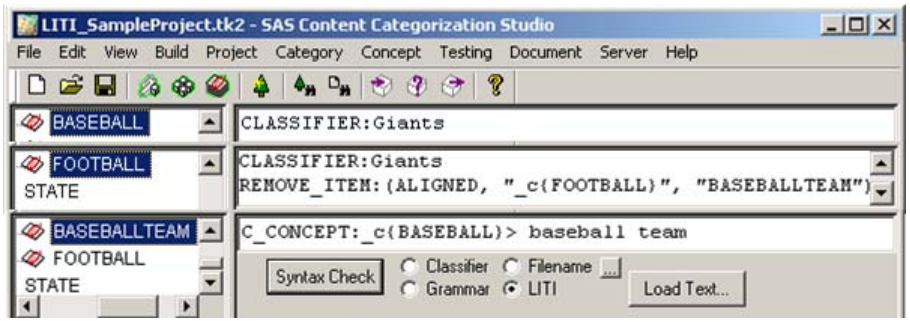


7.8.6 Disambiguating Matches

REMOVE_ITEM definitions differentiate between matches according to their context. This process of differentiation is called *disambiguation*. In SAS Content Categorization Studio disambiguation is specified in a Boolean definition using the __TGIF or __TGUNLESS operator. SAS Enterprise Content Categorization Studio enables you to specify this rule type when you refer to other concepts by writing a REMOVE_ITEM rule. Use this operation to eliminate a match on one rule, while returning a match on another rule.

Hint: If you want to prevent a match within a specified context, write a rule using the UNLESS operator. For more information, see Section *Use the UNLESS operator to set a constraint on possible matches. The UNLESS operator is often used to ensure that an attribute is correctly assigned to the appropriate subject. For this reason the UNLESS operator sets boundary constraints.* on page 185.

Figure 7-10 REMOVE ITEM Rule



The FOOTBALL concept definition includes the REMOVE_ITEM rule to prevent Giants football documents from matching the Giants baseball concept.

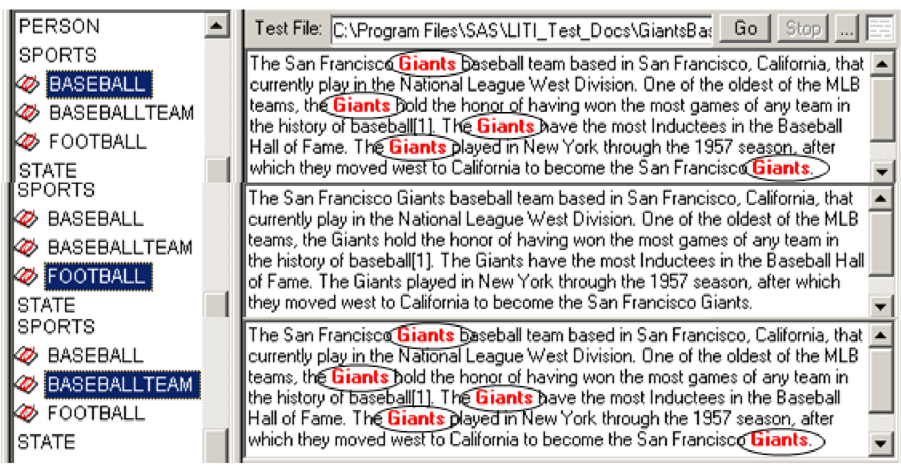
Caution: Use only one _c marker with each REMOVE_ITEM rule to prevent the application from closing.

Example 7-8: Excluding Phrases

Concept Name	**Entry**
`BASEBALL`	`CLASSIFIER:Giants`
`FOOTBALL`	`CLASSIFIER:Giants` `REMOVE_ITEM:(ALIGNED, “_c{FOOTBALL}”, “BASEBALLTEAM”)`
`BASEBALLTEAM`	`C_CONCEPT:_c{BASEBALL} baseball team`

Matches on the word *Giants* are returned for the BASEBALLTEAM concept when the token *Giants* is located in the specified context, Giants baseball team. In this case, this match is not a match for the FOOTBALL concept. The `REMOVE_ITEM` rule specifies that any match on both the BASEBALLTEAM and the FOOTBALL concepts return only matches for the BASEBALLTEAM concept.

Figure 7-11 Disambiguated Matches in Input Documents



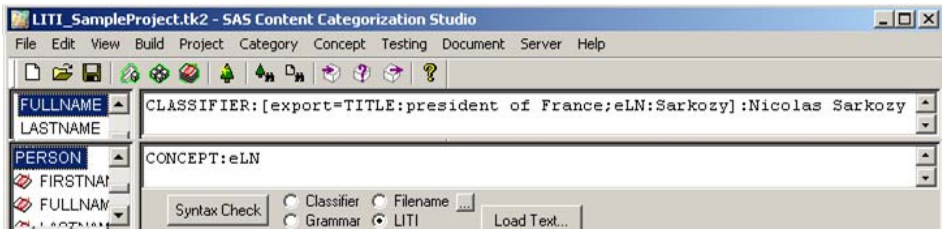
7.8.7 Exporting Classifiers

The CONCEPT rule enables you to export previously unspecified classifier terms to another concept using an acronym that is specified in a concept rule. For example, specify eLN for last name. Alternatively, you can enter the full name of the concept such as LASTNAME.

To write a rule using an acronym, specify this acronym in the destination rule. After an acronym is specified in a CONCEPT rule, other rules can specify this acronym to list the exported term.

The CLASSIFIER rule that specifies the export feature enables you to match incomplete terms in ways that are similar to that of the greater than symbol. For more information, see Section 7.6.8 *The > Symbol* on page 133. However, you can use only the export operation with CLASSIFIER rules.

Figure 7-12 CLASSIFIER Rule with Export Feature



The FULLNAME concept specifies a CLASSIFIER rule that exports matches on Sarkozy to the PERSON concept that has a CONCEPT rule specifying eLN. This rule also specifies its own matching string and the context for matches.

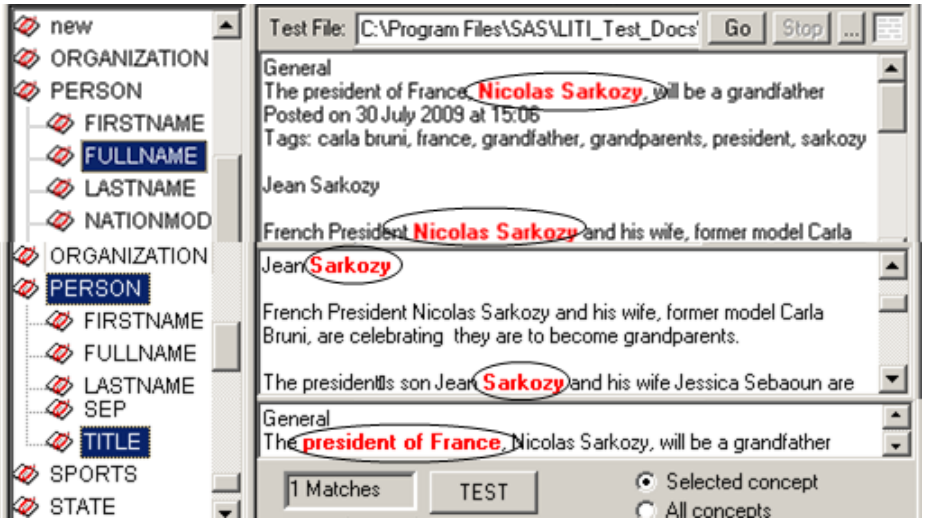
Example 7-9: Exporting Classifiers Example 1

Concept Name	Entry
FULLNAME	CLASSIFIER:[export=TITLE:president of France; eLN:Sarkozy]:Nicolas Sarkozy
PERSON	CONCEPT:eLN

The following matches occur in an input text that has the words *Nicolas Sarkozy* and *President of France* present somewhere in the same document:

- *President of France* is exported to, and matches, the TITLE concept.
- *Sarkozy* matches the PERSON concept. This match occurs because the acronym `eLN` is specified in the PERSON concept.
- *Nicolas Sarkozy* is returned as the match for the FULLNAME concept.

Figure 7-13 Classifier and Export Matches in Input Documents



The export feature works on an internal, per-document basis. In this example, the terms *President of France* and *Sarkozy* only match the TITLE and PERSON concepts if *Nicolas Sarkozy* is present in the input document. The exported terms do not appear in the concept definitions when these terms are exported. The concepts do not have to exist in the taxonomy in order for the export rule to work.

Display 7-2 CLASSIFIER Rule with Export Feature

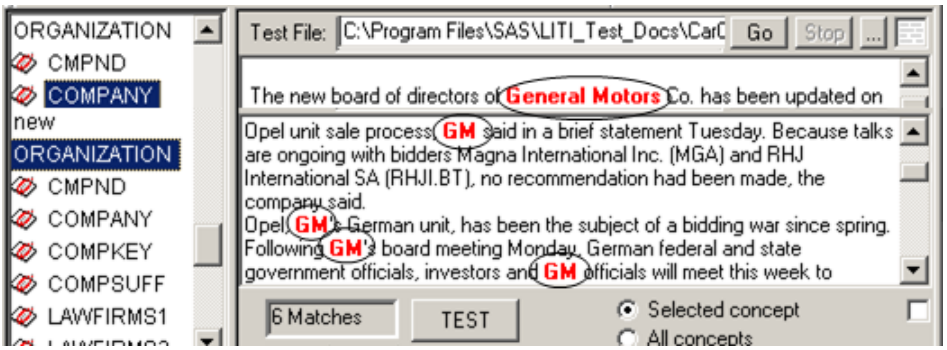
The COMPANY concept specifies that a match on `GM` is exported to the ORGANIZATION concept.

Example 7-10: Exporting Classifiers Example 2

```
COMPANY                 CLASSIFIER:[export=ORGANIZATION: GM]:
                        General Motors
```

If an input text contains the string *General Motors*, the document matches the COMPANY concept. If this document also contains the word *GM*, the token *GM* is recognized as a match on the ORGANIZATION concept. However, if the word *GM* appears in a document without the term *General Motors*, *GM* is not returned as a match to the ORGANIZATION concept.

Figure 7-14 Export Rule Matches in Input Documents



7.8.8 Setting Priorities for Overlapping Matches

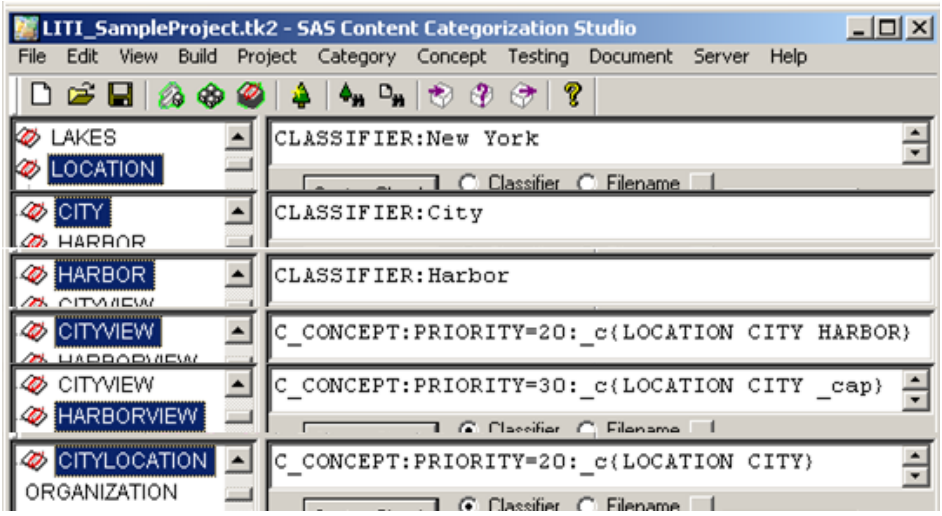
SAS Enterprise Content Categorization Studio enables you to override the **Priority** setting in the Data window for the selected contextual extraction concept. This feature works with `CONCEPT_RULE` definitions and coreference rules when you write a `PRIORITY` specification into a rule. For more information about coreference, see Section 7.11.7 *Rank Coreference Definitions and Eliminate False Positives* on page 211.

To use this feature, select **Best Matches** in the **LITI** tab of the Project Settings window. By default, the **Priority** is set to `10` in the Data window for contextual extraction concepts. (However, this setting is applied to any SAS Content Categorization Studio concepts that you write when you upload a contextual extraction project as a binary file.)

You can also increase the **Priority** setting in the Data window for all of the rules in one definition, or specify a `PRIORITY` in a contextual extraction concept rule. When you specify a `PRIORITY` in a rule, this setting overrides the **Priority** setting in the Data window—for this rule only. The `PRIORITY` specification in a rule applies to the rule, and not to the entire definition. For this reason, any matches on this rule are prioritized over matches on any other rules in this definition, or in any other definitions

These specifications are used to increase the relative rankings between contextual extraction concepts. Priorities are also used to prevent matches on more than one concept. You can also use this setting to prevent matches on terms that are used in different contexts. For example, if `Roche` is specified in the PERSON concept and also in the CORPORATE concept, priorities can be used to determine the appropriate match.

Figure 7-15 C_CONCEPT Rule Specifying a Priority Setting



The HARBORVIEW concept has the highest `PRIORITY` setting. Documents that match this concept, and any of the other concepts shown, are matched to the HARBORVIEW concept.

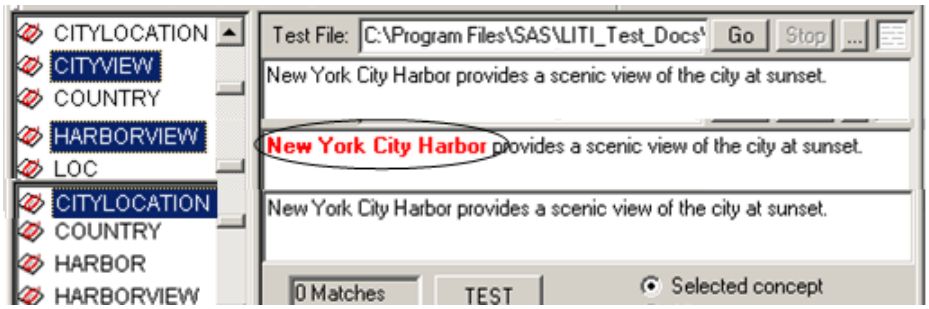
Example 7-11: Setting Priorities

Concept Name	Entry
LOCATION	CLASSIFIER:New York
CITY	CLASSIFIER:City
HARBOR	CLASSIFIER:Harbor
CITYVIEW	C_CONCEPT:PRIORITY=20:_c{LOCATION CITY HARBOR}
HARBORVIEW	C_CONCEPT:PRIORITY=30:_c{LOCATION CITY _cap}
CITYLOCATION	C_CONCEPT:PRIORITY=25:_c{LOCATION CITY}

The following document is returned as a match to the HARBORVIEW concept. This is true even though *New York City Harbor* also matches the CITYVIEW concept and part of this term matches CITYLOCATION.

New York City Harbor provides a scenic view of the city at sunset.

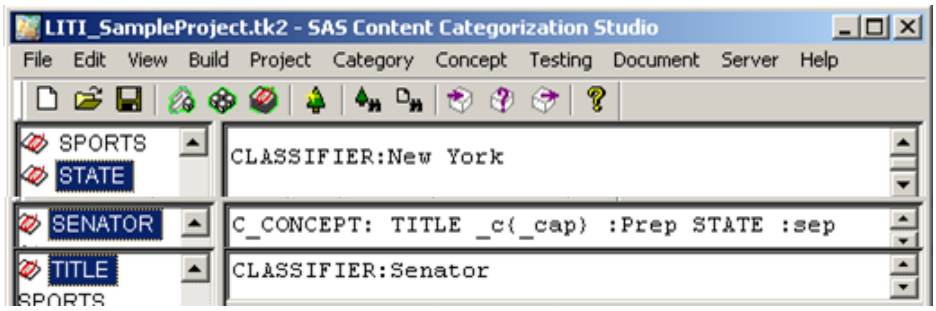
Figure 7-16 A Prioritized Match in an Input Document



7.8.9 Specifying Part-of-Speech Tags

Like SAS Content Categorization Studio, SAS Enterprise Content Categorization Studio enables you to use part-of-speech tags to locate matches. These tags are useful when you want to locate a wide range of matches without specifying a list of dictionary entries. Part-of-speech tags are particularly useful when you know the syntax, but not the wording of, the exact matches that you are seeking.

Figure 7-17 C_CONCEPT with Part-of-Speech Tags



A space is required before the colon (:) that precedes the part-of-speech tag. Specify a lowercase s in the sep part-of-speech tag.

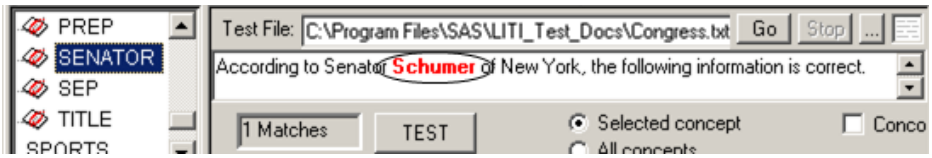
Example 7-12: Using Part-of-Speech Tags

Concept Name	**Entry**
STATE	CLASSIFIER:New York
TITLE	CLASSIFIER:Senator
SENATOR	C_CONCEPT: TITLE _c{_cap} :Prep STATE :sep

Schumer is returned as a match for the SENATOR concept when a preposition (Prep) precedes a match on the CITY CLASSIFIER concept and a separator (sep) character follows this concept. See the following example:

> According to Senator *Schumer* of New York, the following information is correct.

Figure 7-18 A C_CONCEPT Rule with Part-of-Speech Tag Match in an Input Document



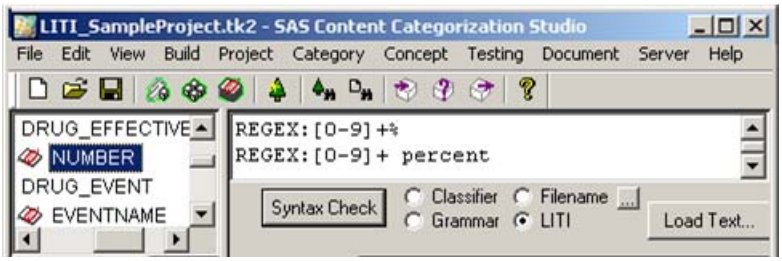
7.8.10 Specifying Regular Expressions

Specify regular expressions to locate matches based on known patterns. For example, telephone numbers, street, and e-mail addresses are all defined using recognizable patterns. When you write regular expressions, you specify a range of letters or numbers inside square braces (`[]`) to form a regular expression rule. For example, type `a-z` or `0-9`. This syntax matches any ASCII character whose value is between a and z or between 0 and 9 inclusive.

If you add a plus (+) sign after the last brace, all lowercase letters are matched. For example, you could write `REGEX:[a-z]+`.

You can also add either the `%` symbol or write out the word `percent`. If you do this after you add the plus (+) symbol all of the instances of percentages in the input document are returned.

Figure 7-19 Regular Expression Rules



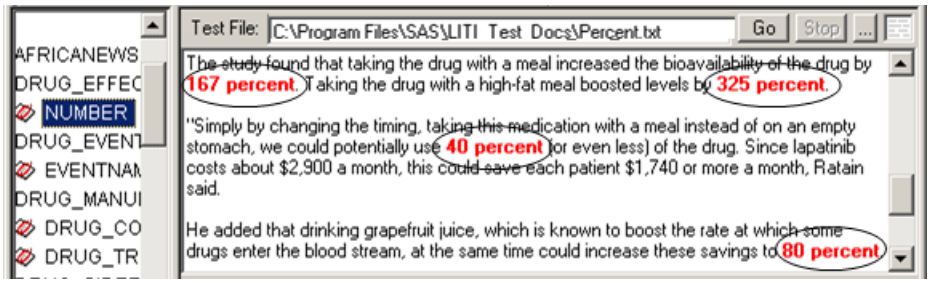
The NUMBER concept has `REGEX` rules. The different specifications for percent ensure wider definition coverage.

Example 7-13: Specifying Regular Expressions

Concept Name	Entry
NUMBER	REGEX:[0-9]+% REGEX:[0-9]+ percent

This regular expression definition specifies that numbers followed by either percentage sign match. For example, matches on both 99%, and 50 percent are both returned.

Figure 7-20 REGEX Rule Matches in an Input Document

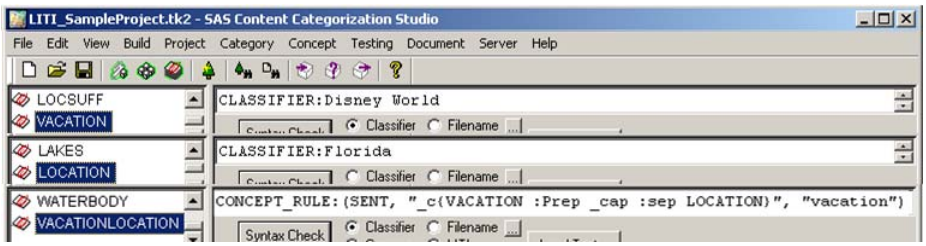


Notes: For more information, see Appendix B: *Using the Directive and Regex Syntax on page 7*.
You can also specify a returned information string after a comma (,). In this case, the returned information is the value for the matched concept. For more information, see *SAS Content Categorization Studio: User's Guide*.

7.8.11 Specifying a SENT Operator

By default, SAS Content Categorization Studio looks for matches within the entire text of an input document. Limit matches to one sentence by writing the SENT operator into a CONCEPT_RULE.

Figure 7-21 CONCEPT_RULE with a Sentence Operator



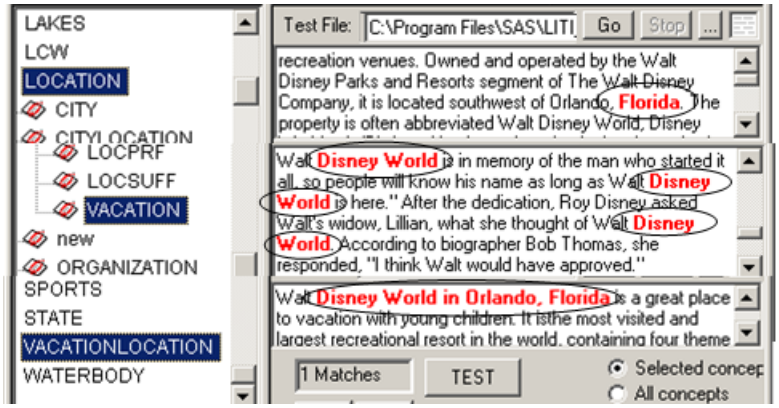
The VACATIONLOCATION concept specifies that a match is returned only when all of the specified elements are located in the context of a sentence.

Example 7-14: Specifying a Sentence Operator

Concept Name	Entry
VACATION	CLASSIFIER:Disney World
LOCATION	CLASSIFIER:Florida
VACATIONLOCATION	CONCEPT_RULE:(SENT, “_c{VACATION :Prep _cap :sep LOCATION}”, “vacation”)

The VACATIONLOCATION definition uses the CONCEPT_RULE to identify a match, when all of the specified components occur within one sentence. These matches occur when a preposition follows a VACATION concept match, a word that begins with an uppercase letter, a separator character, and a match on the LOCATION concept. If this match is followed by a match on the token vacation, a match is returned for the VACATIONLOCATION concept.

Figure 7-22 CLASSIFIER and CONCEPT_RULE Matches



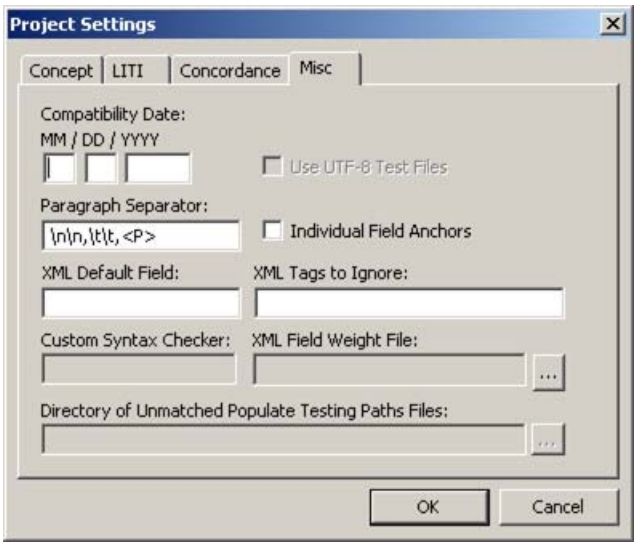
7.8.12 Specifying a PARA Operator

By default, SAS Content Categorization Studio looks for matches within the entire text of an input document. Limit matches to one paragraph by writing the `PARA` operator into the `CONCEPT_RULE`.

Before you specify your concept definitions, specify the paragraph separator that is used in your documents. For example, specify `<p>` for `.html` documents. If you are using multiple types of documents, list the paragraph separator for each type.

To specify the paragraph separator, complete these steps:

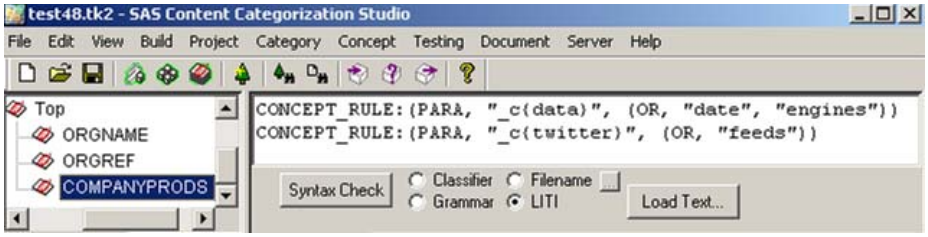
1. Select **Project --> Settings**. The Project Settings window appears.



2. Click **Misc** and the Misc window appears.
3. Enter the paragraph separators for your input documents into the **Paragraph Separator** field. For example, enter `\n\n,\t\t,<P>`.
4. Click **OK** to save your specifications.

After you specify your paragraph separators, you can write the rules for each concept.

Display 7-3 CONCEPT_RULEs with a Paragraph Operator



The PARA operator specifies that a match is returned only when all of the specified elements are located in the context of a paragraph. Each paragraph is delineated by one of these paragraph markers.

Example 7-15: Specifying Paragraph Operators

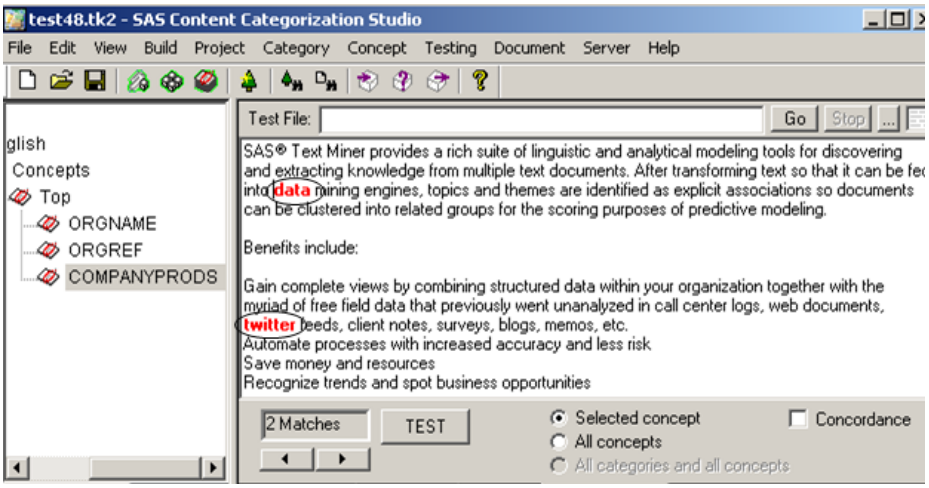
Concept Name	Entry
COMPANYPRODS	CONCEPT_RULE:(PARA, "_c{data}", (OR, "date", "engines")) CONCEPT_RULE:(PARA, "_c{twitter}", (OR, "feeds"))

The COMPANYPRODS definition uses CONCEPT_RULE definitions to identify matches within different paragraphs:

In the first case, a match occurs when *data* and either *date* or *engines* appear in the same paragraph.

In the second case, a match occurs when either *twitter* or *feeds* occur within the same paragraph.

Figure 7-23 CONCEPT_RULE and Paragraph Matches



7.8.13 Specifying a DIST Operator

Specify the maximum number of words in which matches can be located, instead of using the default behavior to search the entire document. The distance (`DIST_n`) operator for `CONCEPT_RULE` enables you to specify the maximum number of words that can occur between matches on the first and the last term. However, this operator does not specify the ordering of the matches.

Figure 7-24 CONCEPT RULE with a Distance Specification

The AFRICANEWS definition specifies that a match is returned if there are no more than 11 words between a match on the LASTNAME and LOCATION concepts.

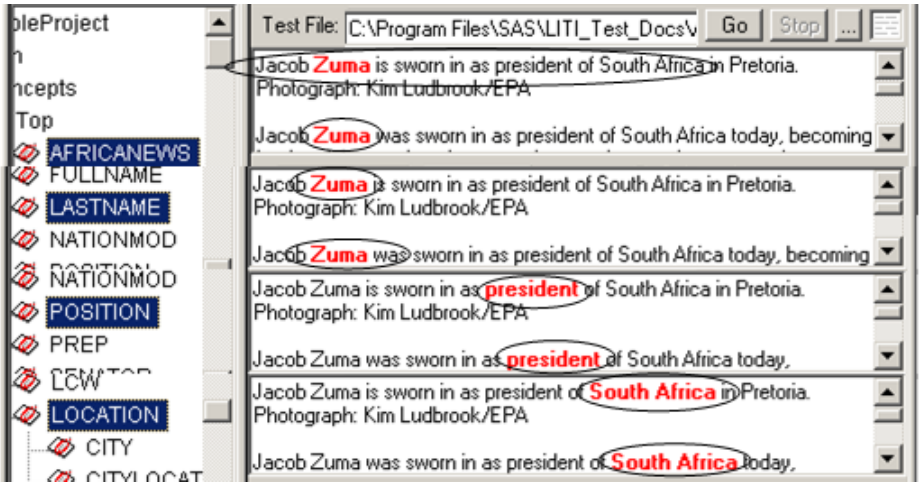
Example 7-16: Specifying the DIST Operator

Concept Name	**Entry**
`LASTNAME`	`CLASSIFIER:Zuma`
`POSITION`	`CLASSIFIER:president`
`LOCATION`	`CLASSIFIER:South Africa`
`AFRICANEWS`	`CONCEPT_RULE:PRIORITY=15:(DIST_11, “_c{LASTNAME}”, “POSITION”, “LOCATION”)`

The AFRICANEWS concept uses the `DIST` operator to specify a distance of 11 words between the location of a match on the LASTNAME concept and the LOCATION concept. This match is returned if there is also a match on the POSITION concept within these 11 words. In addition, this `CONCEPT_RULE` overrides the default **Priority** setting in the Data window. If there were other

rules in this definition, these rules would keep the same priority setting specified in the Data window.

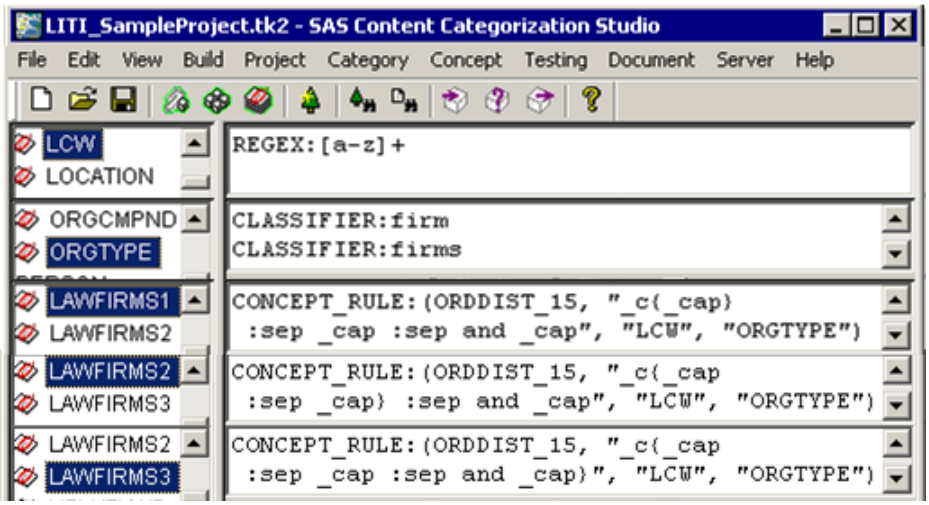
Figure 7-25 CONCEPT_RULE and CLASSIFIER Matches in Input Documents



7.8.14 Specifying an ORDDIST Operator

The ORDDIST_n operator is similar to the DIST operator. However, the ORDDIST operator specifies the order and distance requirements that are necessary to return a match on the CONCEPT_RULE definition.

Figure 7-26 CONCEPT_RULE with ORDIST Operator:



The CONCEPT_RULE for each LAWFIRMS concept places the ending curly brace (}) in a different location to return different results from the same input document.

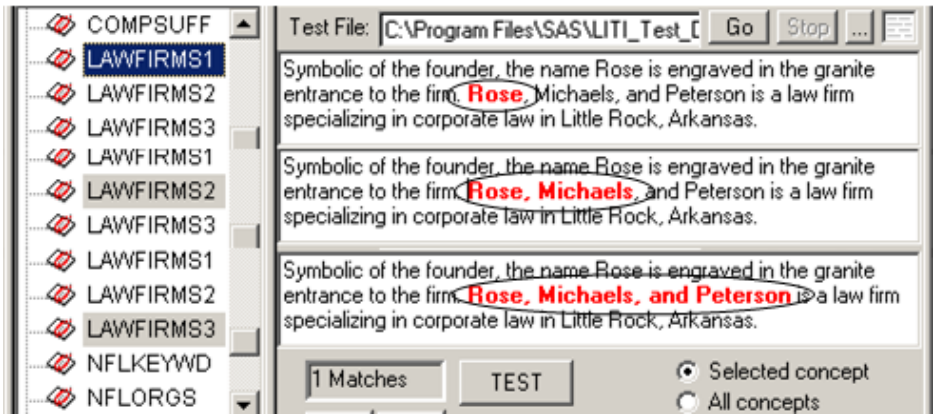
Example 7-17: Exporting Classifiers

Concept Name	Entry
LCW	REGEX:[a-z]+
ORGTYPE	CLASSIFIER:firm
	CLASSIFIER:firms
LAWFIRMS1	CONCEPT_RULE:(ORDDIST_15, “_c{_cap} :sep _cap :sep and _cap”, “LCW”, “ORGTYPE”)

This CONCEPT_RULE states that the following instances return a match if the matches occur in the specified order and within a distance of 15 words. A word begins with an uppercase letter and is followed by a separator character and an uppercase letter. This match is followed by a separator character, the token and, and another word beginning with an uppercase letter. The match is not returned unless the LCW REGEX rule is also matched and a match on the ORGTYPE concept also occurs within 15 words.

When the closing curly brace (}) is moved for the LAWFIRMS2 and LAWFIRMS3 concepts, the following matches are returned.

Figure 7-27 CONCEPT_RULE Matches in Input Documents



You can also change the default **Priority** setting of 10 in the Data window for any of the concept definitions shown above.

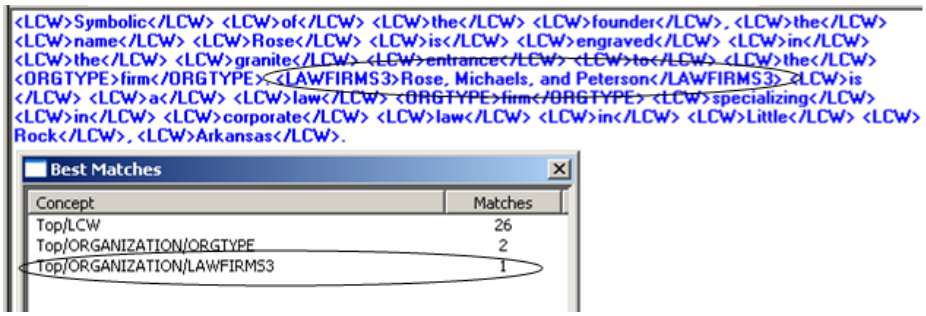
Display 7-4 Project Settings - LITI Settings

Use the Project Settings to affect how matches are returned:

- Select **All matches** and all of the matches for LAWFIRMS1, LAWFIRMS2, and LAWFIRMS3 above are returned. In this case, because the greater than (>) symbol does not end any of the `CONCEPT_RULE` definitions, only one match is returned for each concept.
- Select **Longest** and a match on LAWFIRMS3, only, is returned.
- Select **Best** and a match LAWFIRMS3 is returned. This is true unless you specify a higher priority in either the Data window or within a concept definition.

  See the example shown below that applies to both the **Longest** and **Best** selections:

Figure 7-28 Best Matches Window



- Select **Return all identical matches**, if either **Longest** or **Best** is matched, and all of the instances with the same priority or length are returned.
- The **Remove duplicate facts** operation does not apply. No facts can be specified for `CONCEPT_RULE` definitions.

7.8.15 Specifying the NOT Operator with the AND Operator

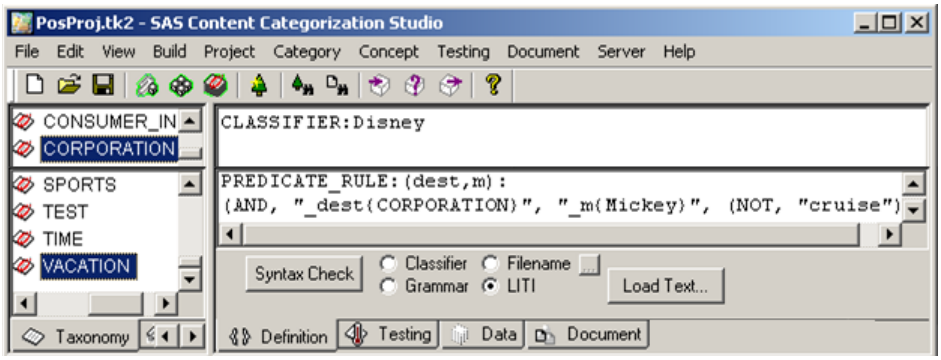
Use the NOT operator to exclude matches on the term that is specified by the NOT operator. Any other matches specified by this rule are returned. When you use this operator, also use the AND operator in a CONCEPT_RULE or a PREDICATE_RULE. When you write a PREDICATE_RULE, do not specify an argument within the NOT operator syntax and be sure to specify the AND operator.

You can choose to specify one argument in a PREDICATE_RULE. For example, write the following rule when you want to match documents that contain the term *Disney*, but not the word *cruise*:

```
PREDICATE_RULE:(dest):(AND, "_dest{Disney}",
                       (NOT, "cruise"))
```

An example of this rule, with two arguments and a reference to the CORPORATION concept, is shown below:

Figure 7-29 The NOT Operator in a Rule



This VACATION concept consists of one PREDICATE_RULE that specifies a NOT operator with the AND operator. In this example, the AND operator takes two arguments.

Hint: You can also use single letters for arguments such as _a and _b.

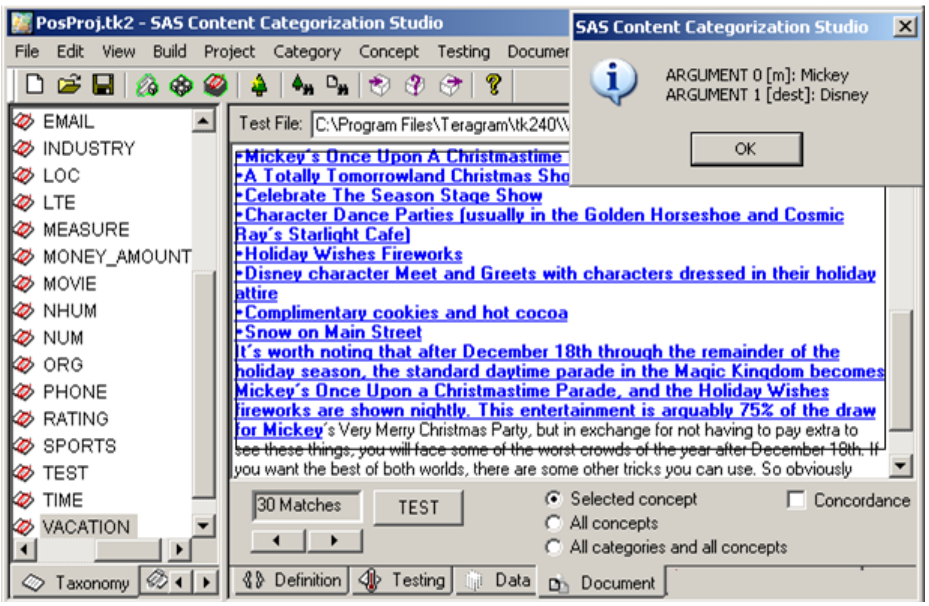
In the following example, the AND operator also takes two arguments.

Example 7-18: Matching Using a NOT Operator

Concept Name	Entry
VACATION	PREDICATE_RULE:(dest,m):(AND, "_dest{CORPORATION}", "_m{Mickey}", (NOT, "cruise"))
CORPORATION	CLASSIFIER:Disney

This VACATION rule specifies that a match occurs when there is a match on the CORPORATION concept and the term *Mickey* is also located in the input document. This statement is true only when the word *cruise* is not located in the same document. See the following match example where *cruise* does not occur:

Figure 7-30 NOT Operator Rule Matches in an Input Document



Note: At this time, the index position of the arguments in the SAS Content Categorization Studio window is arranged in reverse order.

The term cruise is present in the input document below. This statement is true even though the terms *Mickey* and *Disney* are also present in this document. See the following example:

Display 7-5 No Matches Are Returned for This Document

When you specify a `NOT` operator, make sure that you follow these requirements:

- A `NOT` operator is always used with an `AND` operator and the `AND` operator precedes the `NOT` operator. The `NOT` operator cannot be used alone.
- Use the `NOT` operator with a `CONCEPT_RULE` or a `PREDICATE_RULE`.
- When you write a `CONCEPT_RULE`, use only one `_c` marker.
- You cannot specify a `_c` marker in a `NOT` string. For example, see the `NOT` string above `(NOT, "river")`.
- When you write a `PREDICATE_RULE`, do not specify an argument in the `NOT` specification.
- The `NOT` operator accepts only one argument. For this reason, it is necessary to remove either `"river"` or `"rainforest"` from the following rule:

```
CONCEPT_RULE:(AND, “_c{Amazon}”, (NOT, “river”,
            “rainforest”))
```

For more information and examples, see Section 7.7.1.C *The NOT Operator* on page 142.

7.8.16 Specifying the UNLESS Operator

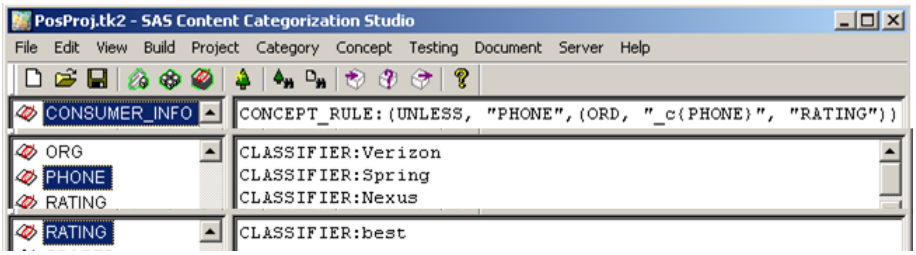
Use the UNLESS operator to set a constraint on possible matches. The UNLESS operator is often used to ensure that an attribute is correctly assigned to the appropriate subject. For this reason the UNLESS operator sets boundary constraints.

Hint: If you want to prevent a match on another rule, write a REMOVE_ITEM rule. For more information, see Section 7.8.6 *Disambiguating Matches* on page 158.

The UNLESS operator places additional restrictions on a Boolean operator. For this reason, this Boolean operator takes two arguments. A match is returned for the UNLESS operator only when the match specified by the UNLESS operator is not located within the window specified by the second Boolean operator. Specify only the following Boolean operators with the UNLESS operator: AND, SENT, DIST, ORD, and ORDDIST.

For example, if two phones are compared and you want to return only the matches for the best phone, the UNLESS operator can be used. In this case, you can specify the UNLESS operator when you want to return only the matches for brand name phones that are rated *best*. For this reason, matches are not returned for phones with any other ratings and the correct brand is associated with the rating. See the following example:

Display 7-6 The UNLESS Operator in a Rule



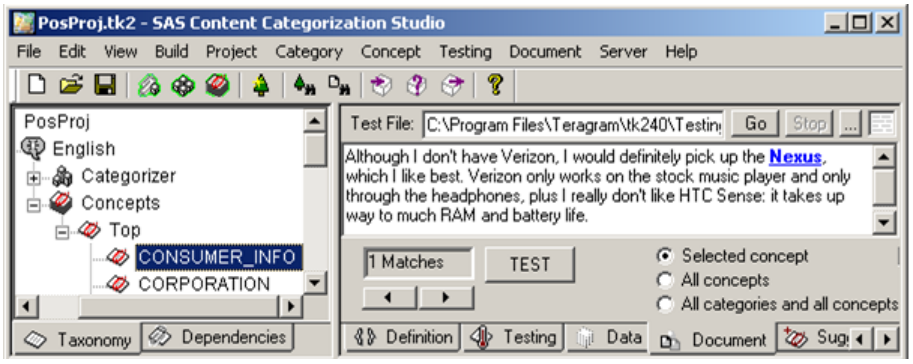
This CONSUMER_INFO concept consists of one CONCEPT_RULE that specifies an UNLESS operator.

Example 7-19: Matching Using an UNLESS Operator

Concept Name	Entry
PREDICATE_RULE	CONCEPT_RULE:(UNLESS, "PHONE",(ORD, "_c{PHONE}", "RATING"))
CLASSIFIER	Verizon
CLASSIFIER	Spring
CLASSIFIER	Nexus
CLASSIFIER	best

See the following example of the matches for this rule:

Display 7-7 UNLESS Operator Matches in an Input Document



The UNLESS argument applies possible matches to its first argument as a constraint against the Boolean argument. In the rule shown above, the Boolean argument is:

```
(ORD, "_c{PHONE}", "RATING")
```

The following requirements apply to a rule that uses the UNLESS operator:

- Use the UNLESS operator with a CONCEPT_RULE or a PREDICATE_RULE.
- The UNLESS operator accepts two high-level arguments. (In Example 7-7 above, these arguments are: "PHONE", and (ORD, "_c{PHONE}", "RATING").) The second argument takes a Boolean operator with at least two arguments.

- The first argument for an UNLESS operator can be a term or a concept such as "PHONE". In this example, PHONE is not a term, PHONE refers to another concept match.
- The second argument contains a Boolean operator and references two existing concepts or specifies two terms to match. (You can also specify one term and one concept, in any order.)

Note: If the second argument does not contain a Boolean operator, a syntax error is returned when you click **Syntax Check** in the Definition window.

7.9 Locating Facts

7.9.1 Overview of Facts

Facts, or predicates, refer to terms that match at least two concepts. Facts consist of at least two arguments. For example, *Harry Truman was president of the United States* is a fact based on several arguments. These arguments are defined by the following concepts NAME, TITLE, and COUNTRY. The following matches *Harry Truman*, *president*, and *United States* are returned to these concepts. By specifying this type of rule, you also locate similar matches in input documents without rewriting your rules.

Note: When you specify an argument such as _title, use only lowercase letters, enclose in quotation marks (“ “), and separate the argument with a comma.

Both SEQUENCE and PREDICATE_RULEs extract facts. SEQUENCE rules specify the order of the matches. PREDICATE_RULEs use Boolean operators, but do not specify the ordering of any matches unless you specify the ORD or ORDDIST operators. For more information, see Section 7.9.2 *A Predicate Sequence Example* on page 188 and Section 7.9.3 *The Predicate Examples* on page 191.

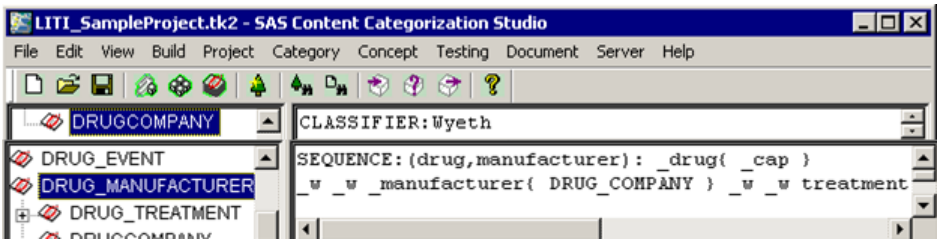
7.9.2 A Predicate Sequence Example

Identify previously unknown relationships, otherwise known as facts or events, in input documents. Predicate sequence, or `SEQUENCE`, rules extract the meaningful relationships between matched concepts and tokens. For example, identify the names and positions that various managers hold within a company. Locate this information even when these relationships are unknown to you, or when the concepts do not directly follow one another.

Predicates are also defined as facts or events. The terms are interchangeable. Facts are always defined by at least two concepts or tokens and one or more parts of speech. The term *sequence* is used to specify the necessary ordering of the concepts and semantic terms that define these facts.

When you specify a predicate sequence definition, you define not only the concepts, but also the arguments that are used with these concepts. Use this rule to also specify the sequence of these entities and any appropriate parts of speech.

Figure 7-31 SEQUENCE Rule



Example 7-20: Writing a Predicate Sequence Definition

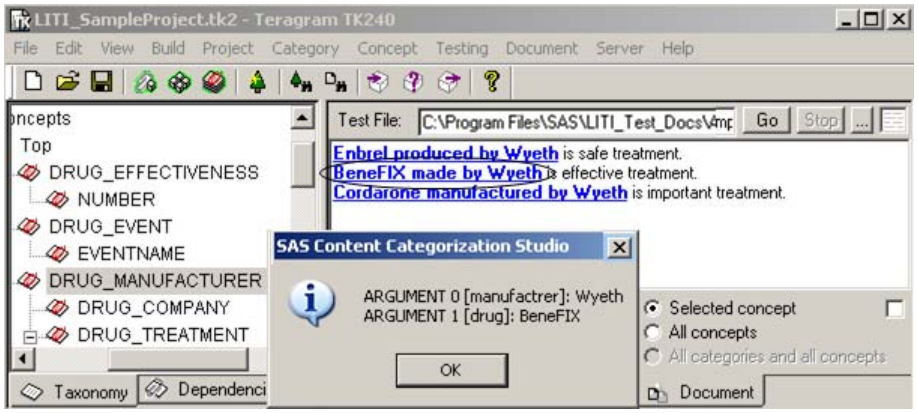
Concept Name	**Entry**
`DRUG_COMPANY`	`CLASSIFIER:Wyeth`
`DRUG_MANUFACTURER`	`SEQUENCE:(drug,manufacturer): _drug{ _cap } _w _w _manufacturer { DRUG_COMPANY } _w _w treatment`

This `SEQUENCE` rule takes the arguments `drug` and `manufacturer`. To locate the `_drug` predicate, locate a word that begins with an uppercase letter that is followed by two tokens. To match the `_drug` predicate, locate the DRUG_COMPANY concept followed by two tokens and the word *treatment*.

However, only the matches within and between the beginning and ending curly braces ({}) are returned as a match for this concept.

For example, the fact *BeneFIX produced by Wyeth* is returned as a match to the DRUG_MANUFACTURER `SEQUENCE` concept along with the matches on the arguments for this fact. You can see the fact matches in the Document window for this testing document. You can also click on one of the returned facts to open a SAS Enterprise Content Categorization Studio status screen. This screen lists the matching arguments for the selected fact.

Figure 7-32 Argument Matches in an Input Document



SAS Content Categorization Server locates this fact and its arguments when you provide the `.li` file. In other words, you can decide to have SAS Content Categorization Server return all of the information shown in the example below. Alternatively, specify different returns. For example, choose to return only the fact, or only its arguments.

Example 7-21: SAS Content Categorization Server Output

```
FACT 0:[0(0)_4(24)]/DRUG_MANUFACTURER/: BeneFIX produced by
Wyeth


ARG 0 [manufacturer]: Wyeth
ARG  [drug]: BeneFIX
```

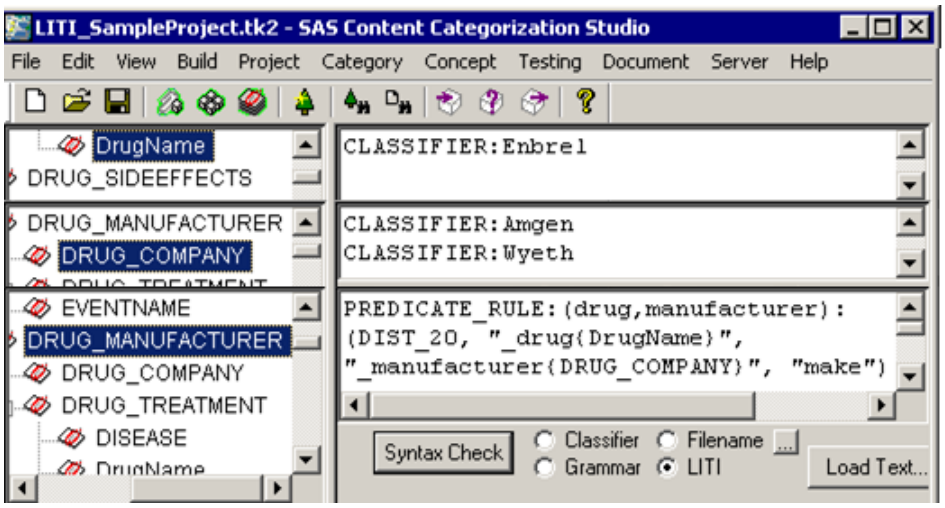
Note: At this time, the index position of the arguments in the SAS Content Categorization Studio window is not specified correctly.

Use the SAS Content Categorization Server Client API to specify fact matching strings, arguments, and the offsets returned by SAS Content Categorization Server.

7.9.3 The Predicate Examples

Like SEQUENCE rules, PREDICATE_RULEs locate facts and their supporting arguments. Unlike SEQUENCE rules, PREDICATE_RULEs do not specify the matching order. Instead, PREDICATE_RULEs use Boolean operators to increase the matching precision within the document. For more information, see Section 7.7 *The Operators* on page 140.

Figure 7-33 PREDICATE_RULE with Logical Operators



Like the preceding SEQUENCE rule, this PREDICATE_RULE defines the arguments drug and manufacturer. However, the DRUG_MANUFACTURER PREDICATE_RULE uses the DIST operator. This operator specifies that a match is returned when the DrugName concept is located within 20 words of a match on the DRUG_COMPANY concept.

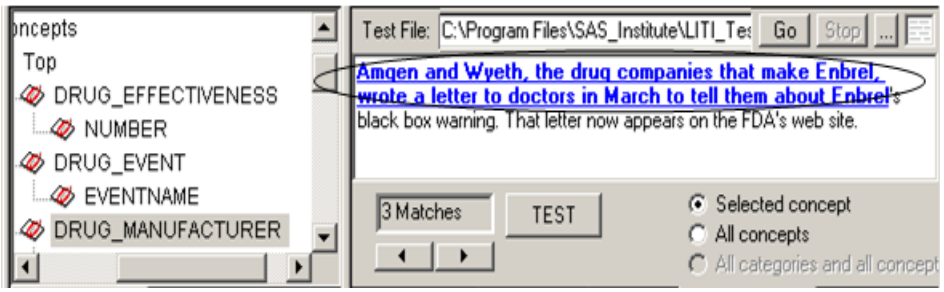
Example 7-22: Viewing a PREDICATE_RULE

Concept Name	Entry
DrugName	CLASSIFIER:Enbrel
DRUG_COMPANY	CLASSIFIER:Amgen CLASSIFIER:Wyeth
DRUG_MANUFACTURER	PREDICATE_RULE: (drug,manufacturer): (DIST_20, "_drug{ DrugName }", "_manufacturer{ DRUG_COMPANY }", "make")

This PREDICATE_RULE defines the arguments drug and manufacturer. Inside the parentheses that follow each argument is the concept that identifies a match. The DIST operator specifies that matches on the DrugName concept can occur within 20 words of a match on the DRUG_COMPANY concept. In addition, a match on the DRUG_MANUFACTURER concept only occurs when the token make is located. Although no other tokens are specified for this PREDICATE_RULE, all of the words located between matches on the concepts DrugName and DRUG_COMPANY are returned as a matching phrase. However, because a PREDICATE_RULE is specified and not a SEQUENCE rule, these matches can occur in any order.

For PREDICATE_RULEs, like other definitions, multiple matches can occur in one document, and multiple facts can be returned.

Figure 7-34 PREDICATE_RULE Match in an Input Document



The results shown above are returned when the default setting, **All matches**, is selected under the **Overlapping Concept Matches** heading in the Project Settings - LITI dialog box.

Click  and  in the Document window to see each of the following matches:

Amgen and Wyeth, the drug companies that make Enbrel

> This fact matches the word *Wyeth* as a token. It is not a match on the *DrugName* concept.

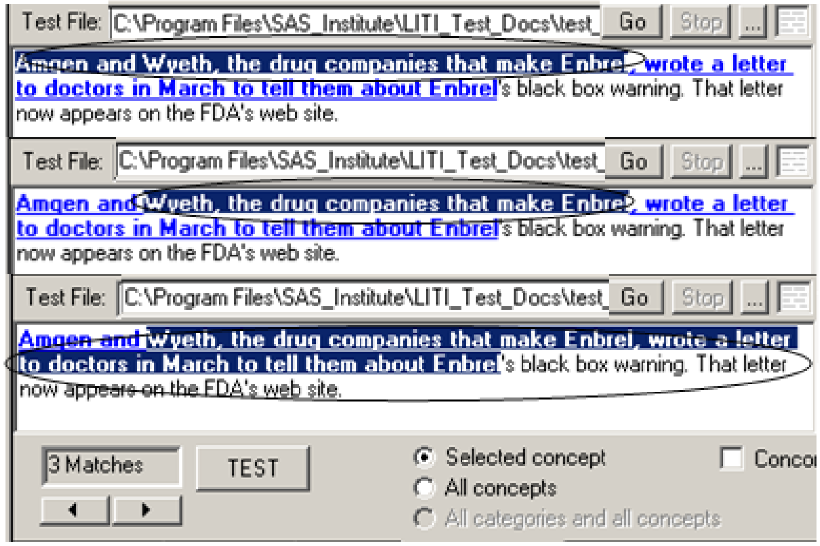
Wyeth, the drug companies that make Enbrel

> This is the shortest of the matches that begin with a match on `Wyeth` in the `DRUG_COMPANY` concept and end with `Enbrel` as a match on the DrugName concept. Also see the following bulleted point.

Wyeth, the drug companies that make Enbrel, wrote a letter to doctors in March to tell them about Enbrel

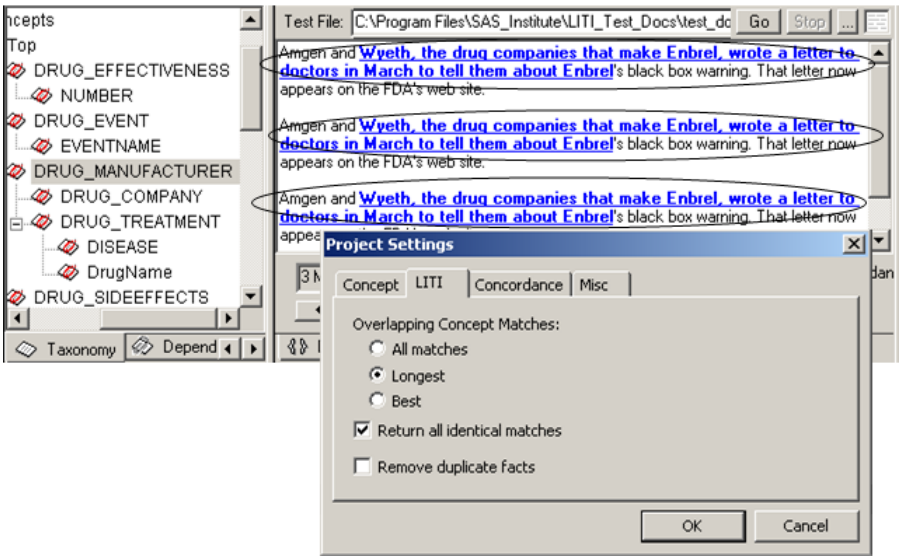
> This is the longest of the matches that begin with a match on `Wyeth` in the DRUG_COMPANY concept and end with `Enbrel` as a match on the DrugName concept. In this case, the first instance of `Enbrel` is matched as a token and not as a match on the DrugName concept. Also see the bulleted point above.
>
> This match is returned when you select **Longest** under the **Overlapping Concept Matches** heading in the Project Settings - LITI dialog box.
>
> These results are also returned when you select **Best**. This statement is true unless you set a **Priority** specification in the **Definition** tab or overwrite the default setting of `10` in the Data window for this concept.

Figure 7-35 PREDICATE_RULE Matches in Input Documents



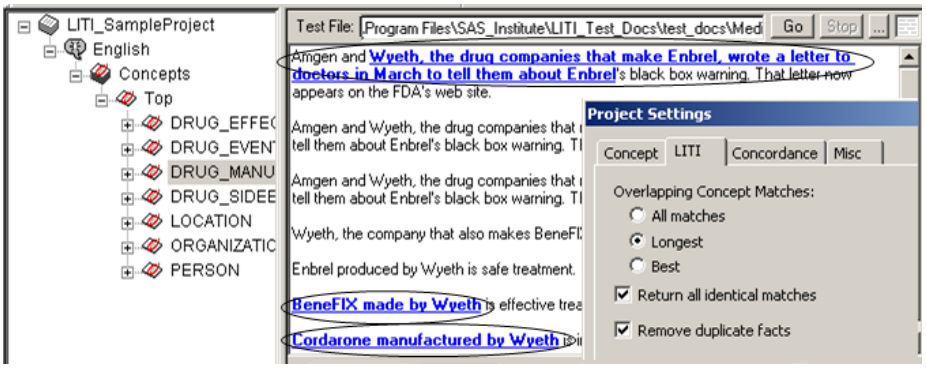
To return all of the instances of the longest fact matches, select **Return all identical matches** in the Project Settings - LITI dialog box. This operation can be selected only if you have also selected either **Longest** or **Best** under the **Overlapping Concept Matches** heading.

Figure 7-36 Several Instances of a Match in an Input Document



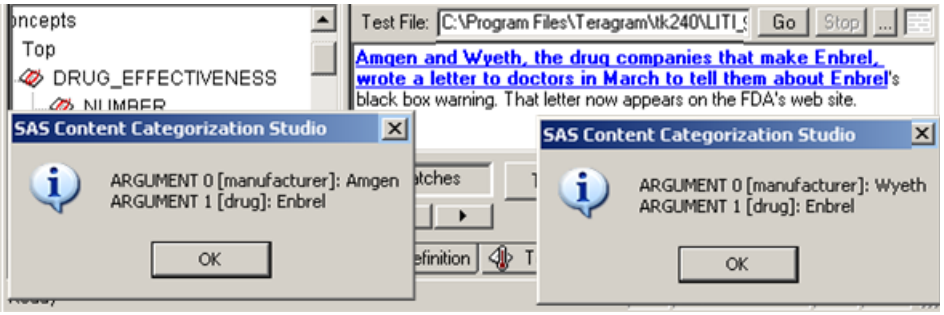
- In the figure below, **Remove duplicate facts** is added to the selections in the figure above. New text is added to the testing document to illustrate the functionality of these interrelated settings. Each instance of a match that is the longest for any of the overlapping matches, but not a duplicate fact, is returned as a match to the selected concept.

Figure 7-37 Longest Unique Matches in an Input Document



All of these facts are highlighted, and initially appear as a single match to the PREDICATE_RULE definition for the DRUG_MANUFACTURER concept. However, there are two sets of arguments, because there are two matches on the DRUG_COMPANY concept and one match on the DrugName concept. It is these matches that define the beginning and end of each fact.

Figure 7-38 Facts and Arguments in an Input Document



Note: At this time, the index position of the arguments in the SAS Content Categorization Studio window is not specified correctly.

7.10 Using Predefined Concepts

7.10.1 Overview of Using Predefined Concepts

Predefined concepts shorten the process of rule-writing by enabling you to reference concept rules that are already defined. For example, choose to reference ORGANIZATION and save the time of creating this concept and writing a list of organizations into CLASSIFIER rules.

Before you use this section, it is important to understand the following information:

- The `ACTIVATE` term that appears with the personal pronoun predefined concepts is necessary. For this reason, no rule type is specified for personal pronoun concepts.
- If you have a concept in your taxonomy with the same name and case as a predefined concept, both rules are applied in the testing process.
- The rules for the predefined concepts are not accessible. For this reason, test your concepts to ensure that you obtain the matches that you expect.
- If you have a term in your document that matches the predefined concept in your rule, this term might also match.
- Predefined Contextual Entities are not available for `CLASSIFIER` rules.
- You can download an additional set of predefined concepts known as predefined dictionary-based entities.

7.10.2 Optional: Download Predefined Dictionary-Based Entities

Before you add predefined concepts to your project, you can download an additional set of predefined LITI concepts at `http://support.sas.com/demosdownloads/setupintro.jsp`. Select the Text Analytics link. These predefined dictionary-based entities are available in a `.zip` file.

To install this file into the folder that makes this feature available to the program, complete these steps:

1. Go to the installation directory, such as:
   `C:\Program Files\Teragram\tk240\data`.
2. Paste and unzip the downloaded `.zip` file into this installation folder.

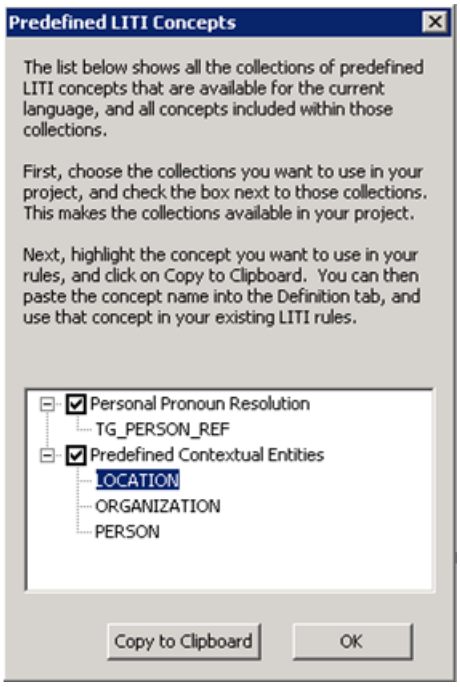
Notes: After you complete these steps, the Predefined Dictionary-Based Entities appear in the Predefined LITI Concepts window. For more information, see Section 7.10.3 *Copy and Paste a Predefined Concept* on page 198.
Make sure that you download the data file for the correct release.

7.10.3 Copy and Paste a Predefined Concept

Use the Predefined Concept window to copy a concept that you can paste into a LITI definition. These steps apply to any of the predefined concepts.

To copy and paste a predefined concept, complete these steps:

1. Go to **Concept --> Show Predefined LITI Concepts List**.



> **Note**: If you followed the steps in Section 7.10.2 *Optional: Download Predefined Dictionary-Based Entities* on page 197, Predefined Dictionary-Based Entities also appear in this window.

2. Choose a concept. For example, select `LOCATION`.
3. Click **Copy to Clipboard**.
4. Click **OK** to save this selection on the clipboard.

5. Paste this concept into your concept definition.

7.10.4 Write a Personal Pronoun Rule

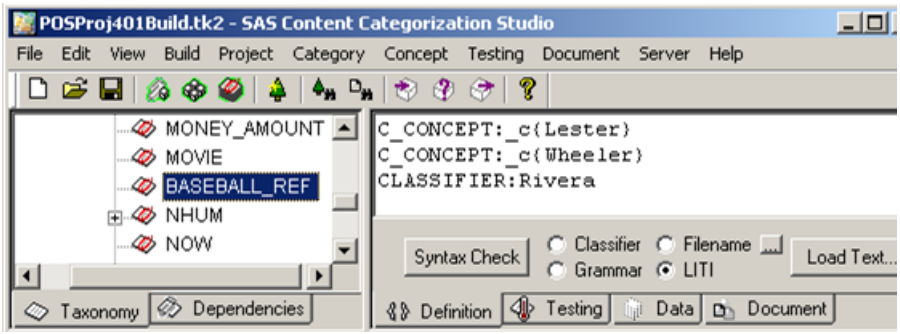
Use the personal pronoun predefined concept (`TG_PERSON_REF`) to locate matches on personal pronouns and the persons that these pronouns reference. (This process is also referred to as coreference. For more information about coreference rules, see Section 7.11 *The Coreference Operators* on page 204.) The `TG_PERSON_REF` predefined concept has no rule type but the syntax is preceded by the `ACTIVATE:` term.

This concept locates the pronoun that references the entities that you define in the related concept. For example, specify a LEADER concept that uses `CLASSIFIER` rules to locate the names of the current world leaders. Specify a LEADER_REF concept that references the LEADER concept using the `TG_PERSON_REF` concept.

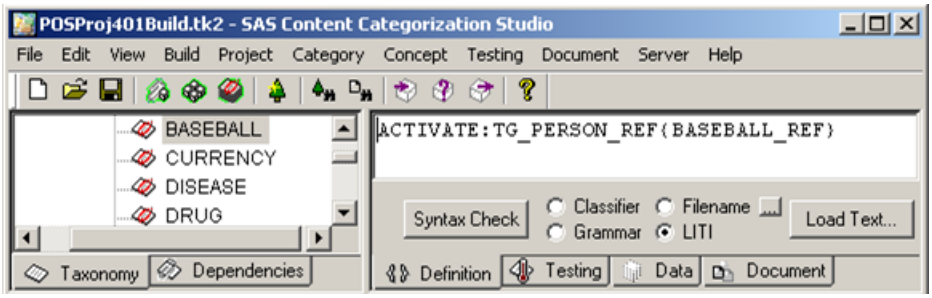
Hint: The pronoun resolution concept uses all of the rules in the referenced concept. For this reason, make sure that the concept you reference specifies only personal nouns.

To define concepts that perform pronoun resolution, complete these steps:

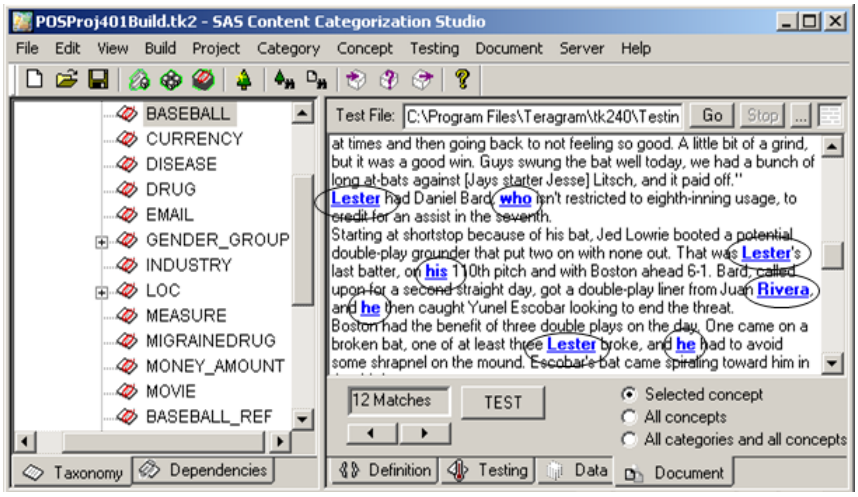
1. Write LITI rules such as a `CLASSIFIER` and `C_CONCEPT` rules that specify the name, or names, of the people that you want to identify coreference for in input documents. See the following example:



2. Use the steps in Section 7.10.3 *Copy and Paste a Predefined Concept* on page 198. Select `Personal Pronoun Resolution --> TG_PERSON_REF` to locate pronouns that might reference the person that you specified. (A personal pronoun concept includes the required `ACTIVATE` statements. You replace the `INSERT_CONCEPT_NAME_HERE` term with the name of the referenced concept.) See the following example:



3. Check the rule syntax, save, and compile your concepts.
4. Test the testing documents in the Testing pane.
5. Double-click a matched document to see the matches in the Document pane.



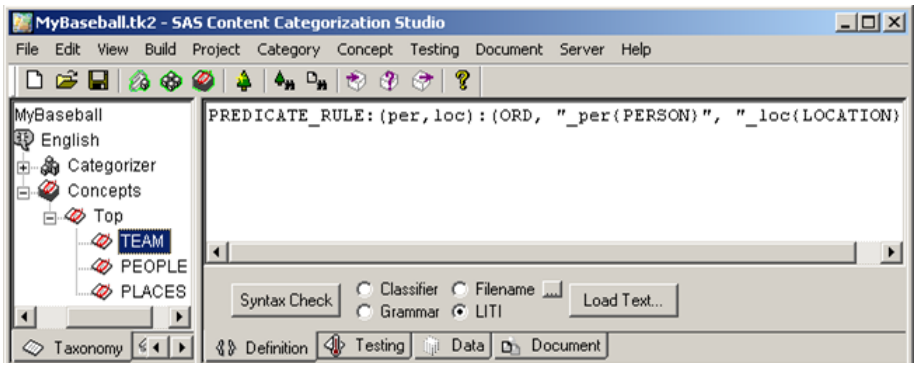
6. (Optional) Right-click on the matched pronoun to see its referenced noun.
7. (Optional) Test **All concepts** to see the matches for all of your concepts.
8. (Optional) Use the Concordance operation to see these matches in context.

7.10.5 Use the Predefined Entities

Paste a predefined contextual entity or a predefined dictionary-based entity (if you downloaded the latter) into your definition when you want to locate a match on a location, organization, or a person. These concepts simplify the rule writing process by enabling you to reference this preexisting concept rule.

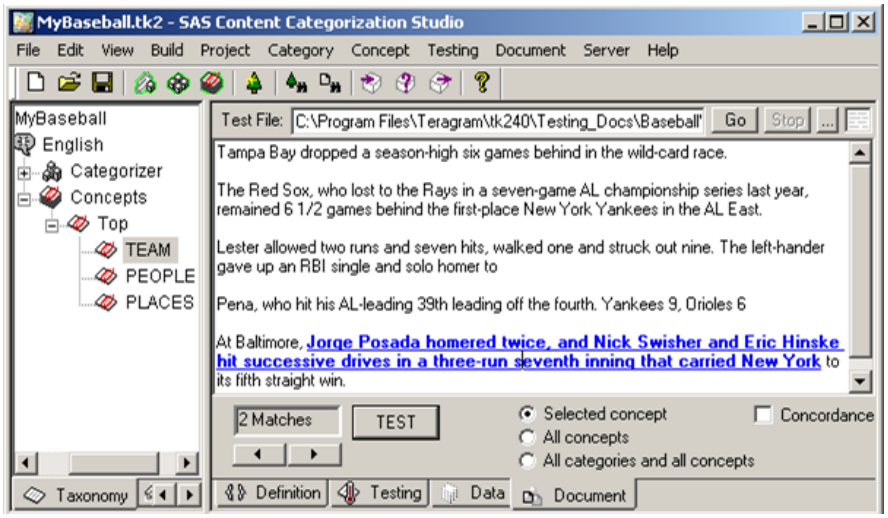
To use a predefined entity in a rule, complete these steps:

1. Write a LITI rule such as a `PREDICATE_RULE` rule.
2. Use the steps in Section 7.10.3 *Copy and Paste a Predefined Concept* on page 198. If you write a rule that specifies more than one concept, select one predefined contextual entity concept at a time. Paste the selected concept into the rule. See the example shown below:

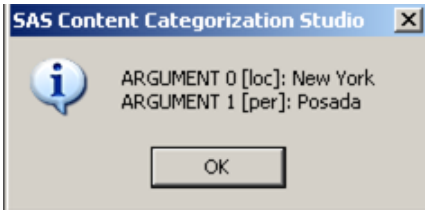


3. Check the rule syntax, save, and compile your concepts.
4. Test the testing documents in the Testing pane.

5. Double-click a matched document to see the matches in the Document pane.



6. Right-click on the rule to see match information about the arguments.



7. Click **OK** to close the SAS Content Categorization Studio
8. (Optional) Use Step 7 and Step 8 on 201.

7.11 The Coreference Operators

7.11.1 Overview of Coreference

Use coreference operators to write rules that return the canonical form of a word along with the referring term. Coreference operators are often used with pronouns, or other words that are called *referring terms*. (This is also known as *anaphora resolution*.) The canonical form of a word can be any term that you choose. For example, return a match on either *Barack Obama* or *President Barack Obama* for each instance of the referring term *Barack* in an input document. Another alternative is to choose to return *President Barack Obama* as the canonical form for each match on the pronoun *he*.

When the tested document is displayed in the **Document** tab, both the canonical word form and the matching term are highlighted. This is because these matches are linked in SAS Content Categorization Studio.

Use the coreference operator (`_ref`) with a `CONCEPT`, `C_CONCEPT`, or a `CONCEPT_RULE` rule. If you want to use a coreference qualifier in a `CLASSIFIER` rule, use `_coref` instead of `_ref`.

> **Note**: The **Overlapping Concept Matches** selections in the **LITI** tab of the Project Settings window do not affect matches made by the export, forward, and preceding operators.

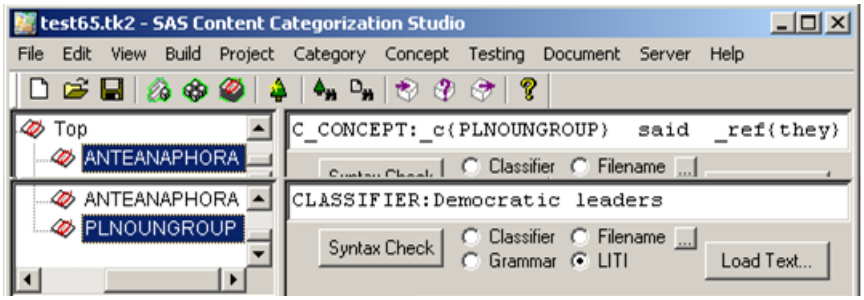
7.11.2 How to Use the Coreference Operator

Use the coreference operator (`_ref`) to link a matched string with its canonical form in an input document.

```
C_CONCEPT:{Jim Goodnight} said _ref{he}
```

In the example above, the canonical form *Jim Goodnight* is returned each time the matching term, *he* is located. This is true when the phrase *Jim Goodnight said he* is located in the text.

Figure 7-39 C_CONCEPT with _ref Operator.



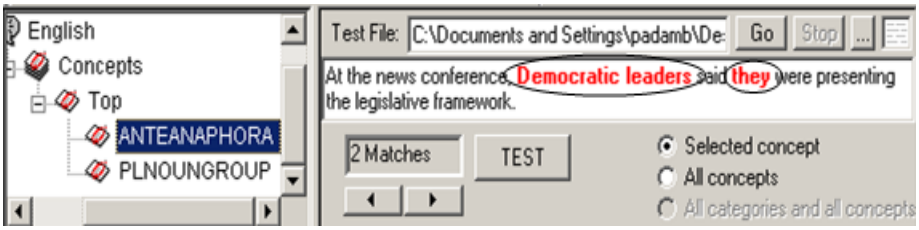
The _c operator is used in a `C_CONCEPT` rule that specifies the canonical form for the coreference specified by the `_ref` operator.

Example 7-23: C_CONCEPT Rule with the _ref Operator

Concept Name	Entry
PLNOUNGROUP	CLASSIFIER:Democratic leaders
PERSON	C_CONCEPT:_c{PLNOUNGROUP} said _ref{they}

When this definition is matched in an input document, a match on the referring term that follows the `_ref` operator returns the canonical form. The canonical form is specified in the bracketed term that follows the context operator (_c). This form is identified in the PLNOUNGROUP concept. In this example, the word that *they* references its specified canonical form *Democratic leaders*.

Figure 7-40 _ref Match in an Input Document



In this example, *Democratic leaders* and *they* are returned as matches in this input document. However, if the document contained other instances of the word *they*, these instances are not matched. You can see these matches in the Document window for this testing document.

7.11.3 How to Use the _ref Operator with the > Symbol

The greater than symbol (>) locates multiple instances of a match specified by the bracketed ({}) coreference operator (_ref) in an input document. For example, you might want to return the canonical form for each matched instance of a first name. In this case, you could specify a rule that identifies any references to *Jim* as a reference to *Jim Goodnight CEO of SAS Institute*. For more information, see Section 7.6.8 *The > Symbol* on page 133.

7.11.4 How to Use the _ref Operator with the Forward or Backward Symbols

7.11.4.A Limiting Matches to Those That Follow or Precede a Coreference Match

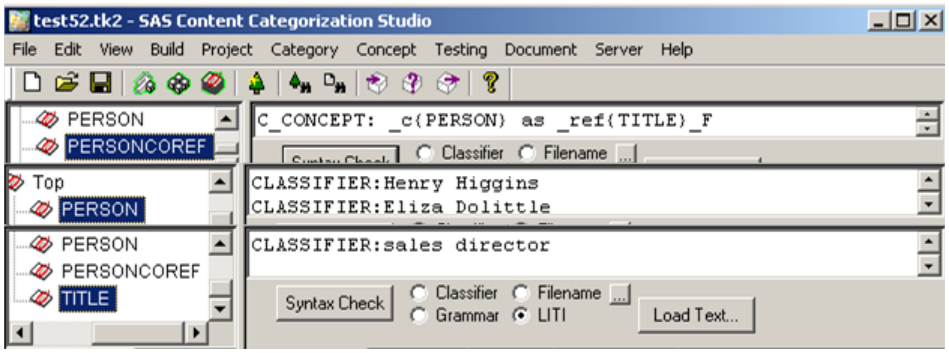
Use the forward (_F) and the preceding (_P) symbols to restrict coreference matches in an input document. When you specify these operators, only the matches which follow or precede the match for the rule, respectively, are returned.

Use these symbols when you want to return all of the matches instead of the one match that follows the rule (coref operator alone). Unlike the greater than (>) symbol, all of the returned matches can occur only before or after the coreference rule match.

7.11.4.B Matching with the Forward Symbol

Use the forward symbol (_F) to return all of the matches that follow a coreference rule match.

Figure 7-41 CONCEPT with _ref and Forward Symbol



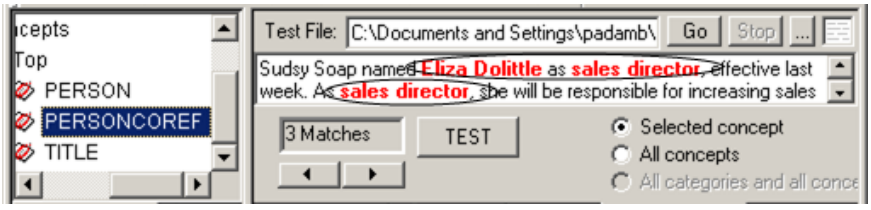
The example above shows a concept with a concept rule with a forward symbol. The rule specifies that all of the instances of matches on the coreference term that follow the coreference match are returned as matches. (Any matches which precede the match on the coreference term are not returned.)

Example 7-24: C_CONCEPT Rules with the _F Symbol

Concept Name	Entry
PERSON	CLASSIFIER:Eliza Dolittle
TITLE	CLASSIFIER:sales director
PERSONCOREF	C_CONCEPT:_c{PERSON} as _ref{TITLE}_F

In this example, a match on the term *Eliza Dolittle* as *sales director* matches. Instances of the term *sales director* that follow are also returned as matches.

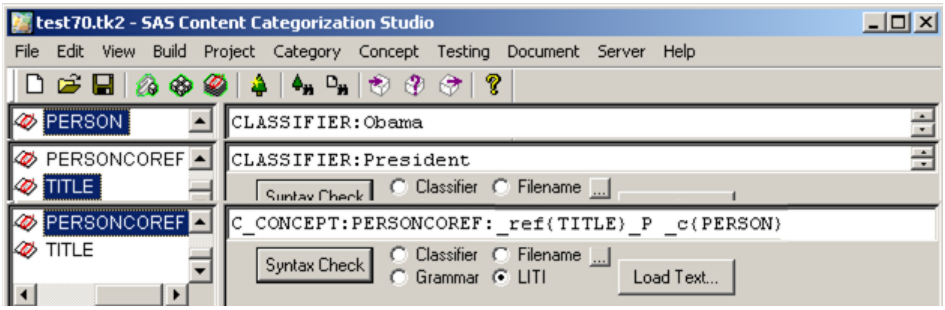
Figure 7-42 _ref and Forward Symbol Matches



7.11.4.C Matching with the Preceding Symbol

Use the preceding symbol (_P) to return matches on all instances of a coreference match that occur before the coreference rule match.

Figure 7-43 CONCEPT with _ref and Preceding Symbol



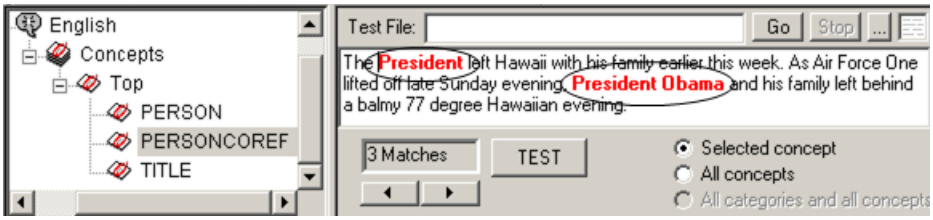
The example above shows a concept with a rule that specifies a preceding symbol. All instances of matches on the TITLE concept that are immediately followed by a match on the PERSON concept are returned as matches. (Any matches that follow the match on the coreference term are not returned.)

Example 7-25: C_CONCEPT Rules with the _P Symbol

Concept Name	Entry
PERSON	CLASSIFIER:Obama
TITLE	CLASSIFIER:President
PERSONCOREF	C_CONCEPT:_ref{TITLE}_P _c{PERSON}

In the example above, all instances of a match on the TITLE concept that precede a match on the TITLE and PERSON concepts are matched in an input document.

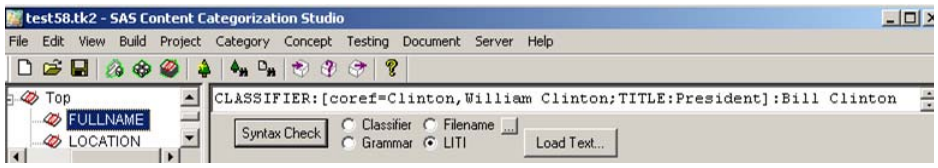
Figure 7-44 Matches on a Rule with the Preceding Operator



7.11.5 Coreference in a Classifier Definition Example

You can use the coreference operator (`coref`) to link a match in a coreference definition to its canonical form. For example, you might want to return *Barack Obama* for a match on any instance of the word *president* in an input document. The `coref` qualifier is used with classifier definitions, only.

Figure 7-45 Coreference Used to Link to Classifier Concept



The example above shows a classifier definition that links matches on the `coref` qualifier to its canonical form.

Example 7-26: A Classifier Concept with a Coreference Qualifier

Concept Name	**Entry**
FULLNAME	`CLASSIFIER:[coref=Clinton,William Clinton;TITLE:President]:Bill Clinton`

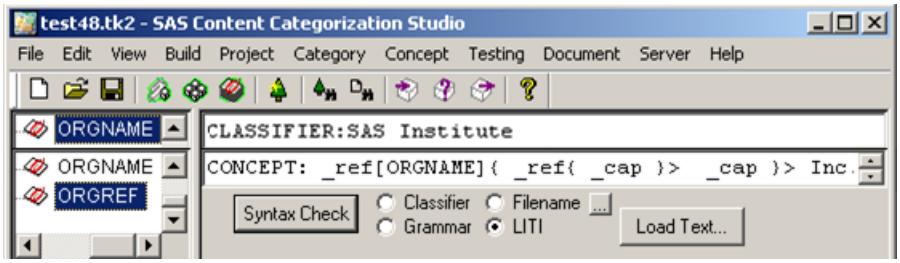
In the example above, if the canonical term *Bill Clinton* is matched once in an input document, all instances of matches on the `coref` qualifier terms also return matches. In this example, *Clinton*, *William Clinton*, and *President* all return matches. The canonical form for each matched term is *Bill Clinton*.

7.11.6 Assigning New Concept Names to Coreference Matches

You can assign a new concept name for a match on a term specified by the `_ref` operator. In this case, any instances of this match are output in SAS Content Categorization Server as a match on this new concept. You can also write a rule that specifies that a match is assigned to an existing concept. For example, you could assign matches on the names of an organization to an existing `CLASSIFIER` definition. In both cases, any matches on the complete definition are returned in the specified canonical form.

Specify a new, or an existing, concept name in square brackets (`[]`) that are preceded by the `_ref` operator. For example, specify `_ref [COMPANY]`.

Figure 7-46 Reassigning a Match



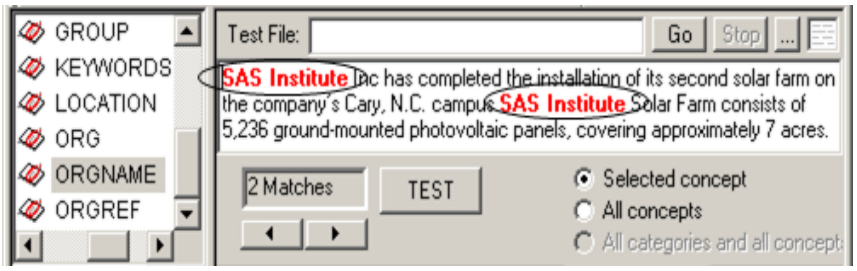
In the example above, if a sequence of two or more words that begins with an uppercase letter is followed by *Inc.*, a match is returned for the ORGREF concept. A sequence of two words that begin with uppercase characters is returned as a match for the concept ORGNAME. The canonical form is returned as a match for the ORGREF concept.

Example 7-27: Assigning a New Concept Name to a Coreference Match

Concept Name	Entry
`ORGNAME`	`CLASSIFIER:SAS Institute Inc.`
`ORGREF`	`CONCEPT:_ref[ORGNAME}] {_ref ( _cap)> _cap}> Inc.`

In the example above, a match on the ORGNAME concept is returned when there is a match on the remainder of the ORGREF rule. For example,

Figure 7-47 Match Returned to Another Concept

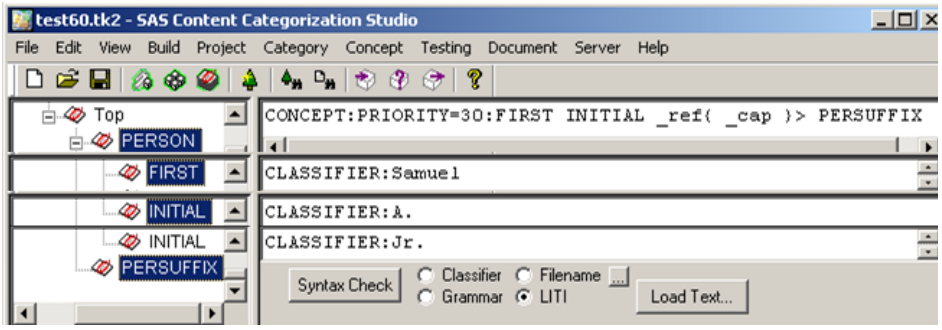


7.11.7 Rank Coreference Definitions and Eliminate False Positives

You can use the `PRIORITY` specification to make matches on one coreference rule rank higher than other rules. Specify a priority to rank matches on the concept that uses coreference higher than other matched concepts. (When you specify a `PRIORITY` in a rule, this setting overrides the **Priority** setting in the Data window—for this rule only.)

You can choose to specify a priority for a concept match that uses the `_ref` operator with the export symbol. You can also use the `PRIORITY` specification to eliminate false positives. For more information about priorities, see Section 7.8.8 *Setting Priorities for Overlapping Matches* on page 163.

Figure 7-48 CONCEPT with _ref and Export Symbol



In this example, if *Samuel A. Alito Jr.* is present once in the document, then every match on *Alito* returns his full name. The canonical form is *Samuel A. Alito Jr.* and the referring term is *Alito*.

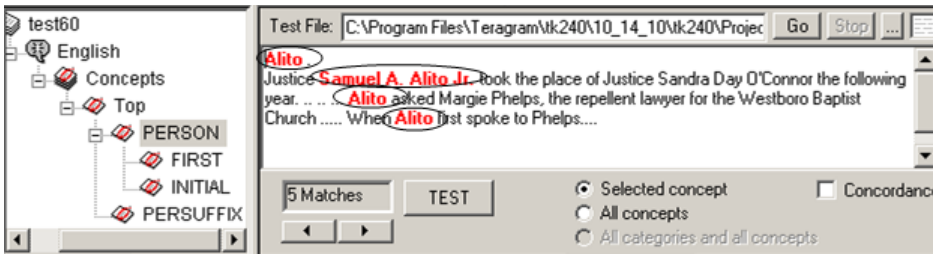
Example 7-28: C_CONCEPT Rule with the Export Symbol

Concept Name	**Entry**
`FIRST`	`CLASSIFIER:Samuel`
`INITIAL`	`CLASSIFIER:A.`
`PERSUFFIX`	`CLASSIFIER:Jr.`
`PERSON concept:`	`CONCEPT:PRIORITY=30:FIRST INITIAL _ref{ _cap }> PERSUFFIX`

In the example above, all instances of *Alito* are matched in an input document when all of the following conditions are met. A match on a first name listed in the FIRST classifier concept is located. This match is followed by a match on an initial specified in the INITIAL concept. When a word beginning with an uppercase letter follows this match, it is the coreference that is matched by all instances that occur in the document. Finally, a match on the PERSUFFIX concept is located.

In the example shown below, all instances of *Alito* are returned as a match. The PERSON concept also has a priority setting of `30`. This means that matches on the PERSON concept rank higher than the matches that are also returned to the FIRST and INITIAL definitions.

Figure 7-49 _ref and Export Symbol Matches



7.12 XML Fields and XPath Expressions

7.12.1 Overview of XML Fields and XPath Expressions

You can choose to specify either fields or XPath expressions in your rules. When you choose to match text in valid XML documents, you can specify the fields, which are also known as elements or nodes, to match. Use any of the following methods to specify these fields:

- Limit matches to the fields that you specify in the **XML Default Field** in the Project Settings - Misc window.
- Specify a field to match in the `CLASSIFIER`, `CONCEPT`, `C_CONCEPT`, `SEQUENCE`, `NO_BREAK`, or `REGEX` rules. Locate matches based on the constraints in multiple fields using either a `CONCEPT_RULE` or a `PREDICATE_RULE` definition. Specify the field name at the beginning of the pattern to be matched. For example, specify the `body` field as the location where all matches occur. The text that is present in other fields such as `link`, `title`, and `description`, cannot be matched.
- XML documents are treated as trees of nodes. The default behavior for XML documents is that the sections that have the same tag names are conflated into one searchable section. By merging multiple sections of the same type, SAS Enterprise Content Categorization Studio optimizes the matching function for rules that use Boolean operators such as `DIST` and `PAR`. The following sections explain how to write a rule that delimits a match on a specific term to the specified field.
- You can also combine the fields specified in the **XML Default Field** with the field, or fields, which you specify in your rules. When you combine these specifications, you enable some definitions to match the text in the default fields. When you specify an XML field in a rule, this field overrides any fields specified in the **Project Settings - Misc** tab.
- XPath is used to navigate through the elements and attributes in an XML document. In XPath, there are these types of nodes: element, attribute, text, namespace, processing-instruction, comment, and document nodes.

Notes: XPath is preceded by an underscore followed by a forward slash (_/). XML syntax is preceded only by an underscore (_).
For information about the XML fields in the **Misc** tab, see the *SAS Content Categorization Studio: User's Guide*.
Matches are returned only if the matches are located within, and not across, fields.

Specify an XPath expression for greater flexibility in locating the exact match, or matches, which your organization requires. For example, specify that a match can occur only on one of several `title` fields.

Each of these specifications is explained in the following sections.

7.12.2 A Sample XML Document

See the following sample, XML document that can be understood as a tree of fields, or nodes. When you specify a default or a matching field, you choose to limit matches to the text located inside these XML tags. For greater flexibility, choose to specify an XPath expression using the rule syntax explained in Section 7.12.3 *SEQUENCE Rules with an XML Field* on page 216. Also see Section 7.12.4 *Matching More Than One XML Field* on page 217.

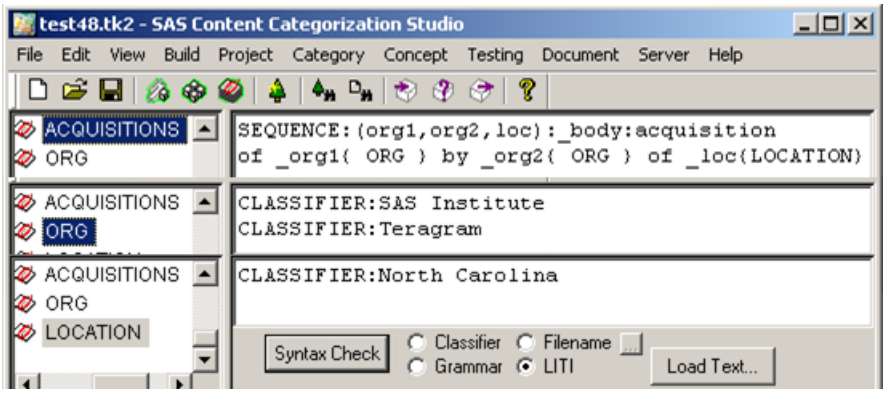
Example 7-29: Sample XML Document

```
<books>
<book number="1">
      <title>Improving Web Site Usability</title>
      <author>Millicent Marigold</author>
      <author>Montana Marigold</author>
      <price>25.99</price>
      <subjects lang="en">
         <subject>Usability testing</subject>
         <subject>Web site development</subject>
      </subjects>
</book>
<book number="2">
      <title>Usability Basics</title>
      <publisher>Ersatz Publications</publisher>
      <price>174.00</price>
      <subjects lang="en">
         <subject>Usability testing</subject>
         <subject>Web site development</subject>
         <subject>Guides and finding aids</subject>
      </subjects>
      </book>
<book number="3">
      <title> Usabilityguy Manuscript Guide </title>
      <author>Millicent Marigold</author>
      <author>Morty Marigold</author>
      <publisher>Ersatz Manuscript Library</publisher>
      <price>21.49</price>
      <subjects lang="en">
         <subject>Computers</subject>
         <subject>Software evaluation</subject>
         <subject>Usability testing</subject>
      </subjects>
</book>
</books>
```

7.12.3 SEQUENCE Rules with an XML Field

When you write a SEQUENCE rule, all of the individual tokens or concepts are matched. These matches occur if all of the tokens and concepts are present within the specified field. SEQUENCE rules do not enable matching across fields.

Figure 7-50 Body XML Field Specified in a Rule



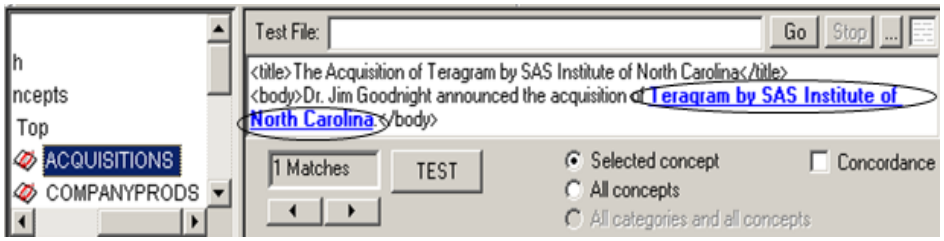
XML field specifications are each preceded by an underscore (_) and followed by a colon (:). In the SEQUENCE rule example above, there are several arguments. A match occurs when each of these arguments is matched in the body field of an input XML document.

Example 7-30: Assigning a New Concept Name to a Coreference Match

Concept Name	Entry
ORG	CLASSIFIER:SAS Institute CLASSIFIER:Teragram
LOCATION	CLASSIFIER:North Carolina
ACQUISITIONS	SEQUENCE:(org1,org2,loc):_body: acquisition of _org1{ ORG } by _org2{ ORG } of _loc{LOCATION}

A match for the ACQUISITIONS concept occurs when the term *acquisition of* occurs followed by two matches on the ORG concept separated by the word *by*. This match is complete when it is followed by a match on the LOCATION concept and all of these matches occur in the body field.

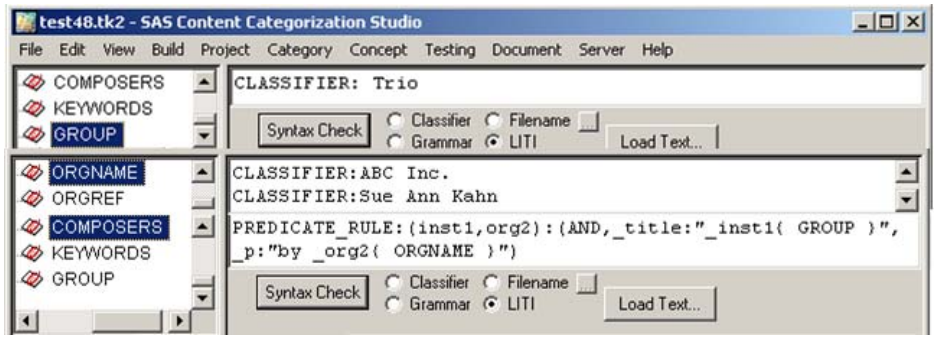
Figure 7-51 Match Located in an XML Field



7.12.4 Matching More Than One XML Field

If you choose to use a `PREDICATE_RULE`, `CONCEPT_RULE`, or a `REMOVE_ITEM` definition, you can specify a separate field for each argument.

Figure 7-52 A Predicate Rule Specifying XML Fields



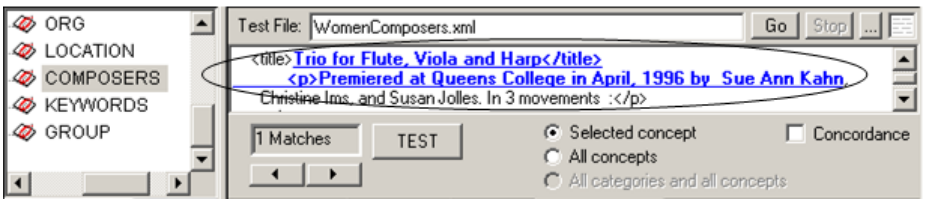
Each XML field is preceded by an underscore (_). For example, type `_title` and _p. The specified matches are enclosed in quotation marks (“”). See the following example:

Example 7-31: Matching XML Fields

Concept Name	Entry
GROUP	CLASSIFIER:Trio
ORGNAME	CLASSIFIER:ABC Inc. CLASSIFIER:Sue Ann Kahn
COMPOSERS	PREDICATE_RULE:(inst1,org2):(AND, _title:"_inst1{ GROUP }",_p:"by _org2{ ORGNAME }")

A match for the COMPOSERS concept occurs when there is a match in the `title` field on the GROUP concept. The match is complete only when there is also a match on the `p` (paragraph) field on the word *by* followed by a match on the ORGNAME concept.

Figure 7-53 Predicate Rule Match in XML Document



7.12.5 Specifying XPath Expressions

7.12.5.A Overview of Specifying XPath Expressions

Use XPath expressions to navigate elements (which are also known as *fields*) and attributes in valid XML documents. Write a rule using XPath expressions for greater flexibility in choosing where to locate matching text. Specify the specific XML field, or fields, to limit matches to this text. For example, refer to the first or last field for elements with the same name, or choose all of the subject and child nodes.

You use the same syntax to specify XPath expressions that you use to specify XML field names:

- For rules that do not specify arguments, place the path, preceded by an underscore (_) and followed by a colon (:) before the rule syntax. For example, type `_//book[2]/author:` followed by the syntax of your rule. This syntax locates any matches that occur in the `author` field of the second `book` element in a `.xml` document.
- For rules that specify arguments, insert the XPath expression before the argument. See the following examples:

```
PREDICATE_RULE:(drug,manufacturer):(DIST_20, _/books/
        book/title:"_drug{ DrugName }", _/books/book/
        title:"_manufacturer{ DRUGCOMPANY }", _/books/
        book/title:"make")
CONCEPT_RULE:(AND, _/books/book/title:"_c{ NFLORGS }",
        _/books/book/title:"NFLKEYWD")
REMOVE_ITEM:(ALIGNED, _/books/book/title:"_c{PERSON}",
        _/books/book/title:"LOCCMPND")
SEQUENCE:(drug,company): _/books/book/title:_drug{
        DrugName } _company{ DRUG_COMPANY }
```

Note: Some of the XPath expressions, like `ancestor`, preceding operators, `.`, and `..` are not supported at this time. In SAS Enterprise Content Categorization Studio, XPath expressions are used to locate matching text in XML elements as an alternative to traversing the XML tree.

7.12.5.B XPath Syntax for Contextual Extraction Definitions

Use the following table to understand the XPath expression syntax that is available for SAS Enterprise Content Categorization Studio. The examples in this table refer to the XML document displayed in Example 7-29 on page 215.

Table 7-3: XPath Syntax for SAS Enterprise Content Categorization Studio

XPath Expression	Description	Example
`/elem_name {/ elem_name}*` **Note**: The forward slash (/) is preceded by an underscore (_).	Specify the path from a root node.	`/books/book/title` can match text in any of the `title` elements.
`//elem_name {/ elem_name}*` **Note**: The forward slashes (//) are preceded by an underscore (_).	Specify the path from an internal node.	`//subjects` can match text in any of the `subject` elements and their children.
`@`	Match text based on attribute name.	`//subjects[@lang46]` can match text in a `subjects` element that has the attribute `lang`.
`*`	Match any element.	`//books/*` can match text in any of the elements under `books`.
`[[0-9]+]`	Matches elements by index location.	`//books/book[2]` can match text in the second `book` element.
`[@attribute_name='value']`	Match elements based on their attribute values.	`//book[@number='3']` can match text in a `book` element where the attribute `number` has a value of `3`.
`[elem_name cmp_op value]`	Match elements conditioned on the child elements value.	`/books/book[price>50.00]` can match text in a `book` element where the `price` attribute is more than `50.00`. **Notes**: No match is highlighted in the Document window when you specify `0` or `1`. You can specify a negative number such as `-25.00`.
`[last() | first()]`	Match the first or last element.	`/books/book[last()]` can match text in the last `book` element.

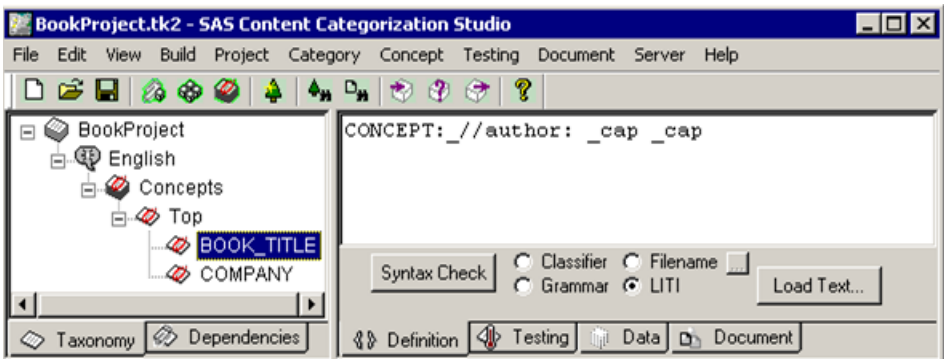
Table 7-3: XPath Syntax for SAS Enterprise Content Categorization Studio

XPath Expression	Description	Example
`[last() | first() +|-[0-9]+]`	Specify the location in which to match text from the first to the last element.	`/books/book[last()-2]` can match text in the second `book` element as counted from the last `book` element.
`[position() <|>|=|>=|<=[0-9]+]`	Select a specified group of elements.	`/books/book[position()>2]` can match text in any of the `book` elements whose index value is greater than `2`.
Notes: In XPath expressions the index begins with `1`, not `0`. Make sure that you do not use `//` slashes in the middle of an XPath element. For example, do not use this rule: `(OR, _/bookstore//book[99]:"Harry Potter")`. This is an unreported syntax error. The XPath wildcard `node()` is not supported.		

7.12.5.C Writing XPath Expression Rules

See the following examples of XPath expressions used in rules. These rules use the XML document displayed in Example 7-29 on page 215.

Display 7-8 An XPath Expression in a CONCEPT Rule



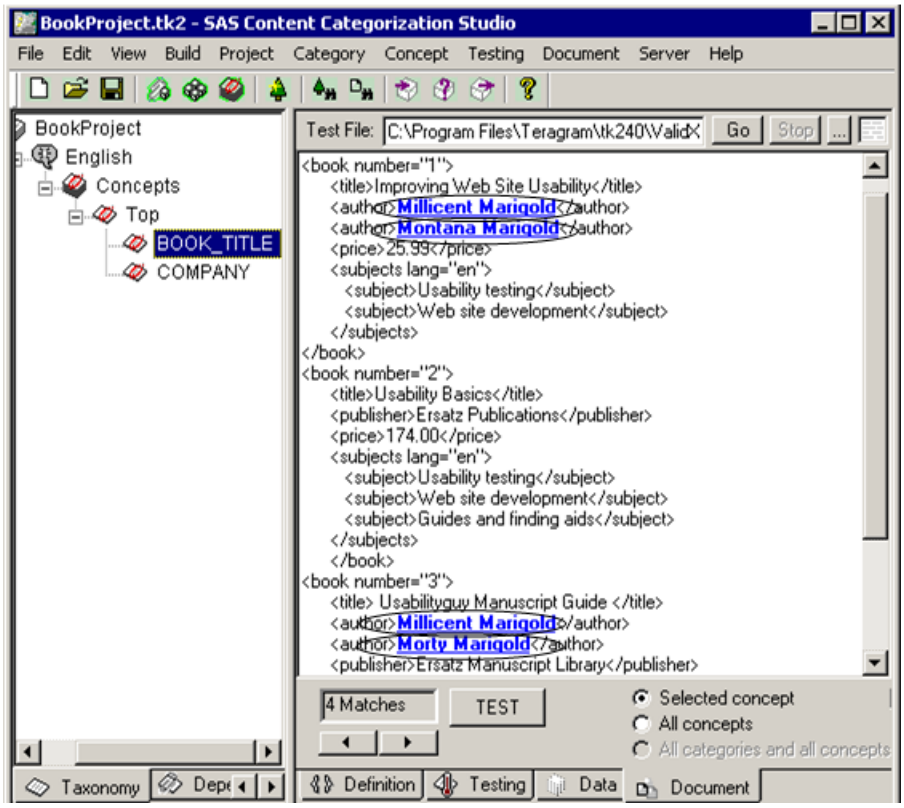
Each XPath expression is preceded by an underscore (_) and followed by a colon (:). For example, `_//author:`, which matches only text in the `author` fields of the input `.xml` document. This expression precedes the body of the `CONCEPT` rule, which in this example is `_cap _cap`:

Example 7-32: Matching an XPath Expression

Concept Name	**Entry**
`BOOK_TITLE`	`CONCEPT:_//author: _cap _cap`

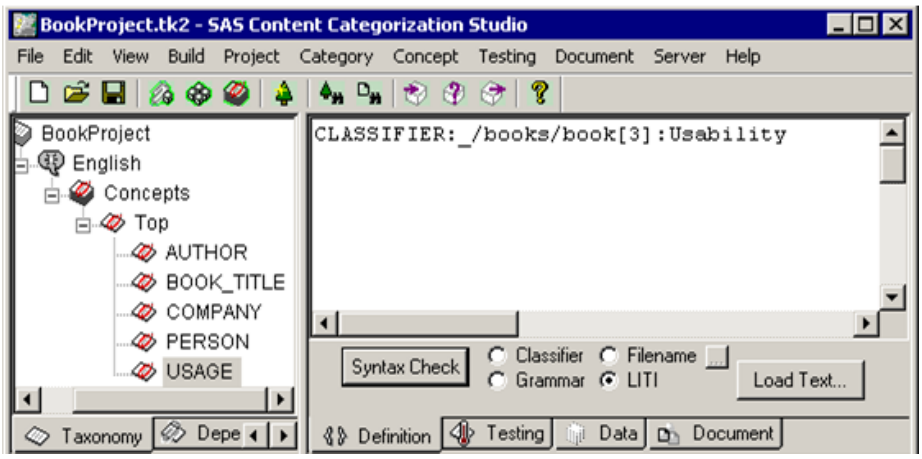
A match for the `BOOK_TITLE` concept occurs when there is a match on two words beginning with an uppercase letter. These matches can occur in any of the `author` fields in the input `.xml` document.

Figure 7-54 An XPath Expression Match for a CONCEPT Rule



See the `CLASSIFIER` example below:

Display 7-9 An XPath Expression in a CLASSIFIER Rule



In this example, `_/books/book[3]:` matches only the word `Usability` in the third `book` field of the input `.xml` document:

Example 7-33: Matching an XPath Expression

Concept Name	**Entry**
USAGE	`CONCEPT:_/books/book[3]:Usability`

The word `Usability` can be matched in any of the child elements of the third `book` element. See the example shown below:

Figure 7-55 An XPath Expression Match for a CLASSIFER Rule

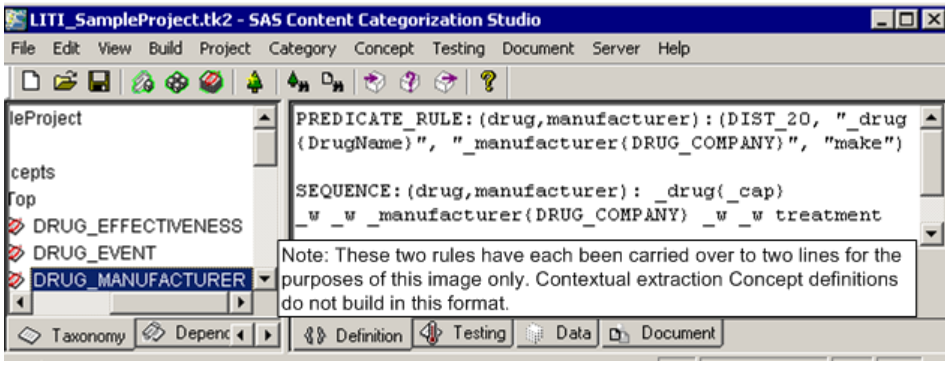
Note: If the word *Usability* occurs in lowercase in the input document, no matches occur. Specify case-insensitive rule matches in the Data window in order to change the case-sensitive default setting.

7.13 Writing Multiple Rules for One Definition

Write multiple rules for each contextual extraction concept. This feature increases the recall of your definitions by enabling you to locate more matches as well as matches based on different specifications.

For example, add the `SEQUENCE` rule shown in Example 7-20 on page 188 to the definition of the DRUG_MANUFACTURER concept to locate matches in documents that might not otherwise match.

Figure 7-56 PREDICATE_RULE with a SEQUENCE Rule



Reference Section

- Appendix A: *Troubleshooting on page 229*
- Appendix B: *Recommended Reading on page 237*
- Appendix C: *Glossary on page 239*

Appendix: A Troubleshooting

- *HTips and Guidelines*
- *HTips and Guidelines*
- *Known Issues*
- *Syntax Error Checking*

A.1 HTips and Guidelines

A.1.1 If You Do Not See the Match That You Expect in a Testing Document

If you do not see the matches that you expect in an input testing document, review the following:

Check Your Settings:

- Check the Data window to make sure that you have not selected **Test Disabled**.
- Did you specify an LITI concept in the Definition window using the **LITI** radio button?
- Open the Project Settings - LITI window and select **All Matches** under **Overlapping Matches**. Save your project, compile your concepts, and test again.
- Open the Document window. Select **All concepts** and click **TEST**.

Check your rule syntax

- Have you checked the rule syntax using the **Syntax Check** button in the **Definition** tab before compiling your concepts? Is this syntax appropriate for the results that you are trying to return, or is there a better syntax or rule type?

- Did you specify the correct rule type?
- Check your spellings in the rule definition, the input text, and any intermediate concepts that your concept references.
- Have you specified your rules to match the upper- and lowercase letters that you want to match? In other words, are your rules specified in a case-sensitive manner?
- Did you enclose the term that you want to return in curly braces (`{}`)?

Check your rule type:

`NO_BREAK` rule:

- Check `NO_BREAK` rules to see whether the match might be starting or ending in the middle of one of these rules.
- Did you specify that partial matches cannot be returned for a term for a `NO_BREAK` rule? If so, did you remember that this rule applies across the entire taxonomy?

`SEQUENCE` or `PREDICATE_RULE` rules:

- Make sure that arguments are specified,
- Are these arguments specified in lowercase letters only?

`REMOVE_ITEM` or `CONCEPT_RULE` rules:

- Use only one `_c` marker with each `REMOVE_ITEM` rule or with a `CONCEPT_RULE` that specifies a `NOT` operator.

`UNLESS` operator:

- When concepts are specified in a rule that uses the `UNLESS` operator, specify concepts that contain only `CLASSIFIER` or `REGEX` rules.

Check rules specifying coreference:

- Did you surround the new, or other, concept to be matched with square braces (`[]`) when you wrote a coreference rule?

Check part-of-speech tags, such as the following:

- `inc`: refers to an unknown term.
- `sep`: refers to punctuation.
- `digit`: refers to an actual sequence of numbers. For example, see 2, 5, and so on.
- `num`: refers to the English word for a digit. For example, see two, five, and so on.

A.1.2 Writing Concept Names

Concept names can consist only of alphanumeric characters and underscores (_) and form a single word. Use all uppercase letters in a concept name to distinguish the concept from a term that you want to match.

> **Note**: These conventions are particularly important if you are referring to another concept from within a `PREDICATE_RULE` or a `CONCEPT_RULE`.

A.1.3 Tokenization

The tokenizer does not return partial matches on a word or on a number that contains a decimal point. For example, if you specify `Pot` in a rule, a match is not returned if the document contains the word *Potter*.

A.1.4 Specifying a CLASSIFIER Definition

If you specify a `CLASSIFIER` definition, but you forget to change the rule type to `LITI` in the **Definition** tab, no syntax error is thrown and no matches are returned. Instead, SAS Content Categorization Studio attempts to match any occurrence of `CLASSIFIER: <specified term>`.

A.2 Known Issues

A.2.1 Remove Rule Types Added with Scroll Operation

When you add a rule type to your Definition window using the Ctrl and up arrow buttons, highlight the rule type and press Backspace or Delete twice. The rule type is removed only after the second Remove operation is performed. (It is not necessary to repeat the Remove operation twice if you enter the rule type manually.)

A.2.2 The Concept Priorities Window

If you write a rule that specifies a `PRIORITY` setting, this specification does not appear in the Concept Priorities window at this time. For this reason, the higher of the two numbers is the number used.

At this time, reverse sorting is not available for priorities when you click the headings in this window.

A.2.3 Index Position of Arguments

At this time, the index position of the arguments in the SAS Content Categorization Studio window is arranged in reverse order.

A.2.4 SAS Content Categorization Studio Windows

A.2.4.A Example One

If you click on the highlighted, matched text for a fact in the Document window, a SAS Content Categorization Studio appears. See the following example:

Figure A-1 SAS Content Categorization Studio Window

INFO

applies to the matched info string for a `CLASSIFIER` or a `REGEX` concept in SAS Content Categorization Studio, not to SAS Contextual Extraction Studio.

CANONICAL

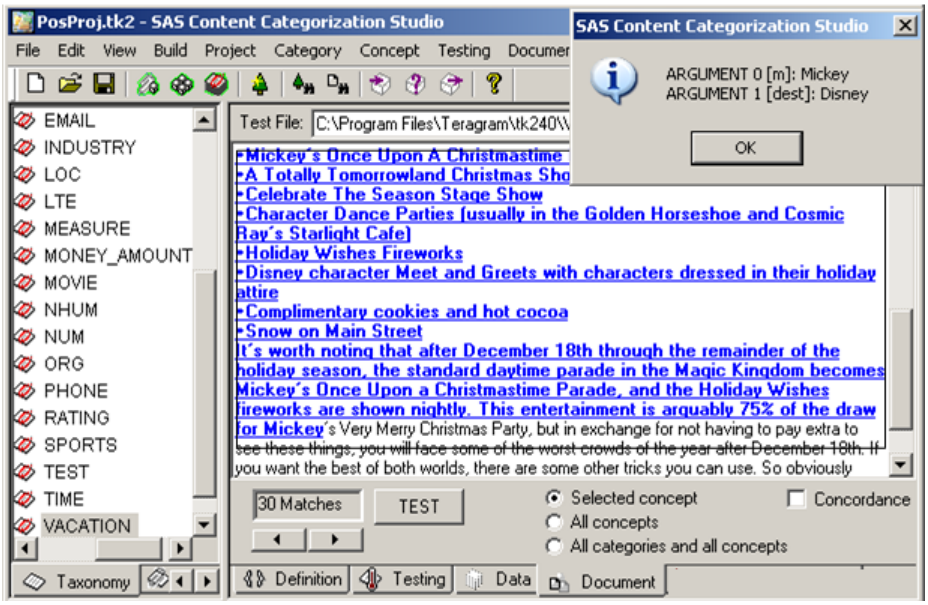
applies only to rules specifying coreference. However, results might appear for rules that do not specify coreference.

For these reasons, this information does not always apply to the matched text.

A.2.4.B Example Two

If you click on the highlighted, matched text in the Document window when there is a match on a rule with arguments, a SAS Content Categorization Studio appears. See the following example:

Figure A-2 NOT Operator Rule Matches in an Input Document



At this time, the index position of the arguments in the SAS Content Categorization Studio window are not specified correctly.

A.2.5 Export Testing Results to a SAS Data Set or a Microsoft Excel Spreadsheet

A.2.5.A Overview of Export Testing Results

The information in the following sections references Section 2.11 *The Export Results Wizard* on page 26 in this manual.

A.2.5.B Heading Report Clarifications

relevancy and **above_rel_cutoff**

There is no support for relevancy at this time.

date

Display the date and time of the export operation for each document. This output appears in SAS timestamp informat. The date and time is the same for every document in the output results.

info_or_fact_args

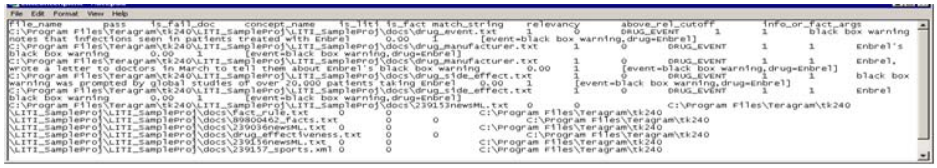
fact_args are the arguments that comprise the located facts. However, these arguments might cause incomplete, misaligned, or other defects in the output file. For information about **info_args**, see *SAS Content Categorization Studio: User's Guide*.

Note: If a testing document does not match a concept, this file might or might not, appear in the results.

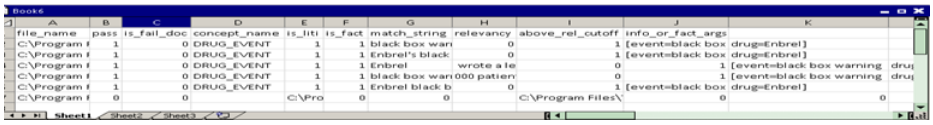
A.2.5.C If Your Notepad Results Look Inconsistent

If the results that you see do not align with the columns displayed, import your results into a *Microsoft Excel* spreadsheet. See the following example:

Display A-1 Results Displayed in Notepad



Display A-2 Results Displayed in Microsoft Excel



Note: In some cases, the results continue to be misaligned in *Microsoft Excel*. This statement is often true when

there are multiple matches for the match string and for facts, or when documents with no results are included.

A.3 Syntax Error Checking

Not all syntax errors are flagged for rules and definitions. For this reason, check your syntax carefully if you do not get the expected matching results.

For example, do not use `//` slashes in the middle of an XPath element such as `(OR, _/bookstore//book[99]:"Harry Potter")`. This is an unreported syntax error.

The XPath wild card `node()` is not supported.

XPath is preceded by an underscore followed by a forward slash (`_/`). XML syntax is preceded only by an underscore (`_`). (If you are specifying a path to a relative node, use two forward slashes instead of one.)

Appendix: B
Recommended Reading

The following books are recommended:

- *SAS Content Categorization Studio: Installation Guide*: Install SAS Content Categorization Studio.
- *SAS Content Categorization Studio: User's Guide*: Create a SAS Content Categorization Studio project, test, and upload the output to SAS Content Categorization Server.
- *SAS Enterprise Content Categorization Studio: Administrator's Guide*: Install and configure the server used for the collaborative operations available in SAS Enterprise Content Categorization Studio.
- *SAS Enterprise Content Categorization Servers: User's Guide*: Install, configure, and use SAS Content Categorization Server, SAS Content Categorization Collaborative Server, and SAS Document Conversion Server. Upload `.li` files using this product.

SAS offers instructor-led training and self-paced e-learning courses to help you get started with the SAS add-in, learn how the SAS add-in works with the other products in the SAS Enterprise Intelligence Platform, and learn how to run stored processes in the SAS add-in.
For more information about the courses available, see `support.sas.com/training`.

For a complete list of SAS publications, see the current SAS Publishing Catalog. To order the most current publications or to receive a free copy of the catalog, contact a SAS representative at

SAS Publishing Sales
SAS Campus Drive
Cary, NC 27513
Telephone: (800) 727-3228*
Fax: (919) 677-8166
E-mail: sasbook@sas.com
Web address:`support.sas.com/pubs`
* For other SAS Institute business, call (919) 677-8000.

Customers outside the United States should contact their local SAS office.

Appendix: C
Glossary

_c

specifies the context for the matches for some LITI rules.

_cap

specifies that a word beginning with an uppercase letter is a match on a LITI rule.

argument

is defined by two or more concepts that are related to each other. When these matches are identified, arguments are returned. See *fact*.

branch

refers to either the category, or the concepts, section of the taxonomy tree. The first node in a branch is either the `Categorizer` or the `Concepts` node. If the project is built with more than one language, each language section is also referred to as a branch.

canonical form

specifies the full name, or form, of the term. For example, SAS Institute Inc. is the canonical form of SAS.

categorization

defines the subject matter of a document, in other words, the main idea or subject of the document.

CLASSIFIER

specifies the terms to be matched. This LITI rule works like a SAS Content Categorization Studio classifier definition because this LITI rule specifies a simple list of terms to be matched.

classifiers

specify a list-based set of terms that are extracted from your documents.

concept

define an autonomous piece of information such as movie, book, title, and so on. Also see *Entity.*

collaboration

work of two, or more, subject matter experts working together on one project. This project is located on a server with a cached copy in the shared projects folder on a local machine.

concordance

displays a list of the matching terms located in a document with the text surrounding them. Specify the number of characters or words that are returned for a match on a concept.

coreference

refers to pronoun resolution. A pronoun is matched to the antecedent that it refers to. Coreference is also known as *anaphora resolution*.

definition

defines a concept, whether the definition consists of one or more rules. *Definition* is used interchangeably with the word *rule*. See *rule*.

document

refers to a printable page. Also see *text*.

event

is used interchangeably with *fact*. See *fact*.

Fact

refers to two or more concepts or tokens that are specified in one `SEQUENCE` or `PREDICATE_RULE` definition. See *SEQUENCE* and *PREDICATE_RULE* below.

node

refers to the visual representation of a concept or a category. Less frequently, this term is used to refer to the `Categorizer`, `Concepts`, `Top`, or another component of the taxonomy tree.

precision

is a measurement of the relevancy of the matched documents. In other words, the concept definition excludes possible matches that do not reflect the subject matter of the concept. For example, texts referring to *rock collections* are not matched for the category *Rock and Roll*.

PREDICATE_RULE

returns matches when an operator is specified with arguments. Unlike the `SEQUENCE` definition, the matches do not need to occur in the order specified by the definition.

priority

ranks concepts. By default, priority is set to `10` in the Data pane for LITI concepts. This specification prioritizes LITI concepts over classifier and grammar concepts.

recall

is a measurement of how well the definition matches all of the relevant texts.

referring term

is a term that refers to a canonical form.

REGEX

specifies regular expression syntax.

regular user

refers to a collaborative user with one of several permission levels for a project. However, this user lacks either database or administrative privileges. The regular developer is added to the project and his or her permission level is set by either the database, or the project, administrator.

remote project

resides on a server instead of your local machine.

rule

refers to a category or a concept definition. There can be many rules for each LITI concept definition. This term *rule* is used interchangeably with *definition*, but properly speaking, one definition can contain many rules. See *definition*.

SEQUENCE

returns facts when matches occur within the specified context.

shared projects folder

stores a cached copy of the project that is located on the server on your local machine. Individual cached copies of the project can be stored on one machine using different shared projects folders.

string

refers to a group of words or characters that you specify for a rule.

syntax changes

edits to a category rule or to a concept definition are defined as changes to the syntax of the rule or definition.

taxonomy

organizes a classification structure that can be either a flat or a hierarchical system.

text

forms a written document, or a Web page. Also see *Document*.

token

is a synonym for a word. *Token* is not a synonym for the word *string* that can refer to several words or characters. *Token* refers only to one word.

XML field

refers to an element, or a node, in a valid XML document.

XPath expression

use XPath expressions for greater flexibility when you want to specify matches on XML fields in valid XML documents.

Index

A

B

C

D

E

F

H

I

L

M

N

O

P

Q

R

S

T

U

V