IBM

LanguageWare Resource Workbench 7.2
Create a parsing rules database

Introduction

- **Module overview**
  - What is a rules database
  - How to create and configure Rules DB
  - Best practices

- **Target audience:**
  - All audiences

- **Prerequisites:**
  - Install LanguageWare® Resource Workbench
  - Create a LanguageWare Project

- **Version release date** LRW 7.2, ICA 2.2, released October, 2010

Create a parsing rules database

Module objectives

After this module you will be able to:

- Create and configure a rules database

- Understand what a rules database is for

- Start learning how to create rules.

Create a parsing rules database

Module roadmap

- **Parsing rules database**

    What is it?

    How to create and configure it?

    How to export and import data from and into a parsing rules database.
- Summary and best practices
- Sample exercises

Create a parsing rules database

What is a parsing rules database?

- A (parsing) rules database stores the parsing rules.
- Parsing rules generate annotations (output types) based on input types coming from different resources: Dictionaries (lexical and custom), other parsing rules, character rules.
- The rules database is compiled into a jar file that will be run inside the pipeline.
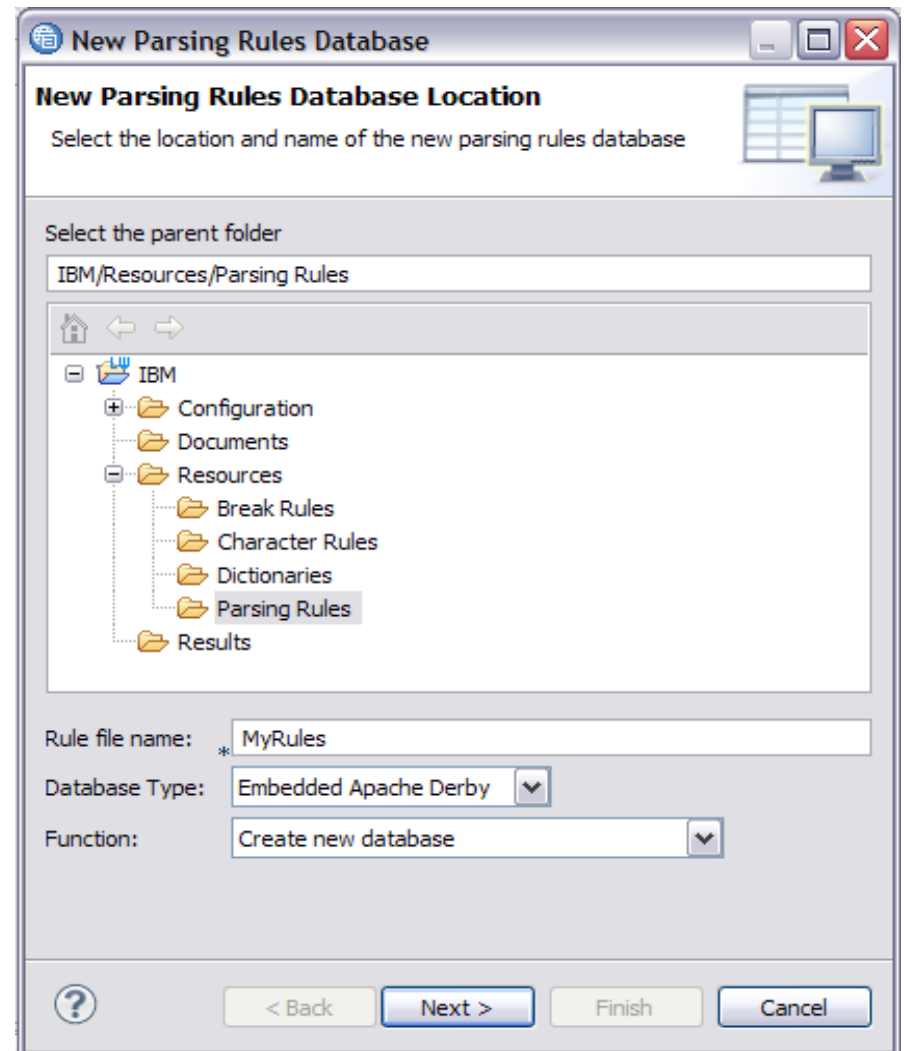
Create a parsing rules database

Rules database
How to create and configure it?

- *Right click on the Parsing Rules folder in the Resources directory, and select New/Parsing Rules Database.*
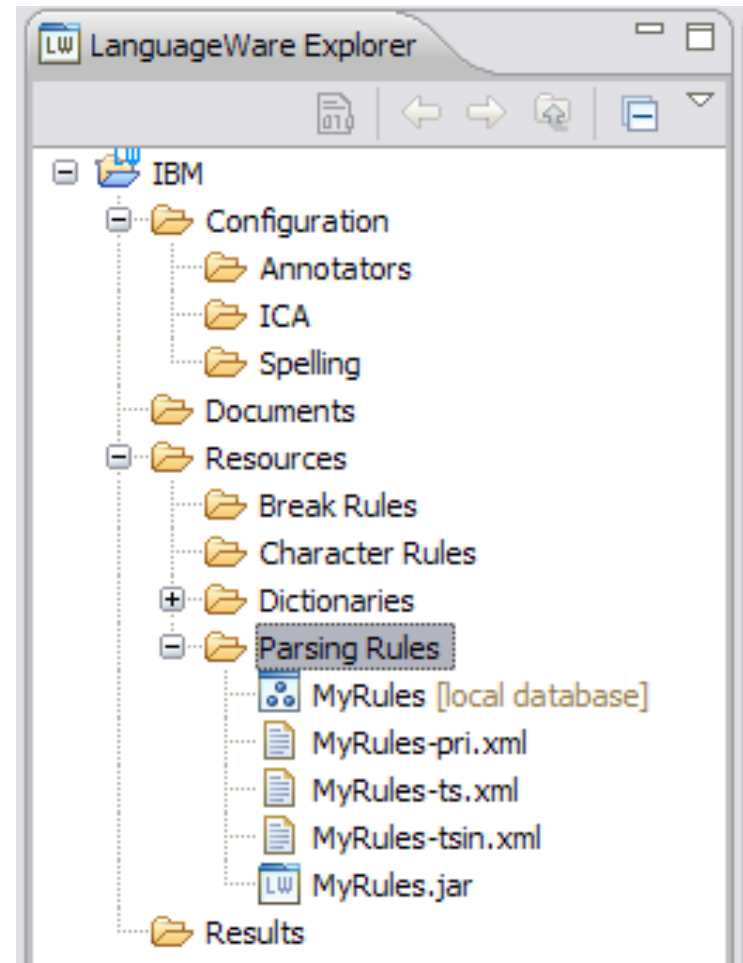
Create a parsing rules database
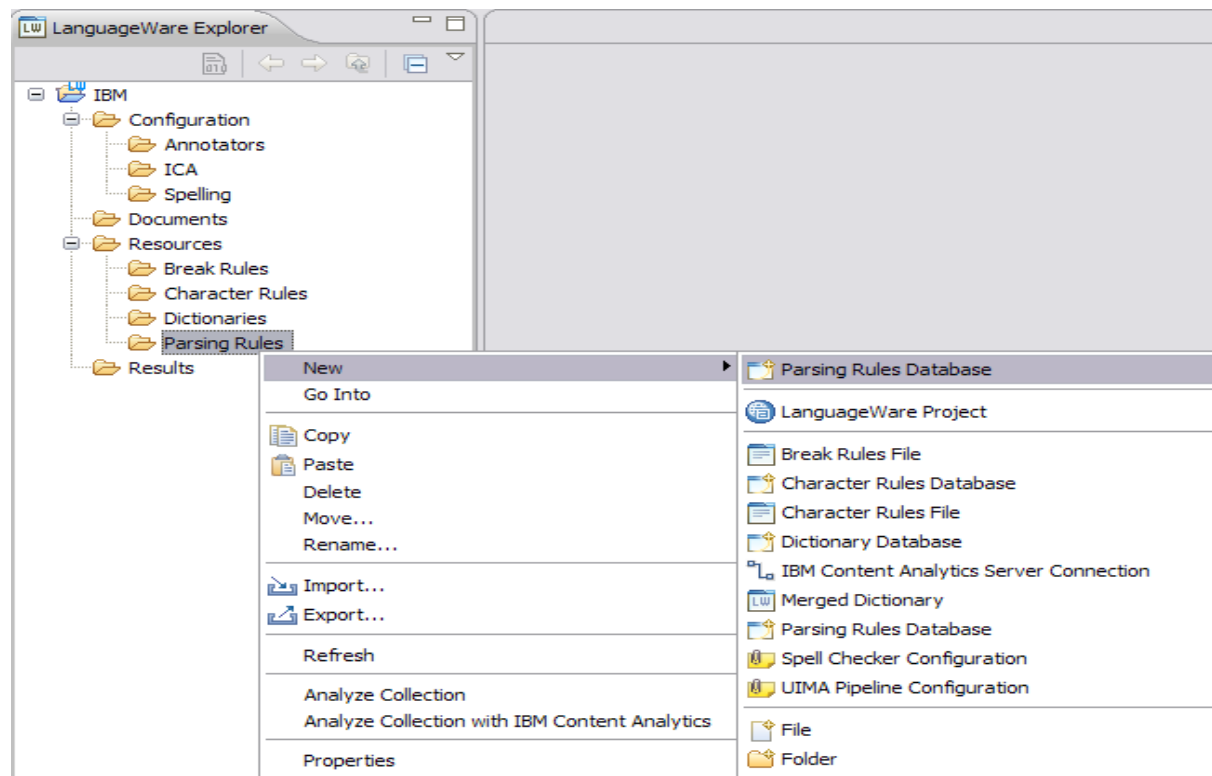
- The Create Rules Database dialog box pops up:
  - Specify the parent folder
  - Specify the name of the Rules Database. It's good to use relevant names (specific to the the rules, for example Person, or relevant to the model, for example Healthcare).
  - Database type: Embedded Apache derby if it is a local DB
  - Function: Create a new database. If you are creating a database from an export file, select the option "create database from database export file"

- Click Next.

New Parsing Rules Database

New Parsing Rules Database Location
Select the location and name of the new parsing rules database

Select the parent folder

IBM/Resources/Parsing Rules

- IBM
  - Configuration
  - Documents
  - Resources
    - Break Rules
    - Character Rules
    - Dictionaries
    - Parsing Rules
  - Results

Rule file name: * MyRules
Database Type: Embedded Apache Derby
Function: Create new database

< Back   Next >   Finish   Cancel

Create a parsing rules database

- The Parsing rules will be compiled into a .jar file. Other files (such -ts.xm and tsin.xml are also generated).
  - UIMA type prefix: the default UIMA prefix of the project.
  - Rule file: the .jar file that will be used in the annotator.
  - Version: the version of the jar. It will be increased every time the rules database is compiled.
  - Copyright: copyright text of the database.

- Click finish to create the database and the related files.

New Parsing Rules Database

Parsing Rules Build
Parsing Rules Build

UIMA type prefix:	com.ibm
Rule file:	*/IBM/Resources/Parsing Rules/MyRules.jar [Select...]
Version:	1.0.0.0
Copyright:	

[< Back] [Next >] [Finish] [Cancel]
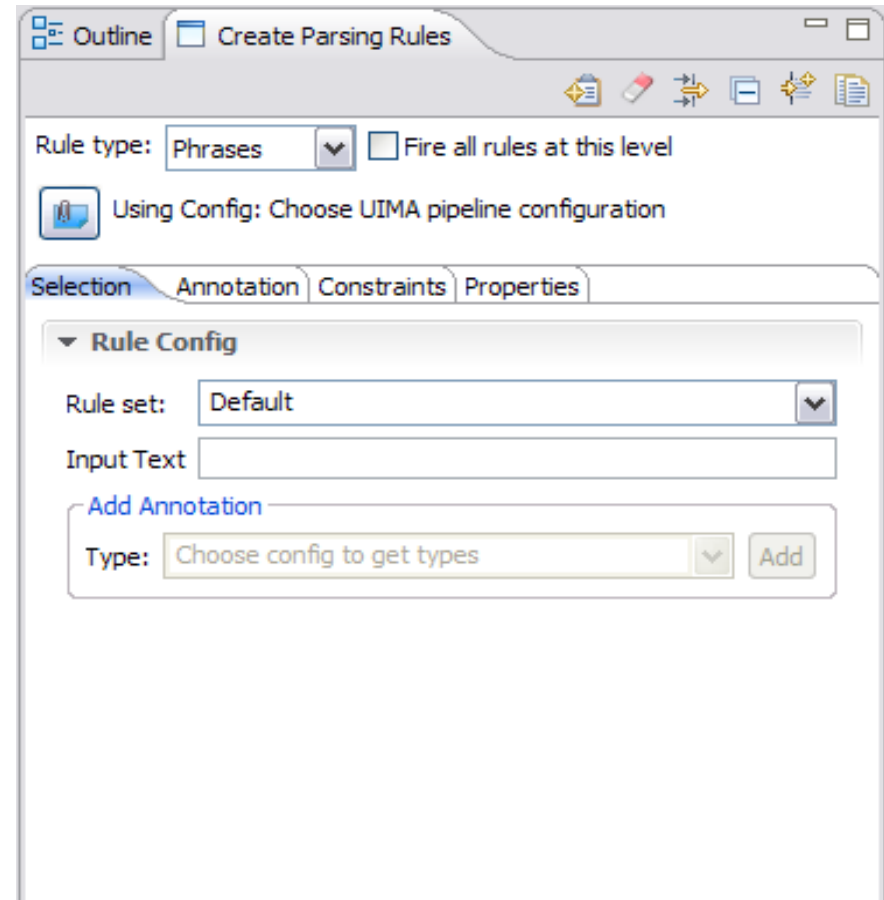
Create a parsing rules database

- These are the files that are generated when you create a new parsing rules database.
  - **Local database**: the rules database.
  - **-pri.xml**: Type priority file. This specifies the relative priorities of Types in UIMA, as required by this Parsing Rule file
  - **-ts.xml**: The output TypeSystem file for this Parsing Rule file. This defines the Types and associated features of all Types created by the rules in this file
  - **-tsin.xml**: The input TypeSystem file for this Parsing Rule file. This defined the Types that are required as input by the rules in this file. These Types must be generated by an earlier stage in the UIMA pipeline.
  - **.jar**: The Parsing Rule file. This contains the compiled rules used by the Parsing Rules annotator.

Opening a parsing rules database

- Double click on the rules database link, and two perspectives will open.
  - The Database Viewer
  - The Create Parsing Rules view

■ The **Database Viewer**: allows you to view the database columns

X	Set	Type	Label	Original Text	Created	Modified	Id

Properties | Problems | Rules

Create a parsing rules database

- The **Create Parsing Rules** view: this view is used to create parsing rules. This will be covered in following presentations.

Create a parsing rules database
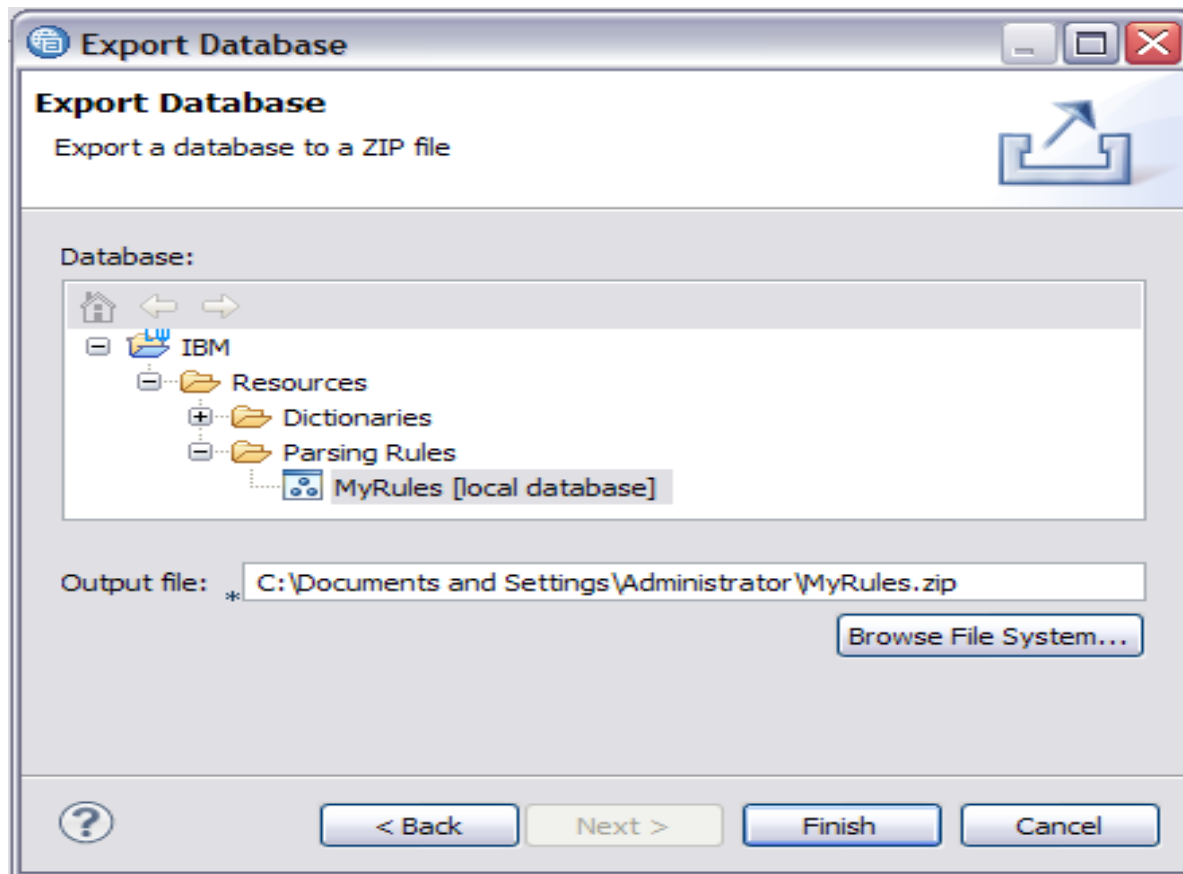
Exporting a parsing rules database

- Data is exported from the database in the form of a zip file, containing **.csv** files. The file containing the data is the **AFSTRULES.CSV**

- Make sure the database is closed (right click, select close),

- Right-click the database link in the LanguageWare Explorer and select export, a dialog box will pop up.

Create a parsing rules database
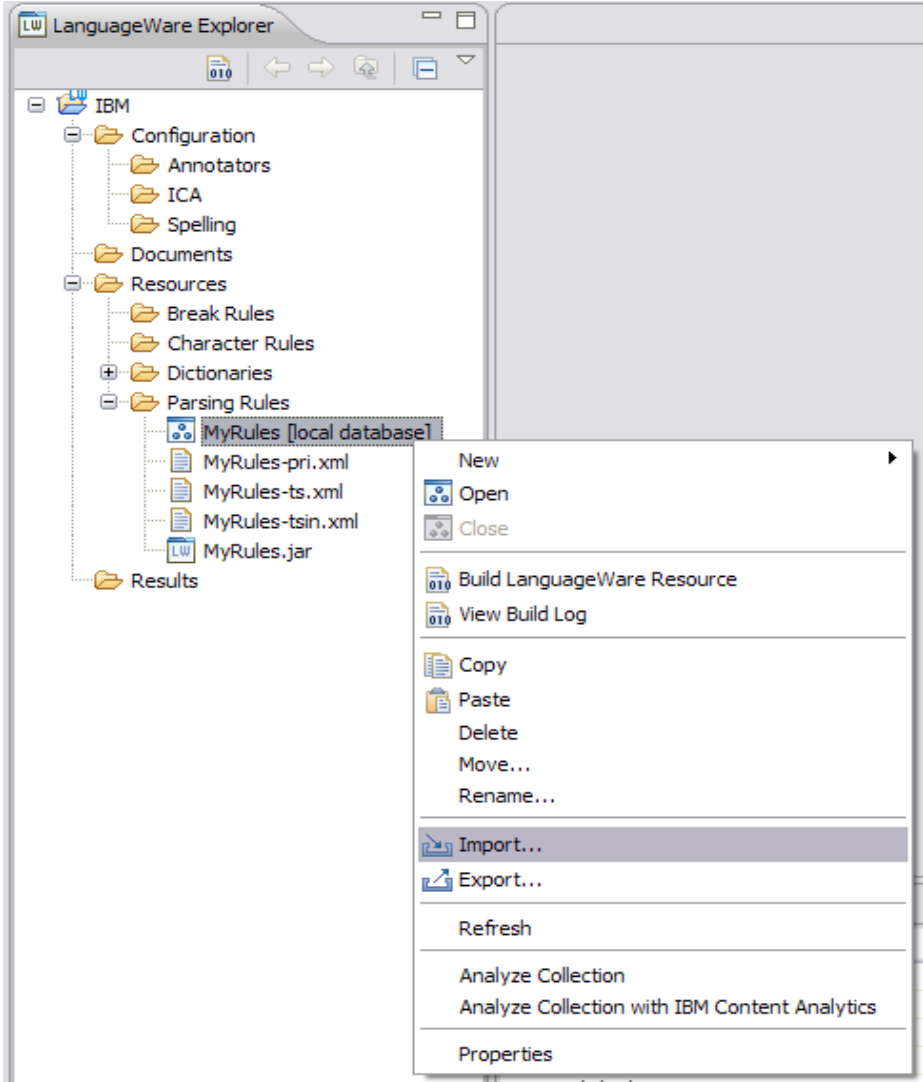
■ Select the "IBM LanguageWare/Export Database" option, then click Next

Create a parsing rules database
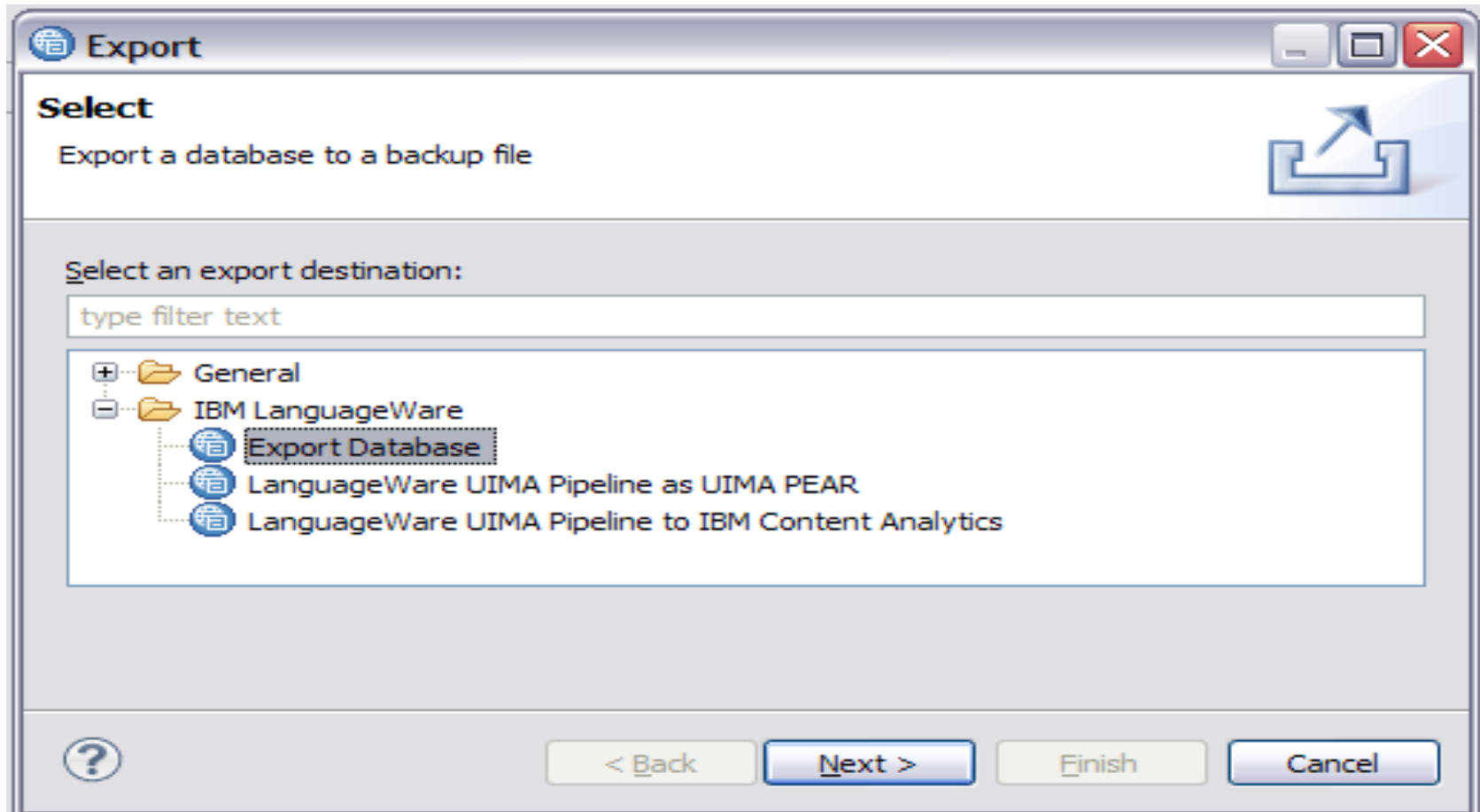
- Select the Database to be exported, select the output file path and name, then click Finish.

Create a parsing rules database

- If you open the exported .zip file, you can see the different **.csv** files. The **AFSTRULES.CSV** file contains the rules data. The remaining files are database tables and properties.

AFSTRULES.CSV Microsoft Excel Comma Separ... 1 KB	**databaselink.properties** PROPERTIES File 1 KB
LRWATTR-AFSTRULES.CSV Microsoft Excel Comma Separ... 2 KB	**LRW-TABLES.CSV** Microsoft Excel Comma Separ... 1 KB

Importing data into a parsing rules database

- Data can be imported into a database in the form of a **.csv** files (comma separated document).

- Make sure the database is closed (right click, select close),

- Right-click the database link in the LanguageWare Explorer and select import: a dialog box will pop up.
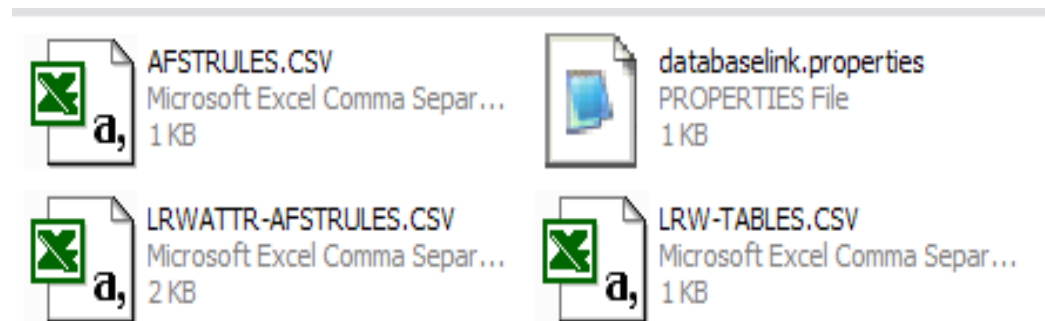
Create a parsing rules database © 2011 IBM Corporation
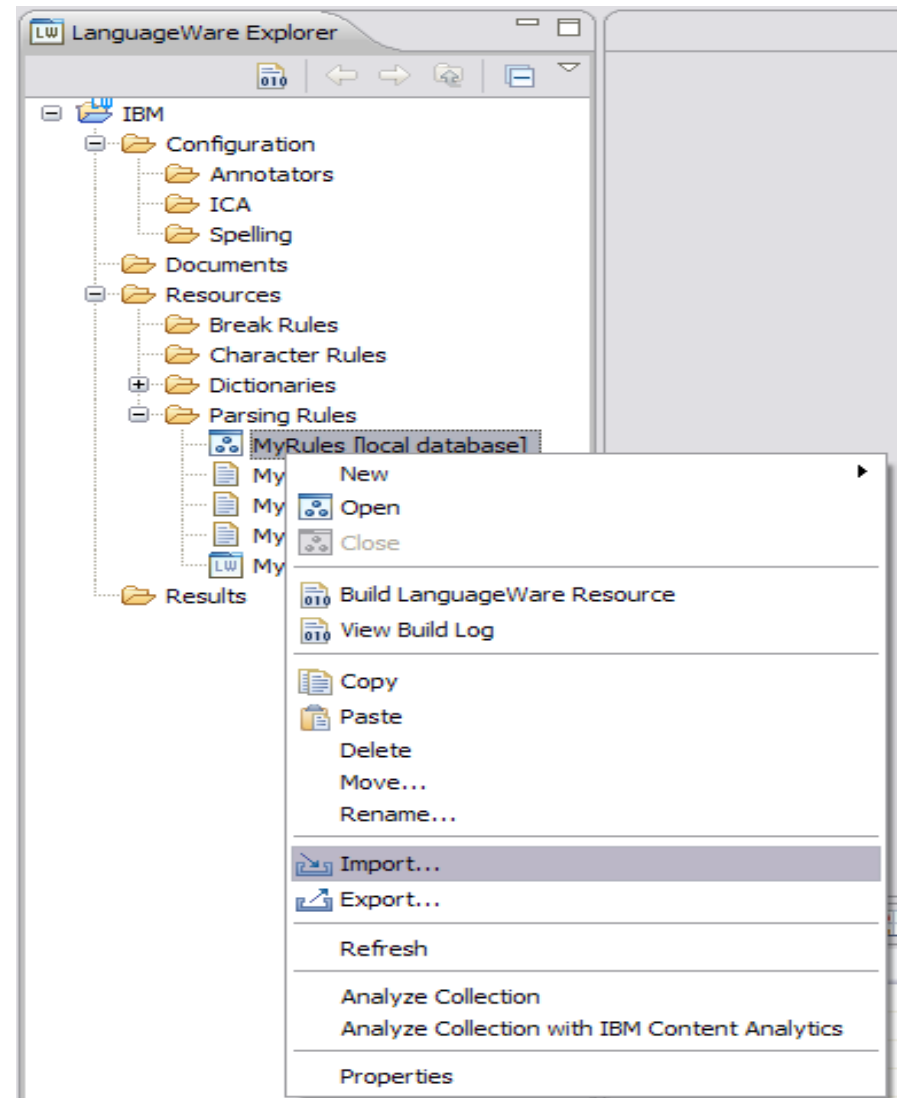
- Select the "IBM LanguageWare/Import into Database" option, then click Next

Create a parsing rules database
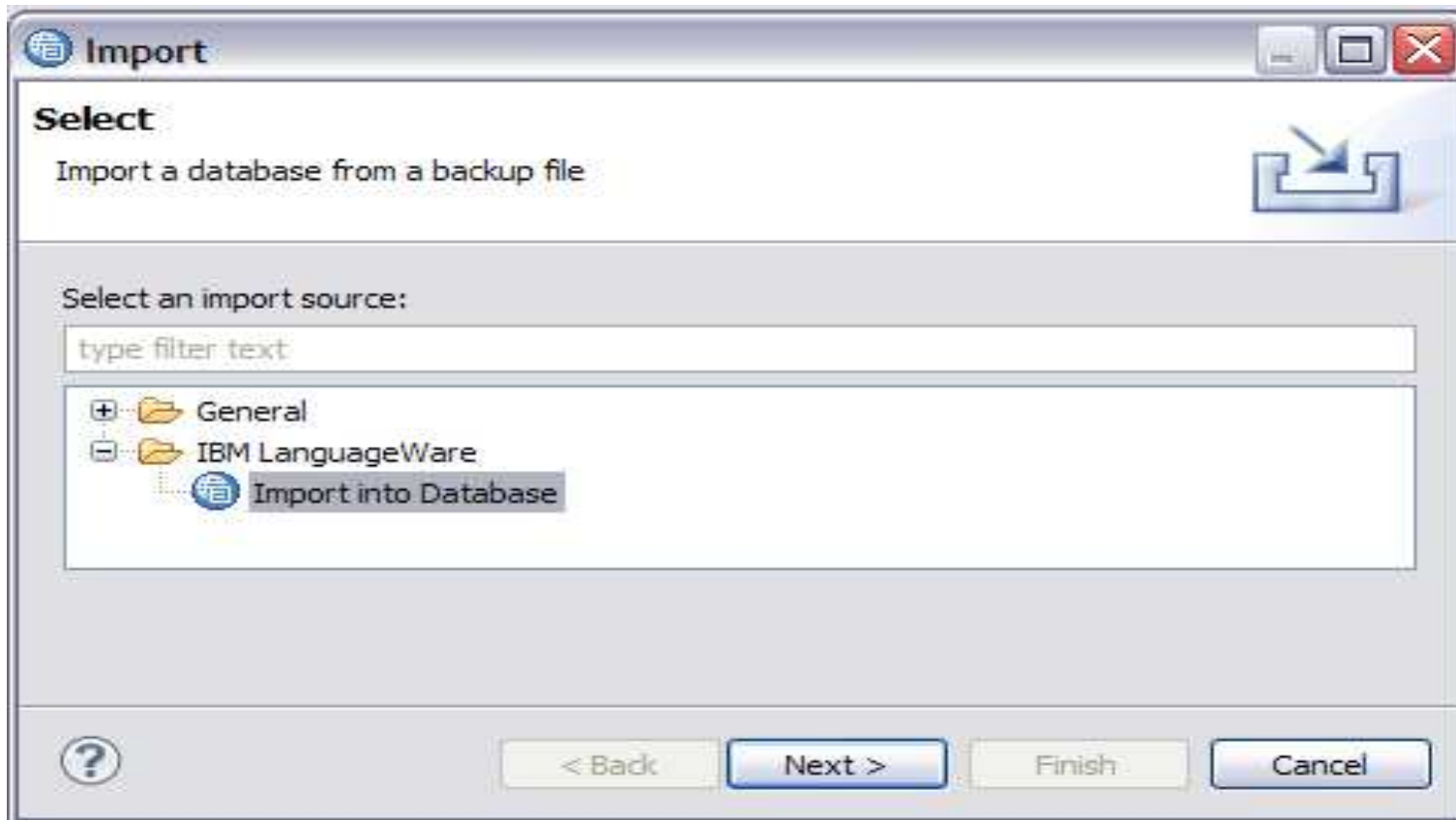
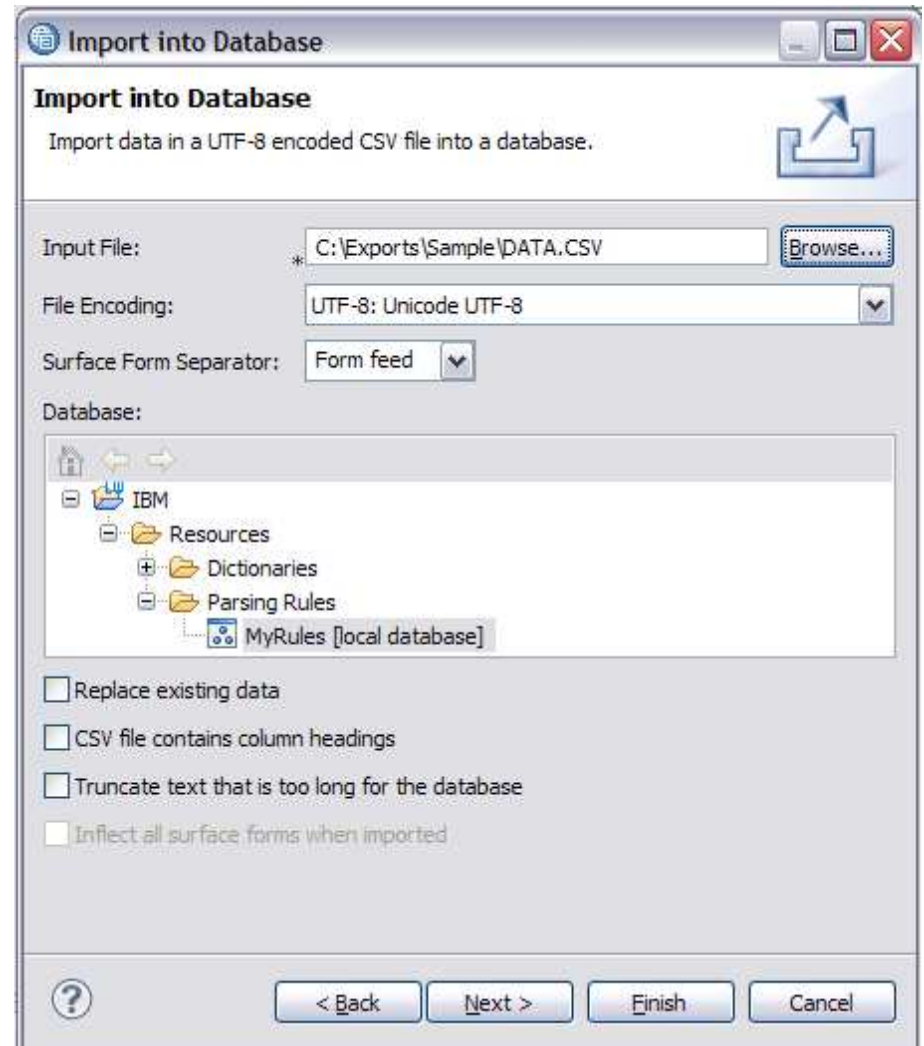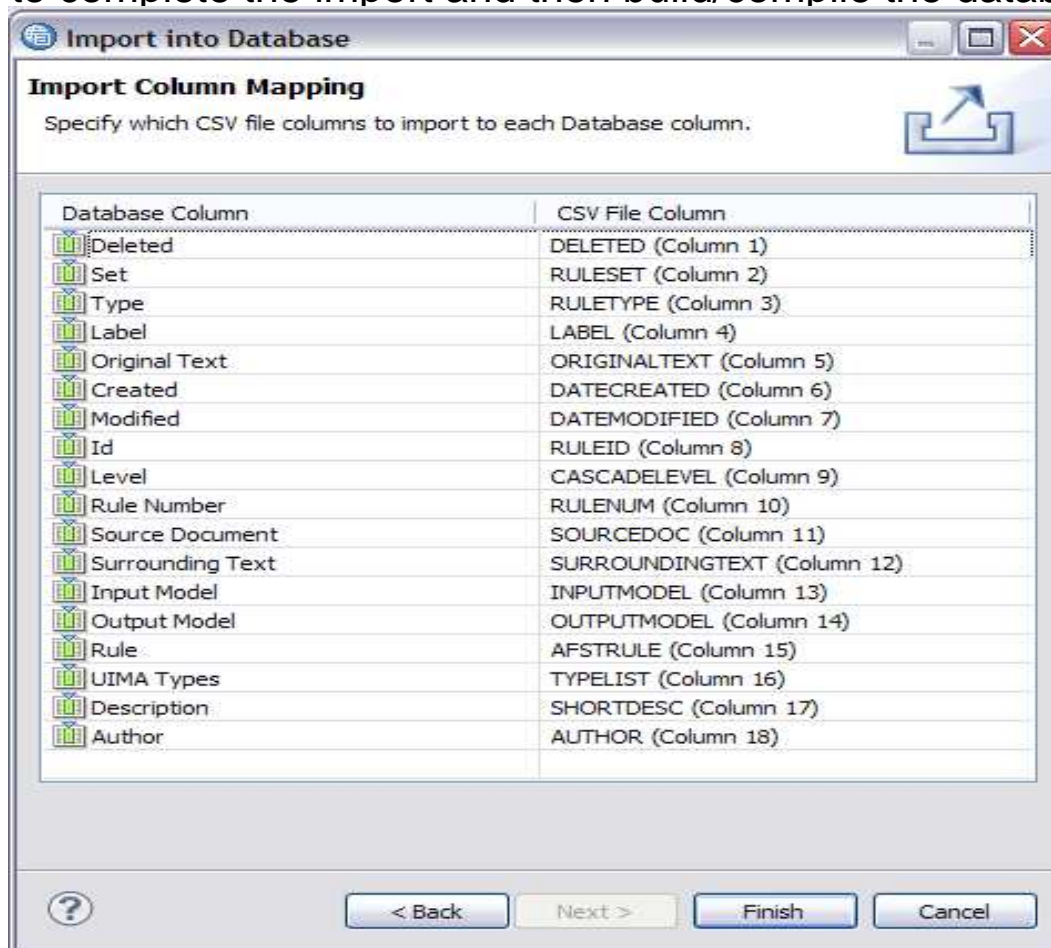- **Input File**: must be a *.csv* file.

- **File Encoding**: the default is UTF-8.

- **Surface form separator**: in the .csv file, the surface forms are in the same column as the lemma, but they are separated with a character different from the comma.

- **Database**: select the database you want to import the data into.

- **Replace existing data**: check this box if you want to overwrite the data in the database with the new imported data.

- **CSV contains column headings**: check this box if the *.csv* files contains headings (especially if it is an export from another custom database).

- **Truncate text that is too long for the database**: if you check this box, the entries that are too long for the default column size will be truncated.

- **Inflect surface forms when imported**: This option is disabled for Parsing Rules databases.

Import into Database

Import into Database

Import data in a UTF-8 encoded CSV file into a database.

Input File: *   C:\Exports\Sample\DATA.CSV   Browse...

File Encoding: UTF-8: Unicode UTF-8

Surface Form Separator: Form feed

Database:

- IBM
  - Resources
    - Dictionaries
    - Parsing Rules
      - MyRules [local database]

☐ Replace existing data

☐ CSV file contains column headings

☐ Truncate text that is too long for the database

☐ Inflect all surface forms when imported

< Back   Next >   Finish   Cancel

- This dialog box maps the columns of the export file to the columns in the database. You can change the mapping by changing the values in the "CSV File Column". Click on the value of the column and choose from the menu.

- Click Finish to complete the import and then build/compile the database.

Import into Database

Import Column Mapping

Specify which CSV file columns to import to each Database column.

Database Column	CSV File Column
Deleted	DELETED (Column 1)
Set	RULESET (Column 2)
Type	RULETYPE (Column 3)
Label	LABEL (Column 4)
Original Text	ORIGINALTEXT (Column 5)
Created	DATECREATED (Column 6)
Modified	DATEMODIFIED (Column 7)
Id	RULEID (Column 8)
Level	CASCADELEVEL (Column 9)
Rule Number	RULENUM (Column 10)
Source Document	SOURCEDOC (Column 11)
Surrounding Text	SURROUNDINGTEXT (Column 12)
Input Model	INPUTMODEL (Column 13)
Output Model	OUTPUTMODEL (Column 14)
Rule	AFSTRULE (Column 15)
UIMA Types	TYPELIST (Column 16)
Description	SHORTDESC (Column 17)
Author	AUTHOR (Column 18)

[< Back] [Next >] [Finish] [Cancel]

Create a parsing rules database

Module roadmap

- **Parsing rules database**

    What is it?

    How to create and configure it?

    How to export and import data from and into a parsing rules database.

- **Summary and best practices**

- **Sample exercises**

Create a parsing rules database

Module summary

You have completed this module and can:

- Create and configure a parsing rules database

- Import and export data from a parsing rules database

- Add a parsing rules file to a UIMA annotator configuration and start learning ow to create parsing rules.

Please refer to the LanguageWare help for more tips and advanced use cases.

Best practices

- A parsing rules database will contain the parsing rules created in a model.

- It is good practice to use explicit names when creating the parsing rules database as this will help know what it does, and it makes it reusable in different models (for example a Dates rules database will be specific to finding dates).

- It is also advisable to have separate rules databases for different (groups of) concepts/Types, this facilitates the maintenance, update and debugging of rules.

Create a parsing rules database

Module roadmap

- **Rules database**

    What is it?

    How to create and configure it?

    How to export and import data from and into a parsing rules database.
- **Summary and best practices**
- **Sample exercises**

Create a parsing rules database

Practice exercises

- Create a Rules DB called "IdentifyQuestion"

- Open the database in the Database viewer.

Create a parsing rules database

Contacts

- If you have any questions, comments or suggestions, contact us using the LanguageWare email address *EMEALAN@ie.ibm.com* or on the developerWorks® Forum.

Create a parsing rules database

Trademarks, copyrights, and disclaimers

IBM, the IBM logo, ibm.com, developerWorks, and LanguageWare are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of other IBM trademarks is available on the web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

Other company, product, or service names may be trademarks or service marks of others.

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS OR SOFTWARE.

© Copyright International Business Machines Corporation 2011. All rights reserved.

Create a parsing rules database