IBM Classification Module: Make It Work for You

Introduces the concept, architecture, tools, and integration

Describes building, training, and fine-tuning the knowledge base

Provides the steps to integrate with other products and solutions

Wei-Dong Zhu
Solomon Barron
Maurizio Gallotti
Vijay Gupta
Xiaomei Wang
Josemina Magdalen
Jane Singer

ibm.com/redbooks
International Technical Support Organization

IBM Classification Module: Make It Work for You

October 2009
First Edition (October 2009)

This edition applies to Version 8, Release 6, Modification 0 of IBM Classification Module (product number 5724-T45).

© Copyright International Business Machines Corporation 2009. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
# Table of Contents

**Figures** ......................................................... ix

**Notices** .......................................................... xix
**Trademarks** ....................................................... xx

**Preface** .................................................................. xxi
**The team who wrote this book** ................................ xxii
**Become a published author** ........................................ xxiv
**Comments welcome** ................................................. xxiv

## Part 1. Core product information .............................................. 1

### Chapter 1. Overview ...................................................... 3
1.1 Introducing Classification Module ..................................... 4
1.2 How Classification Module works ...................................... 5
1.2.1 Input and output .................................................. 6
1.2.2 Configuration process ............................................. 7
1.3 Business use cases ..................................................... 9
1.3.1 Achieving and maintaining regulatory compliance ............ 10
1.3.2 E-mail management .............................................. 11
1.3.3 Customer communication management .......................... 12
1.3.4 Enterprise content management .................................. 13
1.3.5 Other use cases ................................................. 15
1.4 Integration options ................................................... 16
1.4.1 Stand-alone Classification Module ............................... 16
1.4.2 Integration with IBM FileNet P8. ................................. 17
1.4.3 Integration with IBM Content Collector .......................... 18

### Chapter 2. Concepts, architecture, tools, and integration ............ 21
2.1 Classification concepts ............................................... 22
2.1.1 Knowledge base .................................................. 22
2.1.2 Classification workflow .......................................... 23
2.1.3 Core classification technology ................................... 25
2.1.4 Decision plan ..................................................... 26
2.1.5 Field definitions ................................................. 28
2.2 Core components .................................................... 29
2.2.1 Classification Workbench ....................................... 29
2.2.2 Classification Module server ..................................... 30
2.2.3 IBM FileNet P8 integration asset ............................... 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3 Classification Workbench</td>
<td>34</td>
</tr>
<tr>
<td>2.3.1 Create and configure knowledge base project overview</td>
<td>36</td>
</tr>
<tr>
<td>2.3.2 Taxonomy Proposer</td>
<td>37</td>
</tr>
<tr>
<td>2.4 Classification Module server</td>
<td>39</td>
</tr>
<tr>
<td>2.4.1 Management Console</td>
<td>40</td>
</tr>
<tr>
<td>2.4.2 Classification Module server deployment configurations</td>
<td>45</td>
</tr>
<tr>
<td>2.5 Classification Module APIs</td>
<td>48</td>
</tr>
<tr>
<td>2.6 Classification Module integration architecture</td>
<td>49</td>
</tr>
<tr>
<td>2.6.1 Classification Center</td>
<td>50</td>
</tr>
<tr>
<td>2.6.2 Content Extractor</td>
<td>57</td>
</tr>
<tr>
<td>2.7 Additional reference</td>
<td>57</td>
</tr>
<tr>
<td>Chapter 3. Working with knowledge bases and decision plans</td>
<td>59</td>
</tr>
<tr>
<td>3.1 Importance of knowledge bases and decision plans</td>
<td>60</td>
</tr>
<tr>
<td>3.2 Creating, training, and analyzing a knowledge base</td>
<td>60</td>
</tr>
<tr>
<td>3.2.1 Preparing data for import</td>
<td>61</td>
</tr>
<tr>
<td>3.2.2 Building a knowledge base</td>
<td>62</td>
</tr>
<tr>
<td>3.2.3 Analyzing and learning with a knowledge base</td>
<td>76</td>
</tr>
<tr>
<td>3.3 Creating and analyzing a decision plan</td>
<td>82</td>
</tr>
<tr>
<td>3.3.1 Creating a decision plan</td>
<td>83</td>
</tr>
<tr>
<td>3.3.2 Analyzing a decision plan</td>
<td>93</td>
</tr>
<tr>
<td>3.4 Building a knowledge base advanced topics</td>
<td>99</td>
</tr>
<tr>
<td>3.4.1 Using an uncategorized or partially categorized content set</td>
<td>100</td>
</tr>
<tr>
<td>3.4.2 Using keywords</td>
<td>101</td>
</tr>
<tr>
<td>3.4.3 Enabling the knowledge base with feedback processing</td>
<td>102</td>
</tr>
<tr>
<td>3.4.4 Working with offline (deferred) feedback</td>
<td>108</td>
</tr>
<tr>
<td>3.4.5 Handling overlapping categories</td>
<td>109</td>
</tr>
<tr>
<td>Chapter 4. Knowledge base 101</td>
<td>119</td>
</tr>
<tr>
<td>4.1 Use cases</td>
<td>120</td>
</tr>
<tr>
<td>4.2 Working with Classification Module</td>
<td>121</td>
</tr>
<tr>
<td>4.2.1 Typical life cycle of a knowledge base</td>
<td>122</td>
</tr>
<tr>
<td>4.2.2 Natural Language Processing</td>
<td>123</td>
</tr>
<tr>
<td>4.2.3 Classification Module implementation success</td>
<td>124</td>
</tr>
<tr>
<td>4.2.4 Classification Module implementation summary</td>
<td>124</td>
</tr>
<tr>
<td>4.3 Defining categories</td>
<td>125</td>
</tr>
<tr>
<td>4.3.1 Become familiar with the problem and your data</td>
<td>125</td>
</tr>
<tr>
<td>4.3.2 Choosing categories</td>
<td>126</td>
</tr>
<tr>
<td>4.3.3 Getting good sample content for each category</td>
<td>128</td>
</tr>
<tr>
<td>4.3.4 Discover the categories from your data</td>
<td>128</td>
</tr>
<tr>
<td>4.3.5 Separate content and categories for separate uses</td>
<td>128</td>
</tr>
<tr>
<td>4.3.6 A detailed example of the Self-Help/CRM scenario</td>
<td>130</td>
</tr>
<tr>
<td>4.4 Preparing the content set</td>
<td>131</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.4.1 Data sources for knowledge base building</td>
<td>131</td>
</tr>
<tr>
<td>4.4.2 The nature of the content</td>
<td>132</td>
</tr>
<tr>
<td>4.4.3 Cleaning corpus texts</td>
<td>133</td>
</tr>
<tr>
<td>4.4.4 Various types of texts: Their use and content types</td>
<td>134</td>
</tr>
<tr>
<td>4.4.5 Overcoming the lack of data: Using initialization data</td>
<td>137</td>
</tr>
<tr>
<td>4.5 Building and training the knowledge base</td>
<td>139</td>
</tr>
<tr>
<td>4.5.1 Building a knowledge base with existing categories and data</td>
<td>139</td>
</tr>
<tr>
<td>4.5.2 Building a knowledge base when the categories are unclear</td>
<td>140</td>
</tr>
<tr>
<td>4.5.3 Training as opposed to online learning (feedback)</td>
<td>140</td>
</tr>
<tr>
<td>4.5.4 Discover categories from your data using Taxonomy Proposer</td>
<td>141</td>
</tr>
<tr>
<td>4.6 Analyzing knowledge base performance: Identifying the problems</td>
<td>141</td>
</tr>
<tr>
<td>4.6.1 Analyzing your knowledge base</td>
<td>141</td>
</tr>
<tr>
<td>4.6.2 Sources of analysis data</td>
<td>143</td>
</tr>
<tr>
<td>4.6.3 Scores used for analysis</td>
<td>144</td>
</tr>
<tr>
<td>4.6.4 Understanding your analysis data</td>
<td>144</td>
</tr>
<tr>
<td>4.6.5 Running analysis in Classification Workbench</td>
<td>145</td>
</tr>
<tr>
<td>4.6.6 Measures of accuracy</td>
<td>146</td>
</tr>
<tr>
<td>4.6.7 The process of analysis</td>
<td>147</td>
</tr>
<tr>
<td>4.6.8 Automation and accuracy</td>
<td>147</td>
</tr>
<tr>
<td>4.6.9 The number of categories returned by matching</td>
<td>148</td>
</tr>
<tr>
<td>4.6.10 Typical analysis process (report interpretation)</td>
<td>148</td>
</tr>
<tr>
<td>4.6.11 Reviewing Cumulative Success scores</td>
<td>149</td>
</tr>
<tr>
<td>4.6.12 Understanding precision and recall</td>
<td>150</td>
</tr>
<tr>
<td>4.7 Fine-tuning your knowledge base: Fixing the problems</td>
<td>151</td>
</tr>
<tr>
<td>4.7.1 Overlapping categories (categories stealing from each other)</td>
<td>152</td>
</tr>
<tr>
<td>4.7.2 Low scores in all categories</td>
<td>153</td>
</tr>
<tr>
<td>4.7.3 Category does not represent intent of a message</td>
<td>154</td>
</tr>
<tr>
<td>4.7.4 Possible multiple intent categories</td>
<td>154</td>
</tr>
<tr>
<td>4.7.5 Human error (poor manual classification)</td>
<td>155</td>
</tr>
<tr>
<td>4.7.6 Poor sampling (randomization issues)</td>
<td>156</td>
</tr>
<tr>
<td>4.7.7 Identifying hidden subcategories</td>
<td>156</td>
</tr>
<tr>
<td>4.7.8 Poorly performing category with too few examples</td>
<td>157</td>
</tr>
<tr>
<td>4.7.9 Identifying obsolete categories</td>
<td>158</td>
</tr>
<tr>
<td>4.7.10 Automation thresholds</td>
<td>158</td>
</tr>
<tr>
<td>4.8 Maintaining your knowledge base over time: Using feedback</td>
<td>160</td>
</tr>
<tr>
<td>4.8.1 Working with feedback: Avoiding the problems</td>
<td>161</td>
</tr>
<tr>
<td>4.9 Analyzing the knowledge base in production</td>
<td>164</td>
</tr>
<tr>
<td>4.9.1 Adding and removing categories</td>
<td>165</td>
</tr>
<tr>
<td>4.10 Application design considerations</td>
<td>166</td>
</tr>
<tr>
<td>4.10.1 Matching</td>
<td>166</td>
</tr>
<tr>
<td>4.10.2 Feedback</td>
<td>167</td>
</tr>
<tr>
<td>4.10.3 Using Classification Module to route texts</td>
<td>169</td>
</tr>
<tr>
<td>4.10.4 Gathering feedback in an application</td>
<td>170</td>
</tr>
</tbody>
</table>
Part 2. Integration details ...................................................... 175

Chapter 5. Integration with IBM FileNet P8 .......................... 177
5.1 Solution introduction .................................................. 178
  5.1.1 Integration architecture ......................................... 178
  5.1.2 Use case .......................................................... 180
5.2 Integration steps overview ............................................ 182
5.3 Enabling the integration ............................................. 183
  5.3.1 Installing integration component for IBM FileNet P8 ....... 183
  5.3.2 Configuring IBM FileNet P8 for integration .................. 185
  5.3.3 Configuring connectivity between Classification Module and IBM FileNet
       P8 ............................................................................ 198
5.4 Training Classification Module with existing IBM FileNet P8 content ... 204
  5.4.1 Using Content Extractor .......................................... 205
  5.4.2 Creating a knowledge base ..................................... 215
  5.4.3 Creating a decision plan ....................................... 215
  5.4.4 Analyzing a decision plan and its related knowledge base ... 234
  5.4.5 Exporting knowledge bases and decision plans from Classification
       Workbench ................................................................. 235
  5.4.6 Deploying the knowledge base and decision plan using Management
       Console ................................................................. 242
5.5 Configuring and performing classification ........................ 249
  5.5.1 Classification Center overview ................................ 249
  5.5.2 Working with the Classification Center ..................... 252
  5.5.3 Configuring Classification Center ............................ 257
  5.5.4 Verifying IBM FileNet P8 structures ......................... 269
  5.5.5 Performing classification ..................................... 273
5.6 Reviewing and fine-tuning classification ........................ 276
  5.6.1 Classification Center review function overview .......... 276
  5.6.2 Providing feedback to the knowledge base ................. 276
  5.6.3 Configuring the Classification Center for review .......... 277
  5.6.4 Reclassifying a document ..................................... 283

Chapter 6. Integration with IBM Content Collector for File Systems .... 291
6.1 Solution introduction .................................................. 292
  6.1.1 Overview of IBM Content Collector for File Systems ........ 292
  6.1.2 Integration architecture ....................................... 292
How to get IBM Redbooks publications .......................... 434
Help from IBM .................................................. 434

Index ............................................................... 435
# Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Automated classification</td>
<td>4</td>
</tr>
<tr>
<td>1-2</td>
<td>A simple classification scheme</td>
<td>7</td>
</tr>
<tr>
<td>1-3</td>
<td>Classification Module operational schematic</td>
<td>9</td>
</tr>
<tr>
<td>1-4</td>
<td>Automatic mapping to records file plan</td>
<td>11</td>
</tr>
<tr>
<td>1-5</td>
<td>Mapping old structures to new</td>
<td>14</td>
</tr>
<tr>
<td>1-6</td>
<td>Customer support and retention management process</td>
<td>15</td>
</tr>
<tr>
<td>1-7</td>
<td>A simple Classification Module implementation</td>
<td>17</td>
</tr>
<tr>
<td>1-8</td>
<td>Integration with IBM FileNet P8 content repository</td>
<td>18</td>
</tr>
<tr>
<td>1-9</td>
<td>IBM Content Collector integration with IBM Content Manager (CM8)</td>
<td>19</td>
</tr>
<tr>
<td>2-1</td>
<td>A KB project for HR categories</td>
<td>24</td>
</tr>
<tr>
<td>2-2</td>
<td>Inside the IBM Classification technology</td>
<td>25</td>
</tr>
<tr>
<td>2-3</td>
<td>A decision plan project</td>
<td>27</td>
</tr>
<tr>
<td>2-4</td>
<td>Field definition settings within Classification Workbench</td>
<td>28</td>
</tr>
<tr>
<td>2-5</td>
<td>Classification Module server system architecture</td>
<td>31</td>
</tr>
<tr>
<td>2-6</td>
<td>Classification Module server processes: Single server mode</td>
<td>33</td>
</tr>
<tr>
<td>2-7</td>
<td>Classification Workbench</td>
<td>35</td>
</tr>
<tr>
<td>2-8</td>
<td>Taxonomy Proposer</td>
<td>37</td>
</tr>
<tr>
<td>2-9</td>
<td>Taxonomy Proposer Workflow Assistant</td>
<td>38</td>
</tr>
<tr>
<td>2-10</td>
<td>Typical workflow of a Classification Module system</td>
<td>39</td>
</tr>
<tr>
<td>2-11</td>
<td>Management Console connection parameters</td>
<td>41</td>
</tr>
<tr>
<td>2-12</td>
<td>Management Console application</td>
<td>41</td>
</tr>
<tr>
<td>2-13</td>
<td>Management Console: Knowledge bases administration tab</td>
<td>42</td>
</tr>
<tr>
<td>2-14</td>
<td>Management Console: Decision plans administration tab</td>
<td>43</td>
</tr>
<tr>
<td>2-15</td>
<td>Management Console: Field definitions administration tab</td>
<td>44</td>
</tr>
<tr>
<td>2-16</td>
<td>Management Console: Servers administration tab</td>
<td>45</td>
</tr>
<tr>
<td>2-17</td>
<td>Management Console: Servers administration host tab</td>
<td>45</td>
</tr>
<tr>
<td>2-18</td>
<td>Single server configuration</td>
<td>46</td>
</tr>
<tr>
<td>2-19</td>
<td>Multiple server configuration</td>
<td>47</td>
</tr>
<tr>
<td>2-20</td>
<td>Multiple listener configuration</td>
<td>48</td>
</tr>
<tr>
<td>2-21</td>
<td>Classification Module client API architecture</td>
<td>49</td>
</tr>
<tr>
<td>2-22</td>
<td>Classification Module integration process workflow architecture</td>
<td>50</td>
</tr>
<tr>
<td>2-23</td>
<td>Classification Center architecture</td>
<td>51</td>
</tr>
<tr>
<td>2-24</td>
<td>Classification server start command window</td>
<td>52</td>
</tr>
<tr>
<td>2-25</td>
<td>Classification Center landing page</td>
<td>53</td>
</tr>
<tr>
<td>2-26</td>
<td>Classification Center Configuration tab</td>
<td>54</td>
</tr>
<tr>
<td>2-27</td>
<td>Classification Center Dashboard tab</td>
<td>55</td>
</tr>
<tr>
<td>2-28</td>
<td>Classification Center Review tab</td>
<td>56</td>
</tr>
<tr>
<td>3-1</td>
<td>Classification Workbench: Knowledge base development cycle</td>
<td>61</td>
</tr>
<tr>
<td>Page</td>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3-2</td>
<td>File system folders as a content set</td>
<td>62</td>
</tr>
<tr>
<td>3-3</td>
<td>Starting Classification Workbench</td>
<td>63</td>
</tr>
<tr>
<td>3-4</td>
<td>New knowledge base project window</td>
<td>64</td>
</tr>
<tr>
<td>3-5</td>
<td>Creating a project by importing a content set</td>
<td>65</td>
</tr>
<tr>
<td>3-6</td>
<td>Content set format: Files from a file system folder</td>
<td>65</td>
</tr>
<tr>
<td>3-7</td>
<td>Selecting the root folder of a content set</td>
<td>66</td>
</tr>
<tr>
<td>3-8</td>
<td>Applying file filters</td>
<td>67</td>
</tr>
<tr>
<td>3-9</td>
<td>Editing category names for file system folders</td>
<td>68</td>
</tr>
<tr>
<td>3-10</td>
<td>Specifying language and text filter settings</td>
<td>68</td>
</tr>
<tr>
<td>3-11</td>
<td>Project view with imported content items</td>
<td>69</td>
</tr>
<tr>
<td>3-12</td>
<td>The Body field properties</td>
<td>70</td>
</tr>
<tr>
<td>3-13</td>
<td>Options of creating and analyzing a knowledge base</td>
<td>72</td>
</tr>
<tr>
<td>3-14</td>
<td>Create a new knowledge base and delete existing knowledge bases</td>
<td>73</td>
</tr>
<tr>
<td>3-15</td>
<td>Analysis options</td>
<td>74</td>
</tr>
<tr>
<td>3-16</td>
<td>Status information of creating and analyzing a knowledge base</td>
<td>75</td>
</tr>
<tr>
<td>3-17</td>
<td>Content set view with match fields</td>
<td>75</td>
</tr>
<tr>
<td>3-18</td>
<td>Generating Cumulative Success report and Total Precision vs. Recall</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Graph</td>
<td></td>
</tr>
<tr>
<td>3-19</td>
<td>Total cumulative success in Knowledge Base Data Sheet for HR</td>
<td>78</td>
</tr>
<tr>
<td>3-20</td>
<td>Cumulative Success Report for HR</td>
<td>79</td>
</tr>
<tr>
<td>3-21</td>
<td>Total Precision vs. Recall graph for HR</td>
<td>80</td>
</tr>
<tr>
<td>3-22</td>
<td>Create Knowledge base, using existing knowledge base structure</td>
<td>81</td>
</tr>
<tr>
<td>3-23</td>
<td>Learn during analysis</td>
<td>82</td>
</tr>
<tr>
<td>3-24</td>
<td>New decision plan project window</td>
<td>84</td>
</tr>
<tr>
<td>3-25</td>
<td>Creating an empty project</td>
<td>85</td>
</tr>
<tr>
<td>3-26</td>
<td>Project overview of the new empty decision plan</td>
<td>85</td>
</tr>
<tr>
<td>3-27</td>
<td>Add an existing knowledge base project as a referenced project</td>
<td>86</td>
</tr>
<tr>
<td>3-28</td>
<td>Add a knowledge base project to your decision plan</td>
<td>86</td>
</tr>
<tr>
<td>3-29</td>
<td>Defining a rule group</td>
<td>87</td>
</tr>
<tr>
<td>3-30</td>
<td>Adding a new rule</td>
<td>88</td>
</tr>
<tr>
<td>3-31</td>
<td>Rule 1 Properties tab</td>
<td>89</td>
</tr>
<tr>
<td>3-32</td>
<td>Rule 1 Trigger tab</td>
<td>89</td>
</tr>
<tr>
<td>3-33</td>
<td>Rule 1 Add Actions</td>
<td>90</td>
</tr>
<tr>
<td>3-34</td>
<td>Rule 2 Properties tab</td>
<td>91</td>
</tr>
<tr>
<td>3-35</td>
<td>Rule 2 Trigger tab</td>
<td>91</td>
</tr>
<tr>
<td>3-36</td>
<td>Rule 2 Add Action</td>
<td>92</td>
</tr>
<tr>
<td>3-37</td>
<td>Decision plan view with two rules and one knowledge base</td>
<td>93</td>
</tr>
<tr>
<td>3-38</td>
<td>Content set file format: CSV</td>
<td>94</td>
</tr>
<tr>
<td>3-39</td>
<td>Select CSV file</td>
<td>94</td>
</tr>
<tr>
<td>3-40</td>
<td>Editing property of the Message_body field</td>
<td>95</td>
</tr>
<tr>
<td>3-41</td>
<td>Setting the Content type value to the Message_body field</td>
<td>96</td>
</tr>
<tr>
<td>3-42</td>
<td>Launching the Analyze Decision Plan wizard</td>
<td>96</td>
</tr>
<tr>
<td>3-43</td>
<td>Report options of analyzing a decision plan</td>
<td>97</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3-44</td>
<td>Decision plan summary report</td>
<td>98</td>
</tr>
<tr>
<td>3-45</td>
<td>Analyzed content set</td>
<td>99</td>
</tr>
<tr>
<td>3-46</td>
<td>Taxonomy Proposer</td>
<td>100</td>
</tr>
<tr>
<td>3-47</td>
<td>Keywords CSV file for the knowledge base creation</td>
<td>102</td>
</tr>
<tr>
<td>3-48</td>
<td>Management Console logon</td>
<td>103</td>
</tr>
<tr>
<td>3-49</td>
<td>Knowledge base menu</td>
<td>104</td>
</tr>
<tr>
<td>3-50</td>
<td>Feedback: Defer processing</td>
<td>105</td>
</tr>
<tr>
<td>3-51</td>
<td>Restart knowledge base</td>
<td>106</td>
</tr>
<tr>
<td>3-52</td>
<td>Learn using active view</td>
<td>108</td>
</tr>
<tr>
<td>3-53</td>
<td>Analyze knowledge base using active view</td>
<td>109</td>
</tr>
<tr>
<td>3-54</td>
<td>View reports</td>
<td>110</td>
</tr>
<tr>
<td>3-55</td>
<td>Knowledge Base Data Sheet report</td>
<td>111</td>
</tr>
<tr>
<td>3-56</td>
<td>Pairs of categories with overlapping intents</td>
<td>112</td>
</tr>
<tr>
<td>3-57</td>
<td>Example of overlapping categories: Gift Certificates and Gift Wrap</td>
<td>112</td>
</tr>
<tr>
<td>3-58</td>
<td>Category graphs and tables</td>
<td>113</td>
</tr>
<tr>
<td>3-59</td>
<td>Stealing/stolen table</td>
<td>113</td>
</tr>
<tr>
<td>3-60</td>
<td>Content item scoring graph</td>
<td>114</td>
</tr>
<tr>
<td>3-61</td>
<td>Show items of the Gift Wrap category</td>
<td>115</td>
</tr>
<tr>
<td>3-62</td>
<td>Select all items belonging to the Gift Wrap category</td>
<td>115</td>
</tr>
<tr>
<td>3-63</td>
<td>Categorize highlighted items</td>
<td>116</td>
</tr>
<tr>
<td>3-64</td>
<td>Select the Gift Certificates category to apply to highlighted items</td>
<td>116</td>
</tr>
<tr>
<td>3-65</td>
<td>Deleting the Match n fields</td>
<td>117</td>
</tr>
<tr>
<td>3-66</td>
<td>Knowledge Base Data Sheet with improved cumulative success</td>
<td>118</td>
</tr>
<tr>
<td>4-1</td>
<td>Analysis process</td>
<td>143</td>
</tr>
<tr>
<td>4-2</td>
<td>Precision/Recall graph</td>
<td>151</td>
</tr>
<tr>
<td>5-1</td>
<td>Classification Module integration with IBM FileNet P8</td>
<td>179</td>
</tr>
<tr>
<td>5-2</td>
<td>Classification Center Integration Components menu</td>
<td>183</td>
</tr>
<tr>
<td>5-3</td>
<td>Basic or Custom installation</td>
<td>184</td>
</tr>
<tr>
<td>5-4</td>
<td>Features selection</td>
<td>185</td>
</tr>
<tr>
<td>5-5</td>
<td>IBM FileNet P8 folder structure under icm_integration folder</td>
<td>185</td>
</tr>
<tr>
<td>5-6</td>
<td>The Classification Workbench category structure in the Knowledge Base Editor window</td>
<td>186</td>
</tr>
<tr>
<td>5-7</td>
<td>Start IBM FileNet Enterprise Manager</td>
<td>187</td>
</tr>
<tr>
<td>5-8</td>
<td>Select IBM FileNet P8 server to integrate with Classification Module</td>
<td>188</td>
</tr>
<tr>
<td>5-9</td>
<td>Create a new AddOn</td>
<td>188</td>
</tr>
<tr>
<td>5-10</td>
<td>New AddOn configuration</td>
<td>189</td>
</tr>
<tr>
<td>5-11</td>
<td>Open the XML file for the new AddOn</td>
<td>190</td>
</tr>
<tr>
<td>5-12</td>
<td>Verify the correct import file name and path for the new AddOn</td>
<td>191</td>
</tr>
<tr>
<td>5-13</td>
<td>Displaying New AddOn in the list of AddOns</td>
<td>192</td>
</tr>
<tr>
<td>5-14</td>
<td>Install the new AddOn</td>
<td>193</td>
</tr>
<tr>
<td>5-15</td>
<td>Select the AddOn to install</td>
<td>194</td>
</tr>
<tr>
<td>5-16</td>
<td>New AddOn installed confirmation</td>
<td>194</td>
</tr>
<tr>
<td>5-17</td>
<td>Refresh the object store to see the new imported properties</td>
<td>195</td>
</tr>
</tbody>
</table>
5-18 Modify properties of document class .......................... 196
5-19 Document Class Property Definitions tab ......................... 197
5-20 Select all ICM_ prefixed properties ............................... 198
5-21 Edit setupCommandLine.sh ........................................ 199
5-22 The startConnectTest.bat file ..................................... 200
5-23 Edit the WcmConfig.properties file ................................. 201
5-24 Change wcmconfig parameter in the WcmConfig.properties file ......................................................... 201
5-25 Edit the WcmApiConfig40.properties file ......................... 202
5-26 Testing connection ................................................... 203
5-27 Verify Classification Module connection to FileNet P8 is successful ...................................................... 204
5-28 Extractor.properties file path ...................................... 206
5-29 Edit Extractor.properties ............................................ 207
5-30 Default directory for XML output from Content Extractor ................................................................. 208
5-31 Path for the object store from which to extract content ........................................................................ 208
5-32 Output window of running startExtractor.bat .................. 210
5-33 Output of running Content Extractor with the -m parameter ................................................................ 211
5-34 Directory where the output of the Content Extractor file is located ....................................................... 212
5-35 ExtractorOutput.txt ...................................................... 212
5-36 After content extraction completes ................................ 213
5-37 XML output file from the content extraction ...................... 214
5-38 Output binary content .................................................... 214
5-39 Launch Classification Workbench .................................... 217
5-40 Classification Workbench: Workflow assistant .................. 217
5-41 Create a new project (decision plan or knowledge base) .... 218
5-42 Create an empty project ................................................ 219
5-43 New rule for decision plan ............................................. 219
5-44 New Rule Properties tab .............................................. 220
5-45 Set the trigger for the new rule ...................................... 220
5-46 Add actions to the new rule ......................................... 221
5-47 Display decision plan actions ....................................... 223
5-48 Display configured action ............................................. 224
5-49 Decision plan ............................................................. 225
5-50 Create a new Review rule ............................................. 226
5-51 Define the trigger condition for Review rule .................... 227
5-52 Define Review rule based on score result from the select_branch knowledge base ........................................... 228
5-53 Define if score is less than 80% for Review rule ................. 229
5-54 Add action to the Review rule ....................................... 229
5-55 Review rule action ..................................................... 231
5-56 Action for Review rule is created .................................. 232
5-57 Decision plan with new Review rule ............................... 233
5-58 Save the new decision plan: branches ............................. 234
5-59 Open project in Classification Workbench ........................ 235
5-100 Open knowledge base that is referenced by the decision plan . . . . . . . . 271
5-101 View categories for the knowledge base named select_branch . . . . . . . 272
5-102 Start classifying documents ...................................................... 274
5-103 Classifying documents with elapsed time showing ...................... 275
5-104 Click the Review icon to start the document review process ......... 278
5-105 Filter Settings for reviewing documents .................................. 279
5-106 Edit Filter Settings ............................................................... 280
5-107 Document filter ................................................................. 282
5-108 Review documents .............................................................. 284
5-109 Review decision history .......................................................... 285
5-110 Reclassify documents ............................................................ 285
5-111 Reclassify document to additional categories ................................. 286
5-112 Save document in XML format .................................................. 287
5-113 Add a document to IBM FileNet P8 for classification analysis .......................... 288
5-114 Browse to the document location to add it for classification analysis ... 288
5-115 Review the decision history for document classification .......... 289
6-1 Architecture overview of the integration ........................................ 293
6-2 Installing Classification Module client components on the Content Collector server .......................................................... 296
6-3 Validating that the Windows registry entry of
ibm.ctms.utilityconnector.ICMClassificationTask exists ..................... 298
6-4 Classification Module task ............................................................. 299
6-5 Classification Module system metadata properties in Content Collector
Configuration Manager ........................................................................ 300
6-6 Selecting the Add knowledge base menu option ............................... 302
6-7 Adding a knowledge base to the Classification Module server ......... 304
6-8 Field definitions ................................................................. 305
6-9 File system archiving task route to create ........................................ 306
6-10 The Task Routes explore pane and Toolbox of the Content Collector
Configuration Manager ....................................................................... 307
6-11 Add icon for adding a new task route ............................................. 308
6-12 Creating a new task route from a blank task route ............................ 309
6-13 A new blank task route ................................................................ 309
6-14 Selecting the FSC Collector .......................................................... 310
6-15 Adding an FSC Collector to the task route ..................................... 311
6-16 Selecting the P8 4.x Create Document task ..................................... 313
6-17 Adding a P8 4.x Create Document task to the task route ............. 314
6-18 Selecting the IBM Classification Module task ................................. 316
6-19 Adding a Classification Module task to the task route .................... 317
6-20 Configuring the Classification Module task .................................... 318
6-21 Adding the Decision Point to the task route .................................. 319
6-22 Add conditional clause ............................................................... 320
6-23 Configuring the rule of the left branch of the task route .............. 321
7-26 E-mail Services: Web Application Client settings ........................................ 374
7-27 Create New Task Route .................................................................................. 375
7-28 Creating new task route from a blank task route ........................................... 376
7-29 Empty task route showing start and end task nodes ....................................... 376
7-30 Start task node: General tab settings ............................................................. 377
7-31 Selecting EC Collect E-Mail By Rules task node .......................................... 378
7-32 EC Collect E-Mail By Rules: General tab settings ........................................... 379
7-33 EC Collect E-mail By Rules: Schedule tab settings ....................................... 380
7-34 EC Collect E-mail By Rules: Collection Source tab settings ....................... 381
7-35 EC Collect E-mail By Rules: Collection Source addition ............................ 382
7-36 EC Collect E-mail By Rules: Filter tab settings .............................................. 383
7-37 Selecting EC Extract Metadata task node ...................................................... 384
7-38 EC Extract Metadata: General tab settings .................................................... 385
7-39 Selecting EC Prepare E-mail for Archive task node ....................................... 386
7-40 EC Prepare E-mail for Archive: General tab settings .................................... 387
7-41 Adding EC Finalize E-mail for Compliance task node .................................... 388
7-42 EC Finalize for Compliance: General tab settings ......................................... 389
7-43 Adding P8 4.x Create Document task node .................................................... 390
7-44 P8 4.x Create Document: General tab settings .............................................. 392
7-45 Adding Classification Module task node ........................................................ 394
7-46 Utility Classification Module: General Properties ........................................ 396
7-47 Adding a decision point .................................................................................. 398
7-48 Conditional clause for Score < 70 rule ............................................................ 399
7-49 Score < 70 decision rule: General tab settings ............................................. 400
7-50 Add new rule .................................................................................................. 401
7-51 Rule General Properties ................................................................................ 402
7-52 Conditional metadata rule ............................................................................. 403
7-53 Adding Create P8 4.x File Document in Folder task node ............................. 404
7-54 P8 4.x File Document in Folder: General tab settings .................................... 406
7-55 Adding P8 4.x File Document in Folder task node ......................................... 407
7-56 FileNet P8 4.x File Document in Folder: General tab settings .................... 409
7-57 Adding EC Prepare E-mail for Stubbing task node ........................................ 410
7-58 EC Prepare E-mail for Stubbing: General tab settings ................................... 412
7-59 New link .......................................................................................................... 413
7-60 New link connecting P8 4.x File Document in Folder task node to EC Prepare E-Mail for Stubbing task node ................................................................. 414
7-61 Adding EC Create E-Mail Stub task node ....................................................... 415
7-62 E-mail Server - EC Create E-mail Stub: General Properties ......................... 417
7-63 Active flag for task route ................................................................................ 419
7-64 Start Task Routing Engine service ................................................................. 420
7-65 Activating the audit log ................................................................................... 421
7-66 Audit Log Node General Properties ............................................................... 422
7-67 Folder created based on top category Redbook/HR/Stock Options ............... 424
7-68 Folder created based on top category Redbook/HR/Pay ............... 425
7-69 Folder created based on score < 70% - Redbook/manualReview .... 426
7-70 Adding the P8 4.x Declare Record task node ......................... 428
7-71 P8 4.x Declare Record Classification folder selection. ............ 429
7-72 P8 4.x Declare Records: General tab settings ...................... 430
7-73 E-mail declared as records ............................................. 431


Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

The following company names appearing in this publication are fictitious:

Fictional Auto Rental Company A

This name is used for instructional purposes only.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in
any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

**Trademarks**

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at [http://www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

- AIX®
- DB2®
- Domino®
- FileNet®
- Global Business Services®
- IBM®
- InfoSphere™
- iSeries®
- Lotus Notes®
- Lotus®
- Notes®
- OmniFind®
- Quickr™
- Redbooks®
- Redbooks (logo)®
- Sametime®
- Tivoli®

The following terms are trademarks of other companies:

FileNet and the FileNet logo are registered trademarks of FileNet Corporation in the United States, other countries or both.

Java, JDBC, Solaris, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Active Directory, Excel, Internet Explorer, Microsoft, MS, Outlook, SharePoint, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.
IBM® Classification Module (Classification Module) Version 8.6 is an advanced enterprise software platform tool designed to allow organizations to automate the classification of unstructured content. By deploying the module in various areas of a business, organizations can reduce or avoid manual processes associated with subjective decision making around unstructured content. Organizations can also streamline the ingestion of that content into their business systems in order to use the information within the business systems more effectively. At the same time, the organizations can safely remove irrelevant or obsolete information and therefore utilize the storage infrastructure more efficiently. By reducing the human element in this process, Classification Module ensures accuracy and consistency and enables auditing while simultaneously driving down labor costs.

This IBM Redbooks® publication explains what Classification Module does, the key concepts to understand when working with Classification Module, and its integration with other products and systems. With this book, we show you how Classification Module helps your organization to automate the classification of large volumes of unstructured content in a consistent and accurate manner. The topics that are covered include building, training, and fine-tuning the knowledge base, creating decision plans, working with Classification Workbench, and step-by-step integration with other products and solutions.

This book is intended to educate both technical specialists and non-technical personnel in how to make Classification Module work for your organizations.

**Changes in product version and name:** This book is written based on Classification Module Version 8.6. The new Version 8.7 has since been released. The product also has a new name: IBM InfoSphere™ Classification Module. For the purpose of accuracy, Version 8.6 was known in the field as IBM Classification Module (and also Classification Module), and this name is used in the rest of the book.

**What’s new:** You can view what's new from the following Information Center topic:

The team who wrote this book

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Rochester Center.

Wei-Dong Zhu (Jackie) is an Enterprise Content Management Project Leader with the International Technical Support Organization. She has more than 10 years of software development experience in accounting, workflow processing, and digital media distribution. Jackie holds a Master of Science degree in Computer Science from the University of the Southern California. Jackie joined IBM in 1996. She is a Certified Solution Designer for IBM Content Manager and has managed and led the production of many Enterprise Content Management IBM Redbooks publications.

Solomon Barron (Sol) is an Enterprise Content Management IT Specialist in the U.K. He has over 20 years experience in IT technical sales with more than 10 years in Content Management. He studied Law at University in England before moving to Israel where he worked in technical writing within the telecommunications industry. His areas of expertise include records and document management, specifically in the U.K. public sector where he has had extensive experience implementing large scale document and records management systems. Sol joined IBM in 2000.

Maurizio Gallotti is an Enterprise Content Management IT Specialist focused on Discovery and Compliance solutions, in IBM Italy. He has over 29 years of experience in various IT professional fields in IBM, from iSeries® technical support, application change management, to lab infrastructure support and project management. Maurizio has worked in the IBM Rochester lab for a year, in the IBM Tivoli® lab in Rome for five years, and has had both Global Business Services® (GBS) and Software Group (SWG) experience. Maurizio has been in Technical Sales in Information Management since 2005, and has always had direct client technical responsibility in the Italy industrial sector.

Vijay Gupta is a Managing Consultant and Senior IT Specialist in Enterprise Content Management and Discovery products, in IBM U.S. He is on the Professional Services and Enablement team and works directly with clients by providing consulting services for Content Discovery and Enterprise Content Management solutions. Vijay recently started his career with IBM as a Technical Consultant in 2006. He has over twelve years of experience in software and network design and development, validation and verification, and client services. Vijay’s areas of expertise include application development, customer engineering, technical support, and quality analysis. Vijay holds a Master degree in Computer Science from University of Pune, India, and a Master of Business Administration degree from Babson College.
Xiaomei Wang is a Senior Technical Consultant in Enterprise Content Management, Business Partner Technical Enablement team, in IBM U.S. She assists IBM Business Partners with integrating IBM Discovery and Enterprise Content Management products into their solutions. Xiaomei holds a Master of Science degree in Computer Science from York University. She joined IBM in 1999 and has held numerous positions in the company, such as the DB2® UDB Advanced Support Specialist, the Information Management Business Development Manager, and the Technical Sales Specialist. Xiaomei is a Certified IBM IT Specialist Professional and a Board Member. In addition, she is an IBM Certified Solution Experts (IBM Content Manager, IBM FileNet® P8, DB2 UDB Database Administration and Application Development, and Business Intelligence), and a Prentice Hall book author.

The following people from the Classification Module software development team, IBM Israel Software lab, have also contributed to writing this book:

Josemina Magdalen is a Senior Team Leader and Architect at Israel Software Group (ILSL). She has a background in Natural Language Processing (text classification and search, as well as text mining and filtering technologies). Josemina joined IBM in 2005 and has worked in the Content Discovery Engineering Group performing software development projects in text categorization, filtering, and search, as well as text analytics. Prior to joining IBM, Josemina has worked in Natural Language Processing research and development (Machine Translation, Text Classification and Search, and Data Mining) for over ten years.

Jane Singer is a member of the Advanced Engagement Team. She has worked as a Support Engineer in the Software Group with a focus in the Information Management area, supporting IBM OmniFind Discovery Edition, OmniFind Enterprise Edition, and IBM Classification Module. Jane is part of the IBM Israel Software Development Lab in Jerusalem, Israel, and she has worked within the Quality Assurance group and supported presales for IBM Classification Module for many years. Jane holds a Masters Degree in Library and Information Science and a Ph.D. in Musicology, both from the Hebrew University of Jerusalem.

Very special thanks to the entire IBM Classification Module software development team in Jerusalem, Israel. Without their gracious and enthusiastic assistance, we would not have been able to produce this book. Specifically, we would like to thank the following people from the team for their contribution to this book:

Shimon Stark
Adina Taragin
Oren Paikowsky
Yariv Tzaban
Steve Kirshner
Victoria Mazel
Eliahu Ben-Reuven
Aryeh Krassenstein
Tamar Lavee
Boruch Nager
Eitan Porat
Vladislav Rybak
Nir Salansky
Dmitry Shusterman
Denis Voloshin
IBM Software Development Lab, Jerusalem, Israel

We want to thank the following people from IBM U.S. who have contributed to this book project:

Chuck Beretz
Srinivas “Varma” Chitivel
Joshua Payne
IBM Software Development Lab, United States

**Become a published author**

Join us for a two- to six-week residency program. Help write a book dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

**Comments welcome**

Your comments are important to us.

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:
Use the online **Contact us** review IBM Redbooks publications form found at: ibm.com/redbooks

Send your comments in an e-mail to: redbooks@us.ibm.com

Mail your comments to:
IBM Corporation, International Technical Support Organization
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400
Core product information

In this part, we provide a product overview of IBM Classification Module, its architecture, and its value to your business. We also introduce the key concepts that are associated with the classification technology, including the knowledge base and the decision plan, how to create them, and how to work with them to help your organization to automate the classification process.
Overview

This chapter provides both technical specialists and non-technical personnel with an introduction to IBM Classification Module (Classification Module) and the advanced classification technology that it uses. After reading this chapter, you can establish the relevance of Classification Module to your project and have a basic understanding of how to architect a classification solution.

In this chapter, we discuss the following topics:

► Introducing Classification Module
► How Classification Module works
► Business use cases
► Integration options

Changes in product version and name: This book is written based on IBM Classification Module Version 8.6. The new Version 8.7 has since been released. The product also has a new name: IBM InfoSphere Classification Module. For the purpose of accuracy, Version 8.6 was known in the field as IBM Classification Module, and this name is used in the rest of the book.

What’s new: You can view what is new for Version 8.7 at the following Information Center topic:

1.1 Introducing Classification Module

Organizations around the world are struggling to deal with the explosion of information in their storage systems ranging from text files and spreadsheets to e-mails and instant messaging transcripts. The vast majority of this information is unstructured content, and there is no way of knowing what content is important, what is irrelevant, and what constitutes a risk to the business. Using knowledge workers to manually review information in order to categorize and classify this digital landfill is prohibitively costly for most companies. Managing this information becomes a daunting challenge for organizations in both the private and public sectors that are compelled to address issues, such as compliance, freedom of information, accountability, and transparency, yet do not know what is currently held on their own disk drives.

Classification Module is an advanced enterprise software tool that is designed to allow organizations to automate the classification of unstructured content (see Figure 1-1). By deploying the module in various areas of a business, organizations can reduce or avoid manual processes that are associated with subjective decision making around unstructured content and streamline the ingestion of that content into their business systems in order to use the information within those business systems. At the same time, organizations can safely remove irrelevant or obsolete information and therefore utilize their storage infrastructure more efficiently. By reducing the human element in this process, Classification Module ensures accuracy and consistency and enables auditing while simultaneously driving down labor costs.

![Figure 1-1 Automated classification](image-url)
Chapter 1. Overview

1.2 How Classification Module works

Classification Module enables applications to understand free-form communication. Classification Module’s ability to understand is similar to human understanding.

Human understanding is essentially a classification task. When someone describes something new to us, we intuitively identify it based on how it fits into our existing knowledge. If someone shows us a picture of a lion, and we have never seen one before, we can probably identify it as an animal. This conclusion is based upon our existing knowledge that tells us that this creature has hair and other features that we associate with animals. We might even identify it as a cat, because it has whiskers and other features that we associate with cats.

Similarly, Classification Module attempts to understand information based on its existing knowledge. It learns how to distinguish and classify data based on its acquired information. For example, Classification Module learns how to distinguish between texts about dogs and cats, after you provide it with a set of example texts about dogs and another set of examples about cats. After learning, Classification Module will attempt to correctly identify texts related to dogs and texts related to cats. However, when presented with texts about a separate animal (for which no example texts were provided), Classification Module cannot be expected to accurately categorize the text.

Classifying free-form text is inherently difficult, because it is unlikely that a new free-form text will be identical to any of the example texts. Therefore, to understand its content, we classify the text based upon how many features it has in common with the existing knowledge and example texts. Classification Module enables software applications to understand and classify free-form texts. Classification Module uses natural language processing and sophisticated semantic analysis to perform this task.

Classification Module is a software component that enables applications to understand and classify free-form texts by comparing them to a previously defined set of categories, contained in a knowledge base. Classification Module analyzes the free-form text and determines which categories in the knowledge base best match the free-form text.

When the application, powered by Classification Module, understands the intent of the text, it can take appropriate action. For e-mail, or other online applications, the appropriate action might be to route the information to the correct person, to provide an automatic e-mail response, or to look up and return a specific response from a database of answers.
No matter what action the application takes, we want to ensure that the text categorization is as accurate as possible. To ensure accuracy, we build a knowledge base that reflects the nature and structure of the world that it is trying to understand. In addition, the categories in the knowledge base are carefully selected to reflect real-life business processes.

In the next few sections, we explain in simple terms the elements that are required in the classification solution. We provide a more detailed and technical description in Chapter 2, “Concepts, architecture, tools, and integration” on page 21.

1.2.1 Input and output

What goes into Classification Module (the module) and what does it return?

**Input: Content to be classified**
Classification Module accepts any document type to be classified and attempts to analyze its content, as well as extract metadata from the file. Classification Module uses separate filter types for separate document types to extract the information appropriately.

Graphics files, such as .tif and .gif, can be preprocessed through an optical character recognition (OCR) application to extract the textual content from a scanned hard copy document. The resulting text stream can be passed to the module.

**Output: Resulting categories and scores**
Documents are processed individually in a synchronous operation with each call to Classification Module returning a set of metadata consisting of classification scores and optionally other metadata fields consisting of values extracted from the analyzed content. These metadata fields can also be calculated fields based on extracted content data.

In a simple implementation, the classification scores will usually be the top three or four matching categories with the category having the score closest to 100% being the suggested one. This information, together with the rest of the metadata, is what is returned and made available to the calling application.
1.2.2 Configuration process

The key to the Classification Module’s effectiveness is the company’s own domain expertise. The success of automated classification is dependent on your ability to create and train Classification Module, by real-life example, using genuine business documents for each required classification.

**Identifying the required categories**

The first step in configuring Classification Module for usage is deciding on a set of categories. Often, organizations already have categories in use - even informally - and can consolidate and document them. If this situation is not the case, Classification Module can analyze a set of documents, which are representative of the organization’s unstructured content, and propose a *taxonomy* of its own, a classification list, which can form the basis of a new classification scheme. Such classification schemes can be as simple or as complex as is necessary. Figure 1-2 shows a simple classification scheme.

![Example classification scheme]

*Figure 1-2  A simple classification scheme*

**Training and testing**

The next step includes training and testing. For each category in your classification scheme, you need to provide a sufficient set of relevant content so that the module can learn from it. Take content from actual document sets, which are as up-to-date as possible. This information is known as a *content set* and provides the inherent domain knowledge that Classification Module uses to analyze new content.
The learning is maintained in an entity called a *knowledge base*. Classification Module allows you to test your knowledge base with known content so that you can be confident of its accuracy. You can further augment the textual analysis capabilities of a knowledge base with the addition of rule-based analysis, which is provided through a decision plan. A *decision plan* can provide a rule-driven framework to invoke multiple knowledge bases in a single call to Classification Module. Several knowledge bases covering separate domains can be maintained so that separate types of classification can be carried out and multiple classifications can be assigned to a single document.

We describe detailed training and testing in later chapters of this book.

**Deployment and feedback**

After you have created and tested the knowledge base and, optionally, the decision plan, you deploy them for use within the context of a business application. In a typical application, content is submitted to Classification Module and a result set is received and acted upon. The best practice is to perform a review to ensure that the correct classification is maintained even when there are subtle changes in new content that is being submitted. This approach highlights one of the most powerful aspects of IBM Classification Manager. If, during review, a different classification is manually decided upon from that suggested by the module, that decision can be fed back to the knowledge base and can affect future classification decisions. In other words, Classification Module learns from experience and becomes smarter over time. Figure 1-3 on page 9 shows the operational schematic of how Classification Module is configured and deployed.
1.3 Business use cases

You can use Classification Module in many areas of a business to perform a multitude of decision-based tasks through the use of specialized decision plans and knowledge bases. You can use Classification Module in many areas to produce a strong return on investment (ROI):

- **Document classification:**
  - Classifying documents according to their topics

- **Compliance:**
  - Enhancing discoverability of content
  - Automating records classification processes

- **E-mail management:**
  - Automating classification
  - Enabling automated response and prioritization
  - Detecting spam
  - Filtering
Customer communication management:
- Routing customer communication efficiently
- Enabling automated query response
- Providing self-help (FAQ)

Enterprise content management:
- Generating taxonomy
- Ingesting content
- Creating content-centric workflow processes
- Managing records

1.3.1 Achieving and maintaining regulatory compliance

Companies across the world are grappling with the challenges of complying with regulations in information management. The main issues that companies face center on understanding the information (primarily electronic) that they have stored on file shares and information management systems across their organizations.

A partial solution
Classification Module allows you to provide accurate classification and metadata information for these documents to enable timely search and retrieval of the information. Classification Module also allows businesses to apply appropriate access controls on this information. However, these functions are not the full picture. A truly compliant system must also ensure that information is not kept for longer than is necessary and that out-of-date information is disposed of in a timely manner. Because of these requirements, organizations tend to implement a certified records management system.

The full picture
Records management systems typically require the use of a records management file plan: a hierarchical classification scheme, which is often designed around document retention policies within differing parts of a business. This structuring makes it challenging to map document metadata, which is primarily oriented to the business operational profile, to the records management file plan, which is oriented to document types and retention policies.

Traditional solutions to this challenge include developing complex rules engines to determine for any given document which file plan classification to apply, or, alternatively, making the classification process largely manual. A combination of the two approaches is often implemented, resulting in a reduction in accuracy and consistency.
Classification Module for records declaration

Classification Module provides real value in records management scenarios. By running the classification process against existing repository content, Classification Module can further classify any declared records according to the file plan classification scheme. Figure 1-4 illustrates the automatic mapping of classified business content to the file plan.

![Business view of content](image)

![Records file plan view](image)

**Figure 1-4  Automatic mapping to records file plan**

### 1.3.2 E-mail management

The explosive growth of e-mail communication presents the greatest challenge to information management in most organizations. E-mail is used for both business and private purposes. Even when used exclusively for business, there is often little or no contextual information that can be consistently relied on to accurately determine the meaning of an e-mail. Complicating matters further is the fact that important content can often be found buried in lengthy e-mail threads where the original subject has long since been abandoned and the substantive thread focus has drifted significantly.

Classification Module is able to sift through a company’s e-mails and separate the personal from the business, the trivial from the important, and categorize the
important e-mails appropriately. This automation can save hundreds of hours of manual effort and can make the difference between being compliant with regulations or not.

E-mail management goes beyond organizing a company’s stored e-mails. Incoming and outgoing e-mails also need to be actively managed so that they are accurately categorized on receipt/send.

Classification Module can provide spam detection, separating important e-mails from spam or other types of unimportant messages.

You can also use Classification Module for filtering by identifying messages of special interest and pinpointing a small percentage of relevant texts from a large body of texts.

1.3.3 Customer communication management

In both the private and public sector, companies increasingly recognize that each customer interaction is an opportunity to improve the relationship. Because much of this interaction is written, either as e-mail, online chat, or physical paper, the efficient handling of this communication is of paramount importance.

The primary challenges here are:

- Ensuring any inquiry is routed to the correct resource for resolution
- Replying to the communication accurately and quickly
- Using the correct channel according to the inquiry

Classification Module can identify various types of communication, determine their intent, learn to whom to direct the communication, and correctly ascertain the most appropriate response.

For example, Classification Module can be used for e-mail response applications. Use Classification Module to analyze and respond to customer e-mail inquiries. When Classification Module returns a high score for a specific category (for example, above a user-specified threshold), the application automatically sends the appropriate answer to the customer without human intervention.

Classification Module can also serve as an advisor to agents by indicating the most likely e-mail responses that are suggested by the system. You can use it to find and route a message to the most appropriate agent or queue. You can also use it to filter out redundant messages, such as spam.
In the case of inbound e-mails to an inquiry desk, Classification Module can assist in either analyzing the e-mail and directing the inquiry to the appropriate staff, or it can decide which standard response to use, given the nature of the inquiry.

In the case of self-help or frequently asked questions (FAQ) systems, you can use Classification Module to return the most likely answers to an inquiry submitted to a self-help application.

### 1.3.4 Enterprise content management

Many organizations recognize the need to implement an enterprise content management (ECM) system but are confronted by seemingly insurmountable challenges, which delay implementation. Next, we describe a number of these challenges and how you can address them with Classification Module.

**Taxonomy generation**

One of the primary objectives of moving to an ECM system is the ability to apply structure and more granular control to the unstructured information in an organization. Deciding on an appropriate classification scheme can be extremely time-consuming.

Classification Module includes a tool specifically designed to propose a category list (a *taxonomy*) based on an analysis of a representative set of content. It divides the content into conceptually similar clusters and then proposes names for them (which organizations have the option to rename later). You can run the process in an unattended manner, or you can run it interactively with a user who can guide the system to fine-tune the clustering and naming exercise.

Taking this approach, an organization can save significant time and effort in arriving at a classification scheme, which can be used as the basis for intelligently importing the large numbers of files that typically reside on network file shares.

**Initial content loading**

Most often, the biggest line item in the cost of implementing a new content management system is the effort that is required to move content from an existing system to a new repository, because typically, there is no easy way of mapping from one organizational structure to another organizational structure (see Figure 1-5 on page 14). Manual approaches to solving this problem are not only costly but also present one of the major barriers to user acceptance of the new system. Users do not trust the new system to accurately find existing information, because they know human error is unavoidable and misclassification will result.
The use of Classification Module in this scenario can ensure that all content is moved into the new environment accurately and consistently regardless of the size of the task.

![Figure 1-5  Mapping old structures to new](image)

After this stage of content loading is complete, Classification Module can continue to provide an interface to the new system for content being produced by line of business applications that do not have their own direct interface into the new content repository.

### Content-centric workflow processes

An important benefit of ECM is the introduction of content-centric workflow processing. This feature has the potential for vastly reducing the time taken to complete tasks, such as opening an account at a financial institution or initiating a benefit claim at a government body, by presenting the relevant documentation at each decision point in the process. Content-centric processing relieves the knowledge worker of the need to search for and retrieve the documentation that is needed to make a proper decision. However, human interaction itself introduces delay and invariably an element of inaccuracy or inconsistency across multiple manual touch points of the process.

Classification Module can be used to provide a score to a piece of content and to correctly route content or messages based on the score. This process replaces the human element in many workflows, thus reducing end-to-end process
latency, as well as eliminating the potential for introducing errors. Figure 1-6 shows how you can use the module in the context of such a process.

**Records management**
Records management is a critical element of an ECM system’s ability to deliver compliance within an organization. We describe compliance and records management in 1.3.1, “Achieving and maintaining regulatory compliance” on page 10.

### 1.3.5 Other use cases

The use cases just mentioned do not represent an exhaustive list of the ways in which Classification Module can be put to use. Organizations can deploy the technology to solve many unstructured text management challenges, including:

- Auto-tagging text for various search technologies
- Lotus® Quickr™ location suggestions for saved Sametime® instant chat dialogs
- Automatic routing of mail from a mailroom scanning facility
1.4 Integration options

You can implement Classification Module in a number of ways. In this section, we briefly describe the integration options available to implement Classification Module solutions with or without other products.

1.4.1 Stand-alone Classification Module

Classification Module has been designed as a stand-alone tool, which can be integrated with other applications through the SOAP interface (Figure 1-7 on page 17).

What it provides

When implemented as a Web service or as part of a service-oriented architecture (SOA), all applications within an organization can use Classification Module for immediate classification or metadata tagging of a specific text stream. Both the provision of input content and the handling of the output results are the responsibility of the implementer. Classification Module has extensible and well-documented application programming interfaces (APIs) in Java™, C, and Component Object Model (COM), as well as Web Services Description Language (WSDL).

Sample code that is delivered as part of the installation process provides examples of integration with Microsoft® SharePoint® Web parts and Microsoft Word 2007. As an example, for a Microsoft SharePoint integration, you can enable a workflow to call IBM Classification Manager and add metadata, such as a category, to a Microsoft SharePoint document. For a Microsoft Word 2007 integration, you can automatically suggest a category for a document when it is either open or saved.

Typical use cases

Many applications for Classification Module require the ingestion of the classified document into a content repository. At the time of writing this book, the supported content repositories for Classification Module are IBM Content Manager and IBM FileNet Content Manager. If you work with other repositories, use the APIs that are provided by your repositories to invoke Classification Module capabilities.
1.4.2 Integration with IBM FileNet P8

Classification Module Version 8.6 provides integration with IBM FileNet P8.

**What it provides**

IBM FileNet P8 integration provides a Web-based application called the *Classification Center*. The Classification Center enables IBM FileNet P8 documents from one or more folders to be submitted to Classification Module for automatic classification. Through the decision plan and its IBM FileNet specific functions, documents in IBM FileNet P8 content repository can be moved from their existing folders to other locations, metadata can be added to the documents, and documents can even be declared as records with an automatically chosen file plan location.

**Typical use cases**

IBM FileNet P8 clients who produce documents outside IBM FileNet Workplace XT need to consider using the following integration. You can easily integrate applications that produce documents with IBM FileNet P8, because you might not need to ascertain document classes and other metadata prior to the documents being ingested into the content repository. After the documents are in the IBM FileNet P8 system, the Classification Center can determine the correct document classes, folder locations, and even the correct file plan locations if the documents are declared records (see Figure 1-8 on page 18) and process them appropriately.
1.4.3 Integration with IBM Content Collector

Many organizations need to move large numbers of documents from file systems or e-mail systems into an ECM system, both during initial data loading, as well as on an ongoing basis. IBM Content Collector provides a fully comprehensive and standardized method of moving content from file shares and e-mail systems into either IBM FileNet P8 or IBM Content Manager content repositories. IBM Content Collector is licensed as two separate products: IBM Content Collector for File Systems and IBM Content Collector for E-Mail. This section addresses them both as a single entity.

Typically, an ingestion process comprises several stages with data filtering or analysis occurring at each stage. Consequently, IBM Content Collector is designed as a task-oriented framework. One of the tasks is calling Classification Module to determine a category for an individual file or e-mail.

Note: Starting from Version 8.7, Classification Module also provides predefined integration with IBM Content Manager.
IBM Content Collector integration benefits
IBM Content Collector provides a common task-driven framework for crawling file systems or e-mail systems and for ingesting the contents into an ECM platform, such as IBM FileNet P8 or IBM Content Manager. IBM Content Collector can analyze files or e-mails in context and make use of filenames, filename structures, directory and path information in the case of files, and mailbox and e-mail metadata in the case of e-mails. IBM Content Collector can evaluate filename structures and make task route decisions on those structures or other contextual file information.

When calling Classification Module, IBM Content Collector can add a further criterion to its task route decision. It can decide which route to take based on the classification score of the file in question. For example, if the score is above 80%, continue with the task; if the score is not, discontinue the route (ignore the file). These thresholds are configurable.

Using IBM Content Collector, organizations can quickly integrate e-mail and file producing systems with an enterprise content repository without using programming (see Figure 1-9).

Figure 1-9   IBM Content Collector integration with IBM Content Manager (CM8)
**Typical use cases**

Use IBM Content Collector integration where a significant number of documents need to be loaded into an ECM system with full metadata tagging for folder identification, categorization, or even records declaration and classification.

IBM Content Collector, together with Classification Module, can intelligently filter out irrelevant documents and provide alternative processing options for required documents based on source directory name, file name, or even classification score.
Concepts, architecture, tools, and integration

To get the most value out of IBM Classification Module (Classification Module), you need to understand the key classification technology concepts, classification server architecture, Classification Module core components, and integration components to other enterprise content management (ECM) systems.

Classification Module provides capabilities to automatically classify documents in categories. It comes with several graphical interface tools to help you set up, administer, and monitor an automated classification system for your enterprise content needs.

This chapter discusses the following topics:
- Classification concepts
- Core components
- Classification Workbench
- Classification Module server
- Classification Module integration architecture
2.1 Classification concepts

Classification Module uses natural language processing and sophisticated semantic analysis techniques to analyze and categorize text. Natural language analysis requires the use of terminology and concepts that might not be familiar to you. In addition, Classification Module uses specific terms to identify several of its own unique built-in features.

When you configure automated classification for your enterprise, your first and foremost task is to provide Classification Module server with a sample set of data and the categories to which that it belongs. You need to create and train knowledge bases and a decision plan, and you need to set appropriate content fields. We describe the following key concepts in this section to serve as the foundation for understanding Classification Module:

- Knowledge base
- Classification workflow
- Core classification technology
- Decision plan
- Field definitions

2.1.1 Knowledge base

A knowledge base is an object containing the learned information that Classification Module needs in order to perform matching, training, and online learning. This knowledge base is filled with relevant statistical and semantic information derived from a number of sample texts (known as a corpus) by using a process known as training or online learning (also known as feedback).

The sample texts need to represent the kinds of text that the system is expected to handle. The statistical information can consist of entities, such as words in the text, the number of occurrences of those words, hints about the text, and the distance between words. The information reflects the kind of data that the system is expected to handle and can be stored in a file or a database.

Categories are the basic elements within the knowledge base. Categories can represent the textual content of a text, or they can indicate another attribute, such as its source. A category has a set of features, which are known as concepts, that characterizes a category and distinguishes it from other categories. The creation and maintenance of these concepts are internal to Classification Module and are not controlled by the user. Instead, the creation and maintenance of these concepts are affected only by the training or learning processes.
2.1.2 Classification workflow

The following scenario describes Classification Module’s typical classification workflow:

1. A Classification Module-enabled software application sends input text to Classification Module.

2. Classification Module analyzes the input text and computes relevancy scores for each category in the knowledge base. It might also extract information, such as order numbers or zip codes from the text. This process is called matching or categorization.

3. You can maintain and improve the performance of a knowledge base over time by providing it with online feedback text and its correct category. This process of automatic self-adjustment is called online learning.

Classification Module works with an adaptive knowledge base to analyze and categorize texts. Before the knowledge base can analyze texts, it must be trained with a sufficient number of sample texts that are properly classified into categories. A trained knowledge base can take a text and compute a numerical measurement of its relevancy to each category. This process is called matching or categorization. The numerical measurement is called relevancy, score, or relevancy score.

Client applications use categories to denote the intention of text. When text is sent for matching, the knowledge base data is used to select the category that is most likely to match the text.

You can maintain and improve the accuracy of a knowledge base over time by providing it with feedback (confirmation or correction of the current categorization). The feedback is used to automatically update and improve the accuracy of the knowledge base over time. This process of automatic self-adjustment is called learning. Learning is an incremental process that simulates real life, where recently learned information has a greater impact on classification.

The Classification Module platform provides Classification Workbench, a core component, that can quickly create, analyze, tune, and maintain knowledge bases. Figure 2-1 on page 24 shows a knowledge base (KB) project using the Classification Workbench tool.
Figure 2-1   A KB project for HR categories

Note: For step-by-step instructions to create a knowledge base using Classification Workbench, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.
2.1.3 Core classification technology

The IBM classification technology uses a three-phase process to classify input texts:

- **Document Filter Manager**: Strips out the rich formats that are applied to documents, such as Word, Excel®, and PDF documents. The document filter identifies the language encoding and converts any other language encoding into the standard Unicode Transformation Encoding Format (UTEF) of the plain text.

- **NLP (Natural Language Processing) Engine**: Extracts concepts from free-text fields and generates a Semantic Modeling Language (SML) document. A concept is a basic unit of linguistic or quantitative information that is derived from input text (and its context) that can influence classification.

- **Semantic Modeling Engine**: Performs statistical pattern matching on the SML by comparing it with the content of categories residing in the KB to generate the relevancy scores top matching categories.

Figure 2-2 illustrates the classification technology.
2.1.4 Decision plan

In addition to classification using the knowledge base, Classification Module can perform rule-based classification with the help of decision plans. A decision plan is a collection of rules that you configure to determine how Classification Module classifies content items. Unlike a knowledge base, which can be built automatically by supplying a sample set of precategorized content items, you build a decision plan by creating and configuring one or more Rules.

Rules consist of Triggers and Actions. A trigger determines the conditions that must be met to initiate an Action. Each rule has exactly one trigger; however, it can have multiple actions. An action specifies what action Classification Module will perform if an associated rule is triggered and returns true. The action can be to store a document in a specific folder or document class, to move or copy a document from one folder to another folder, to declare a document as a record so that it can be placed under the control of a records management system, or to extract and update metadata information from the document.

You can also configure triggers and actions based on content or metadata. A decision plan can use one or more knowledge bases for a combination of rule-based and knowledge-based classification. For example, you can add a knowledge base to your decision plan project, and you can define rules that are based on matches and scores.

You can combine multiple rules into a group. A group is a logical collection of triggers and actions designed to achieve a certain task. You can create multiple groups within a decision plan and sequence them. Each group can then be dynamically set to be activated or not to be activated based on the rule evaluation of any previous group in the sequence.

You can use the Classification Workbench to configure decision plans and groups of decision plans. Figure 2-3 on page 27 shows a decision plan project within Classification Workbench tool, which uses a KB as a reference project.

**Note:** It is beyond the scope of this book to discuss in further detail how the technology works internally to perform classification. We highlight the three-phase process instead of going into more detail.
Figure 2-3  A decision plan project

**Note:** For instructions to create a decision plan using Classification Workbench, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.
2.1.5 Field definitions

To support flexible text input, the system needs to identify the content types and data types of the text that it receives. The Classification Module receives text as a series of fields. A field definition (a name-value pair or NVP) defines the data type and content type of each field.

Content type definitions are used in determining the final relevancy score of a document. Higher weight is assigned if matches are found against a field with a subject content type as opposed to matches found against a field with a body content type.

Figure 2-4 shows a sample field definition property association dialog box using the Classification Workbench tool.

![Figure 2-4 Field definition settings within Classification Workbench](image)
2.2 Core components

Classification Module is an enterprise platform for a wide range of applications that require unstructured content to be automatically categorized. Applications can include document classification, taxonomy automation within ECM systems, e-mail classification and routing, auto-response and archive tagging, records class declaration and retention policy assignment for records management, and Web self-service systems.

Classification Module consists of the following core components:

- Classification Workbench:
  - Taxonomy Proposer
- Classification Module server:
  - Management Console
  - Client APIs
- IBM FileNet P8 integration asset:
  - Classification Center
  - Content Extractor

In this section, we describe these components and how Classification Module is put to work in a typical workflow.

2.2.1 Classification Workbench

Classification Workbench is a Windows application that allows you to create, analyze, and evaluate the performance of knowledge bases and decision plans. The result is an optimized knowledge base that can be used with the Classification Module server and other applications using built-in APIs. Whether you develop your own applications using APIs to access Classification Module server or if you use existing applications, you use Classification Workbench extensively to create, train, and analyze knowledge bases and decision plans.

The Classification Module server continuously learns by gathering feedback, improving the effectiveness and accuracy of the knowledge over time. Classification Workbench provides a way to pre-train the system, so that you do not have to wait for the incremental learning process to take effect. You can also...
use Classification Workbench to apply learning to an existing knowledge base (KB), and improve its effectiveness.

Typically, you create a KB in stages. Prior to using Classification Workbench, you collect sample data (for example, e-mails and documents) representative of the data that you expect to classify using your own applications or scripts under a file system or an enterprise content management system. Then, you arrange the collected sample data under categories or logical groups that you define for your enterprise.

You import this data (including sample documents and categories) into Classification Workbench to create a content set file. Classification Workbench provides a variety of features and techniques that allow you to fine-tune the content set to optimize knowledge base accuracy. The knowledge base is a single file encapsulating data that the Classification Module requires for accurate classification. The KB will change and improve over time as it learns new categories and adapts to changes in data received by the system.

Refer to 2.3, “Classification Workbench” on page 34 for a description of the functions of Classification Workbench. Chapter 3, “Working with knowledge bases and decision plans” on page 59 provides a detailed overview and step-by-step instructions to work with Classification Workbench.

2.2.2 Classification Module server

Figure 2-5 on page 31 shows the Classification Module server system architecture. The system architecture consists of the following components:

► IBM Classification Server
► Data Server
► SOAP Layer
► Client APIs
► Management Console

Note: In case a predefined category does not exist, there are several ways in which Classification Module can help you with categorization needs. One way is to use a tool called *Taxonomy Proposer*. For more information about Taxonomy Proposer, refer to 2.3.2, “Taxonomy Proposer” on page 37.
IBM Classification Server

IBM Classification Server is the server for client applications that interact with Classification Module. It is the core of the system. It provides various services, such as language analysis, instance pooling of knowledge bases, and management of the distribution of tasks across computers.

The IBM Classification Server contains the following server components:

- Listener
- Read/write process
- Read/only process
- Administration process

Listener

This component serves as the entry point to the system. Client requests are sent to the listener, which then dispatches them to the appropriate server-side component for processing. Requests related to a specific knowledge base are
routed to the corresponding read-write instance, and administration requests are routed to the administration component.

**Read-write process**
This process handles requests to the knowledge base or decision plan, such as matching, feedback, language identification, and modifications made to the knowledge base or decision plan. There is only one read-write instance for each knowledge base (KB R/W) and one read-write instance for a decision plan (DP R/W). Depending on the workload, the read-write instance might use a load-balancing algorithm to forward read-only requests to an available read-only instance.

**Read-only process**
This optional process handles read-only requests that are forwarded by the read-write instance. Common read-only requests include matching, language identification, and viewing the structure of a knowledge base or decision plan. You can configure any number of read-only instances for a given knowledge base or decision plan and configure them to run on any number of computers. You can have multiple knowledge base read-only instances (KB R/O) and multiple decision plan read-only instances (DP R/O).

**Administration process**
This process handles all global administration requests, including requests that originate from the Classification Module administration tool called Management Console. Refer to 2.4.1, “Management Console” on page 40 for a discussion of this administration tool. The administration process is configured to run on a specific server when the Classification Module is installed. Communication between the various server-side components is accomplished through SOAP.

Figure 2-6 on page 33 shows a sample configuration of a Classification Module server in a single server mode.
Chapter 2. Concepts, architecture, tools, and integration

**Data server**

The *data server* (D-server) is a proprietary data storage mechanism that is used by the Classification Module server to persist information that is required by the Classification Module server. Types of information include server configuration information, knowledge base information, decision plan information, and feedback information.

**SOAP layer**

Applications can interact with the Classification Module system by using SOAP. The SOAP layer wraps the native Classification Module function calls. The definition for the SOAP interface is provided by a Web Services Description Language (WSDL). For example, the Classification Module might be included as a Web reference in any .NET application using the WSDL.

**Client APIs**

Classification Module server provides client APIs to interface with applications to perform data layout and formatting. Classification Module includes C, Component Object Model (COM), and Java client APIs. Classification Module supports several configuration options to accommodate varying client needs. Regardless of the type of system configuration that you choose, you can administer all system components from a single point, because all configuration data is stored in a common data server. In most cases, you develop your own clients that fit in with the rest of your application suite.
Management Console
Management Console is the application that you use to administer a Classification Module system.

2.2.3 IBM FileNet P8 integration asset

There are a wide variety of applications that you can build using the Classification Module server layer and the core classification technology. Several predefined applications come with the product. In other cases, clients and users can use the rich set of APIs provided through the SOAP interface to build their own applications to use classification services. Classification Module can be provided with applications, such as IBM Content Collector. It can also integrate into existing IBM ECM products, such as IBM FileNet P8. Classification Module can integrate with the following applications (or platform):

- IBM FileNet P8
- IBM Content Collector for file systems
- IBM Content Collector for E-mail

Special integration assets are provided as IBM services offerings to integrate Classification Module with the following products:

- IBM OmniFind Enterprise Edition
- Microsoft SharePoint Web Environment

Refer to 2.6, “Classification Module integration architecture” on page 49 for a discussion about the Classification Module integration architecture.

Classification Center
If you integrate Classification Module with IBM FileNet P8, you can use the Classification Center to manage the classification processes.

Content Extractor
The Content Extractor is a command-line tool that you use to extract the content from an IBM FileNet P8 object store.

2.3 Classification Workbench

Classification Workbench is a Windows® application that allows you to create and manage a knowledge base and decision plan, analyze the knowledge base performance using reports and graphical diagnostics, and work with a collection of sample text known as a corpus for analysis, training, and learning.
With Classification Workbench, you can import a knowledge base or decision plan for analysis and fine-tuning. After fine-tuning the performance, you can export the knowledge base or decision plan back to the Classification Module server.

Figure 2-7 shows you the Classification Workbench application, which was started using **Start** → **Programs** → **IBM Classification Module 8.6** → **Classification Workbench**.

Client applications for Classification Module interact with knowledge bases and decision plans in a variety of ways. In this section, we discuss how you can perform these tasks:

- Create and configure knowledge bases
- Create and configure decision plans
2.3.1 Create and configure knowledge base project overview

You can create a knowledge base from an existing set of organized data. Using the Classification Workbench tool, the typical process of creating a knowledge base is summarized in the following steps:

1. Prepare and import sample content.
   Begin by collecting and organizing sample content, for example, documents or e-mails that are similar to the type of content that you plan to classify using Classification Module. Then, import your content into Classification Workbench by using the Import wizard.

2. Configure fields for natural language processing.
   In Classification Workbench, each content item is defined by one or more fields. You must configure at least one field that contains textual content for natural language processing.

3. Create and analyze a knowledge base.
   You create and analyze a knowledge base using the Create, Analyze and Learn Wizard. Typically, half of your content set is used to train the knowledge base, and the other half of your content is used for testing.

For detailed steps, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.

Create and configure decision plan project overview

You create a decision plan for rule-based classification by configuring triggers and actions that are based on content or metadata. Your decision plan can refer to one or more knowledge bases for a combination of rule-based and content-based classification.

Using the Classification Workbench tool, the typical process of creating a decision plan is summarized in the following steps:

1. Define rule triggers and actions.
2. Analyze the decision plan using sample content.
3. View reports to check performance.
4. Tune the decision plan if necessary.

You can create a decision plan that includes knowledge base classification. The steps that you follow to create a decision plan that includes knowledge base classification are:

1. Create a knowledge base.
2. Add the knowledge base to your decision plan project.
3. Define rules based on matches and scores.
You can run your sample content set items through the decision plan to see if the appropriate action was taken on the required content set items. You can check the performance of a decision plan in production by exporting analysis data from the Classification Module server and importing the data into Classification Workbench.

For detailed steps, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.

2.3.2 Taxonomy Proposer

The Taxonomy Proposer, which is installed with the Classification Workbench, allows you to discover new categories in an uncategorized or partially categorized body of documents. The Taxonomy Proposer uses custom clustering algorithms to analyze and group similar documents to help you to create a taxonomy for your content: the division of content into ordered groups or categories. Figure 2-8 is the pictorial representation of the Taxonomy Proposer.

![Figure 2-8  Taxonomy Proposer](image)

After you run the Taxonomy Proposer, you can create a knowledge base for Classification Module by manually reviewing documents in each cluster or suggested category and renaming the suggested categories based on document content. You import the categorized content set into Classification Workbench to build the knowledge base and, later, fine-tune the knowledge base performance.

The Taxonomy Proposer is installed with Classification Workbench and runs in a Windows environment.
To start the Taxonomy Proposer, double-click the `TaxonomyProposer86U.exe` file that is located in the `ICM_home\Classification Workbench\Program Files` directory.

The Taxonomy Proposer displays the Workflow Assistant, which guides you through typical scenarios. You can also access the Workflow Assistant from the Help menu. Figure 2-9 shows the Taxonomy Proposer Workflow Assistant.
2.4 Classification Module server

Classification Module server needs to be aware of knowledge bases and a decision plan before it can act upon documents to classify and provide suggestions with confidence scores. Figure 2-10 shows a typical workflow of a Classification Module system. Figure 2-10 explains the key steps that are involved in making automatic classification using the Classification Module platform work.

Classified documents (with well-defined structured categories) are imported into the Classification Workbench application as a content set. Unclassified documents can be assigned automatically generated categories by processing these documents through Taxonomy Proposer and then importing them into Classification workbench as a content set. In the Classification Workbench, the content set creates (or trains) a knowledge base. The knowledge base can be analyzed by another content set to understand the accuracy and effectiveness of the knowledge base. Active feedbacks can be applied to help the system to learn about recent changes to the knowledge base. You can also apply other techniques as described in Chapter 3, “Working with knowledge bases and decision plans” on page 59 to fine-tune the knowledge base. This process is called tuning.

Note: We encourage you to have your own predefined categorization. It is beyond the scope of this book to cover Taxonomy Proposer in detail. We recommend using the Workflow Assistant for guidance on using the Taxonomy Proposer.
The result of the Classification Workbench process of training, analyzing, and tuning is a knowledge base that, when combined with a decision plan, can then be used by a Classification Module server to run the training, analyzing, and tuning as internal processes. After the Classification Module server is running with knowledge base read-only or read-write processes, decision plan read-only or read-write processes, applications, and other clients can invoke functions to this Classification Module server using APIs asking for classification services, such as suggest and decide. Certain IBM ECM applications, such as IBM FileNet P8, can connect directly to the Classification Module server when the integration is enabled. Classification Module in this case can classify documents and automatically file them in the correct folder structure in IBM FileNet P8. We discuss other use case scenarios in the second half of this book.

When it comes to the management capabilities of Classification Module server, we discuss the administrative graphical interface of the Classification Module system and the various Classification Module server deployment configurations. In this section, we describe Management Console and Classification Module server deployment configurations.

### 2.4.1 Management Console

Management Console is the application that you use to administer a Classification Module system. For example, you can add knowledge bases and a decision plan to the Classification Module server, add and modify field definitions, and view information about the servers that host Classification Module components.

With Classification Module Management Console, you can perform the following functions:

- Knowledge base administration
- Decision plan administration
- Content field definition administration
- Classification server administration

To start the Management Console application, click **Start → Programs → IBM Classification Module 8.6 → Management Console**. Figure 2-11 on page 41 shows the initial connection parameter window, which confirms the connection parameters to the Management Console server.
After you connect to the Management Console, you see the Management Console application interface. Figure 2-12 shows the Management Console application with all the administration tabs available.

**Knowledge base administration**

You create and train knowledge bases by using Classification Workbench and importing them to the Classification Module server. You use the Management Console to administer knowledge bases. With the Management Console, you can perform the following tasks:

- Add and remove knowledge bases to the Classification Module server. The knowledge base that is created using the Classification Workbench can dynamically be added to and removed from the Classification Module server.
- View a list of all knowledge bases on the Classification Module server.
- Start and stop knowledge bases. In order for the Classification Module server to start classifying documents based on certain set of knowledge bases, the knowledge bases need to be started. To turn off a knowledge base inside the Classification Module server, you can stop the knowledge base.
- Export a knowledge base to a file system folder.
- Modify knowledge base properties.
Remove knowledge bases from the Classification Module server.

Figure 2-13 shows the Knowledge bases administration tab inside the Management Console application. For detailed administration, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.

**Decision plan administration**

You create a decision plan by using Classification Workbench. You use the Management Console to administer the decision plan. With the Management Console, you can perform the following tasks:

- Add decision plans to the Classification Module server. The decision plan that is created using the Classification Workbench can be added to and removed from the Classification Module server dynamically.
- View a list of all decision plans on the Classification Module server.
- Start and stop decision plans. In order for the Classification Module server to start classifying documents based on a certain set of decision plans, those decision plans need to be started. To turn off a decision plan inside the Classification Module server, you can stop the decision plan.
- Export a decision plan to a file system folder.
- Modify decision plan properties.
- Remove decision plans from the Classification Module server.

Figure 2-14 on page 43 shows the Decision plan administration tab inside the Management Console application. For more information, refer to Chapter 3, “Working with knowledge bases and decision plans” on page 59.
Content field definition administration

The Classification Module receives texts as a series of fields. The field definition defines the data type and the method of language processing that is performed on the field. A single set of fields is used throughout the system. When you use the Management Console to administer field definitions, you can perform these tasks:

- Add a field to the Classification Module server.
- View and modify field properties.
- Import or export a set of fields.
- Delete a field from the system.

Figure 2-15 on page 44 shows the Field definitions administration tab inside the Management Console application.
Classification server administration

You can use the Management Console to see which computers are running Classification Module components.

To view a list of servers that are running Classification Module components, in the Management Console application, click **Servers**. The host names are displayed in the right pane. Your local computer automatically appears in the list of available servers.

To view information about the knowledge bases and decision plans on a specific server, double-click a server name in the right pane. Information about each knowledge base or decision plan is displayed with the following information:

- The knowledge base or decision plan name
- Whether the knowledge base or decision plan runs as a read-write or read-only instance on the server
- The server port assigned to the knowledge base or decision plan
- The status of the knowledge base or decision plan (running or not running)

Figure 2-16 on page 45 shows the Servers administration tab inside the Management Console application. It lists all the host names that are running the Classification Module server.
Figure 2-17 shows the Servers administration tab inside the Management Console with the host name view. It shows all the knowledge base and decision plan processes running on a particular host.

2.4.2 Classification Module server deployment configurations

There are various deployment scenarios in which Classification Module server can be configured to maximize your enterprise IT capital.

You can configure Classification Module server in the following three ways:

- Single server configuration
- Multiple server configuration
- Multiple listener configuration

Single server configuration

In the single server configuration, all server-side components and processes, such as the listener process, administration process, read-write processes (for
knowledge base and decision plan), and read-only processes (for knowledge base and decision plan), run on the same computer. A single server configuration is useful on multiprocessor computers and in small-scale or development environments. On a single server, there can only be one listener, one knowledge base read-write (KB R/W), and one decision plan read-write process (DP R/W). However, you can have multiple read-only processes (KB R/O) for the same knowledge base or multiple read-only processes (DP R/O) for the same decision plan on a single server.

Figure 2-18 shows the single server configuration mode of the Classification Module server.

Clients are either custom applications that use the client libraries or applications for integration with IBM FileNet P8.

The applications communicate with the listener that runs within the server through SOAP. The listener forwards the request to the appropriate server-side component for handling. Administration requests are forwarded to the administration process, and requests on a specific knowledge base or decision plan are routed to the read-write instance of that knowledge base or decision plan. Depending on the nature of the request and the workload, the read-write process either handles the request by itself or forwards the request to a read-only instance of the knowledge base or decision plan.
Multiple server configuration

To promote scalability, server-side components can be distributed across several servers. A multiple server configuration enables seamless load-balancing between multiple computers to accelerate response time and increase throughput. The load-balancing is done at various process levels. For example, the listener process can load-balance read-write processes for separate knowledge bases on multiple servers. The read-write processes can load-balance read-only processes on multiple servers. Distribution of the processes is flexible, because the system does not require that any two components run on the same computer.

Figure 2-19 shows the multiple server configuration mode of the Classification Module server.

Multiple listener configuration

You can install and run multiple listener components on separate computers. A typical use for a multiple listener configuration is to ensure system availability for processing read-only requests, such as suggest requests and language identification requests. In this configuration, you run listener components under two or more separate servers, use an external tool to balance the load, and switch between the listeners in the event that one of the servers crashes. To
ensure the availability of a specific knowledge base, read-only instances are distributed on all of the servers.

Figure 2-20 shows the multiple listener configuration mode of the Classification Module server.

![Multiple listener configuration](image)

**Figure 2-20 Multiple listener configuration**

### 2.5 Classification Module APIs

The Classification Module Application Program Interface (API) is a C++ class library that is used to create applications that will interact directly with Classification Module. These applications can link their code with the Classification Module library and gain full access to Classification Module functions.

Client applications for Classification Module interact with knowledge bases and decision plans in a variety of ways. For example, using the available APIs, your applications can perform one or combination of the following tasks:

- Match texts against a knowledge base.
- Retrieve decision results from a decision plan.
- Submit feedback to a knowledge base or decision plan.
- Identify the language of texts.
- Support multilingual applications.
The Classification Module provides several client API libraries to enable the rapid development of various client applications in several programming languages, such as C, C++, C#, and Java, as well as scripting languages, such as Active Server Page (ASP) or VBScript. Samples that demonstrate system functionality and how to use the various client libraries are provided for several programming languages as part of the code samples with Classification Module installation.

Figure 2-21 shows the Classification Module client API architecture. The .NET clients interface with the Classification Module server using the WSDL directly with the SOAP layer. Other clients, such as Java, C, and COM use their corresponding APIs to invoke SOAP function calls on the Classification Module server.

**Figure 2-21  Classification Module client API architecture**

## 2.6 Classification Module integration architecture

As discussed in Chapter 1, “Overview” on page 3, the value of Classification Module can perhaps be best described by its integration into various IBM ECM products. Figure 2-22 on page 50 shows the Classification Module Integration diagram, listing all its components, tools, and integration components to ECM systems.
In this section, we describe applications that get installed along with the Classification Module FileNet P8 Integration platform. We discuss the following two applications:

- Classification Center
- Content Extractor

### 2.6.1 Classification Center

If you integrate Classification Module with IBM FileNet P8, you can use the Classification Center to manage the classification processes. You use this Web application to select the content to be classified, configure classification options (such as the decision plan to use and various runtime preferences), monitor classification activity, and view the classification results. You can also use the Classification Center to reclassify documents if you determine that other knowledge base categories or decision plan actions are more applicable.

Figure 2-23 on page 51 shows the Classification Center application architecture diagram. The Classification Center Web application runs inside its own open-source JETTY Web server. The Classification Center is based on Model-View-Controller (MVC) architecture and uses the Spring framework. The IBM FileNet P8 Connector component connects to the IBM FileNet P8 server.
whereas the Spring beans use the Classification Module APIs (using the SOAP layer) to invoke Web service methods to Classification Module server.

The Classification Center includes three major tasks that you use to configure classification options, start and monitor the classification processes, and review classification decisions. In this section, we describe each of the Classification Center tasks:

- Configuration
- Dashboard
- Review

Before you browse to the Classification Center application Web page, start the Classification Center server by running **Start → Programs → IBM Classification Module 8.6 → Classification Center → Start Classification Center server**. Figure 2-24 on page 52 shows the command window that shows the status of the Classification Center server.
After the Classification Center server is started, launch the Classification Center application by running Start → Programs → IBM Classification Module 8.6 → Classification Center → Classification Center. Figure 2-25 on page 53 shows the landing page of the Classification Center application.
You can specify how you want Classification Module to classify documents into folders and document classes in IBM FileNet P8.

To configure the setting, click the Configuration icon from the Classification Center main page.

You can configure the following settings:

- General Settings: Specify settings, such as the decision plan to use, the IBM FileNet P8 object store to use, the metadata mappings between the IBM
FileNet P8 document and the fields in Classification Module, and whether to declare documents as records.

- Content to Classify: Specify from what folders or document classes the content is to be classified.
- Runtime Settings: Select settings, such as how resources are used and the number of threads to use for classifying content.

Figure 2-26 shows the Configuration tab in the Classification Center application.
Dashboard
With the dashboard, you can start, stop, and monitor the classification processes. You can view process statistics, such as how many documents were classified. You can view summary information, such as which folders or document classes received the greatest number of documents. You can also view the event log and error log.

Figure 2-27 shows the Dashboard tab in the Classification Center application.

![Classification Center - Dashboard - Windows Internet Explorer](image)

Figure 2-27 Classification Center Dashboard tab

Review
With the Review tab, you can review the classification results and, if necessary, reclassify documents.

When reviewing a document, you can either confirm that the document is correctly classified or select various categories and actions and reclassify the document. By reviewing documents, you help verify that the system performs as expected and help ensure that the correct folders, document classes, and decision plan actions are applied during classification. In addition, when you review documents and manually select appropriate categories, the system learns from your selection, thereby improving future classification.
You can also add documents to the IBM FileNet P8 repository. When you add a document, Classification Module analyzes its content and suggests how to classify it. You can then review and confirm the actions or reclassify the document just as any other document that is available for review.

Figure 2-28 shows the Review tab in the Classification Center application.

Figure 2-28   Classification Center Review tab

**Note:** For more details about the usage of Classification Center, refer to 5.5, “Configuring and performing classification” on page 249.
2.6.2 Content Extractor

The Content Extractor is a command-line tool that you use to extract the content from an IBM FileNet P8 object store. The Content Extractor generates the XML output based on the document metadata. You can import the extracted content XML output into the Classification Workbench to be used as the content set to create and train the knowledge base.

Note: For more information about the usage of Content Extractor, refer to 5.4.1, “Using Content Extractor” on page 205.

2.7 Additional reference

The Classification Module Information Center provides excellent information about the product. We recommend that you use the information provided there in conjunction with this book. At the minimum, review the terms and definitions provided in the Glossary in the Information Center. You can use the terms and definitions as a reference when you encounter new terms or to review the terms that are used in this book.

You can access the IBM InfoSphere Classification Module Version 8.7 Information Center at this Web site:

http://publib.boulder.ibm.com/infocenter/classify/v8r7/
Working with knowledge bases and decision plans

Knowledge bases (KBs) and, optionally, decision plans are the essential components of any IBM Classification Module (Classification Module) system. This chapter introduces the basic techniques that are available in Classification Workbench to create, train, or analyze knowledge bases and decision plans.

This chapter describes the following topics:
- Importance of knowledge bases and decision plans
- Creating, training, and analyzing a knowledge base
- Creating and analyzing a decision plan
- Building a knowledge base advanced topics:
  - Using an uncategorized or partially categorized content set
  - Using keywords
  - Enabling the knowledge base with feedback processing
  - Working with offline (deferred) feedback
  - Handling overlapping categories
3.1 Importance of knowledge bases and decision plans

The most important aspect of creating a successful Classification Module implementation is defining the correct knowledge base and decision plan if required. The accuracy of your knowledge bases or the effectiveness of your decision plans directly determines how well the system performs the classification operation for your application.

A knowledge base provides a content-based classification approach to categorize your content by understanding the full intent of the content. It reflects your specific environment as illustrated in the following examples:

- If we want to classify files in a file system, we can build a list of the file categories and then teach Classification Module how to recognize them.
- If we want to classify customer e-mails, we can build a list of the topics that customers discuss via e-mail and then teach the system how to recognize them.

A decision plan provides a rule-based approach for analytical classification. It is a collection of rules and optionally with one or multiple knowledge bases. With the combination of the content-based and rule-based classification approaches, a decision plan enables the system to classify documents in various scenarios. For example, your system can use a decision plan to identify all documents relating to patent application requests, take actions to auto-classify those documents into the correct folders in the IBM FileNet P8 repository, and declare them as records under the management of IBM FileNet Records Manager at the same time.

This chapter describes the methods for creating a knowledge base and a decision plan and analyzing and tuning them. Classification Workbench is the tool that we use to perform these tasks.

3.2 Creating, training, and analyzing a knowledge base

The process of creating, training, and analyzing a knowledge base consists of multiple steps.

Prior to using Classification Workbench to create a knowledge base, you gather and pre-categorize data that forms a content set. Then, you use Classification Workbench to perform these tasks:

1. Import data.
2. Create and edit a content set.
3. Create, train, and analyze a knowledge base.

4. Evaluate knowledge base performance by generating and viewing summary reports and graphs.

5. As required, improve knowledge base performance by reanalyzing and learning.

Figure 3-1 illustrates the typical stages of knowledge base development. Your workflow can vary.

3.2.1 Preparing data for import

Classification Workbench uses a set of content items, such as files or e-mails, as the basis for creating and analyzing a knowledge base.

A typical example of preparing data for import is to create a file system folder structure with subfolders representing categories and, then, to place a set of sample files under each subfolder. The sample files that you choose must be representative of the information that will be classified using Classification Module in real life. When you import this file system folder structure into Classification Workbench, each file becomes a content item.

In this chapter, we plan to create an HR knowledge base with seven predefined categories, such as Health Care, Leave of Absence, and Stock Options. We gather and place a set of representative sample files into the folder structure, as shown in Figure 3-2 on page 62, where the folder names are consistent with the category names to be created in the knowledge base. After the files are imported into Classification Workbench, their corresponding content items are assigned with category names based on the file system folder location. For example, the
content items derived from all files in the Health Care folder have a category name called Health Care.

![Figure 3-2](image)

**Figure 3-2** File system folders as a content set

### 3.2.2 Building a knowledge base

The following steps summarize the creation of a knowledge base in Classification Workbench:

1. Start Classification Workbench.
2. Create a new knowledge base project.
3. Import a content set.
4. Assign the content type value to a content item’s field.
5. Create a knowledge base via the Create, Analyze and Learn Wizard.

**Starting Classification Workbench**

Click **Start → Programs → IBM Classification Module 8.6 → Classification Workbench**. Classification Workbench starts with the *Workflow Assistant* open, as shown in Figure 3-3 on page 63.
Creating a new knowledge base project

To create a new knowledge base project, follow these steps:

1. In the Workflow Assistant, as shown in Figure 3-3, click **Create a knowledge base project** to display the new project window.

2. In the New Project window, as shown in Figure 3-4 on page 64, perform the following steps:
   a. Enter the project name in the Name field. In this example, we enter HR.
   b. Select the project type in the Type field. In this example, we choose the project type of **Knowledge Base**.

![Figure 3-3 Starting Classification Workbench](image)
**Importing a content set**

Import a content set from a file system folder into Classification Workbench:

1. Select one of the following three options to create a new project:
   - Create a project by importing a content set.
   - Create a project by importing an existing knowledge base file.
   - Create an empty project.

   In this example, we select **Create a project by importing a content set**, as shown in Figure 3-5 on page 65.

![Figure 3-4 New knowledge base project window](image)
2. In the Import Content Set window, as shown in Figure 3-6, determine the type of content set that you want to import. In this example, we select the Files from a file system folder radio button.
3. In the content set location window, as shown in Figure 3-7, specify the fully qualified path for the root folder from which files are to be imported. All files in the root folder and its subfolders are imported. In this example, click **Browse**, and locate the **C:\HR** root folder.

![Import Content Set](image)

*Figure 3-7  Selecting the root folder of a content set*

4. In the file filter window, when the list of files and folders is displayed, select or clear the check boxes for the files and folders that you want to import. In addition, the **Apply Filters** button allows you to perform these tasks:
   - Select the folders that you want to include or exclude.
   - Select the file extensions for the files that you want to include or exclude.
   - Select an option to activate random file select.

In this example, we use the default filter settings to include all files for import, as shown in Figure 3-8 on page 67.
5. In the window to specify category names for each file system folder that you selected, you can edit the Category Name column with a new category name for the folder. In this example, we accept the default folder names as the category names, as shown in Figure 3-9 on page 68.
Figure 3-9  Editing category names for file system folders

6. Specify the language and text filter settings for the content set. In this example, we accept the default settings, as shown in Figure 3-10.

Figure 3-10  Specifying language and text filter settings

Figure 3-11 on page 69 shows the project view after importing the HR content set into the Classification Workbench.
Assigning content type

It is important to select an appropriate content type value for one or more fields containing meaningful text that you expect Classification Module to analyze. Do not assign a content type value for content fields containing non-textual values (for example, account numbers or telephone numbers) or non-meaningful text (for example, a content field containing arbitrary administrative comments about each content item).

Note: To create a knowledge base, you must specify at least one field (from the content set) to be analyzed by Classification Module. The field must contain meaningful text. To tell Classification Module that this field is to be analyzed, you assign a content type value to the field.

Depending on the content set that you have, you have various fields associated with the content set. For our example, the imported content set has the following three fields:

- Categories
- Body
- FileName

The Categories field contains the non-meaningful text of administrative comments, and the FileName field contains the non-textual value of file names. Only the Body field contains meaningful text that we intend for Classification Module to analyze.
In this example, we decide to assign the content type value to the Body field by double-clicking the **Body** field in the Project Details pane, as shown in Figure 3-11 on page 69, to display its Field Properties window.

In the Field Properties window, the *Content type* pull-down list box provides the following five options:

- **Body**: For e-mail environments only. Select this option for content fields that contain the main body text.
- **DocTitle**: Select this option for content fields that contain document titles.
- **Plain Text**: Select this option for content fields that contain textual content.
- **Sender**: For e-mail environments only. Select this option for the *Sender* or *From* field that contains the e-mail address.
- **Subject**: For e-mail environments only. Select this option for a *Subject* field that contains the subject of an e-mail message.

In this example, we select the **Plain Text** content type, as shown in Figure 3-12.

![Field Properties Window](image)

*Figure 3-12* The Body field properties

## Creating a knowledge base

When you use Classification Workbench to create and analyze a knowledge base in a single process, Classification Workbench trains and analyzes the knowledge base at the same time that it creates the knowledge base.
Run the *Create, Analyze and Learn Wizard* to create a knowledge base:

1. Click **Create, Analyze and Learn Wizard** on the toolbar to start the wizard.
2. Specify how to split the content set to create and analyze the knowledge base. Classification Workbench divides the content set in various ways, as indicated by the option names, as shown in Figure 3-13 on page 72.

   In this example, we select **Create and analyze knowledge base using active view** and the option of **Create using even, analyze using odd**. With these settings, the content set in the active view is divided into two parts. One part of the content set is used to create and train the knowledge base, and the other part of the content set is used to analyze the performance of the knowledge base when classifying data. In addition, the **Create using even, analyze using odd** option allows you to create and train a knowledge base using the content items listed in even-numbered positions and analyze its performance using the items in odd-numbered positions.

   **Note:** When you use **Create using all, analyze using all**, your test results will not be representative of knowledge base performance in a live environment, because the items you use to test the knowledge base are the same items as those that are used to create it. Because the knowledge base already “knows” these items, it can produce better-than-expected results.
3. Choose the default option to create a new knowledge base, deleting any existing knowledge base in this project if there is any, as shown in Figure 3-14 on page 73.

Figure 3-13 Options of creating and analyzing a knowledge base
4. Optionally, add match fields to content set items. In this example, we accept the default settings of adding and displaying five match fields, as shown in Figure 3-15 on page 74.

When the Add match fields option is selected, the content set view window displays additional match columns after the analysis. These columns show the top matches identified by the Classification Module as most applicable to each tested content item. In this example, we choose to display five match fields. For each match field, the category’s relevancy score appears in parentheses after the category name. Figure 3-17 on page 75 shows an example of the content set view with match fields. For example, the system determines that the 401k category is most applicable to content item 14; therefore, 401k appears in the Match1 column, followed by the score (99.29), which is the same category that you assign to this content item. Using match fields is a good way to verify your initial categorization, especially if the system-suggested top category differs from the category that you assigned initially.
Figure 3-15  Analysis options

5. The Status Information window allows you to view the create and analyze processes as they progress, as shown in Figure 3-16 on page 75.
6. After the knowledge base creation and analysis, Classification Workbench displays a content set view with the match fields added, as shown in Figure 3-17.

Figure 3-17  Content set view with match fields
3.2.3 Analyzing and learning with a knowledge base

Next, we summarize how to analyze and learn with a knowledge base:

1. Generate accuracy reports.
2. View and interpret the reports.
3. If required, improve the knowledge base performance by reanalyzing and recreating the knowledge base based on the existing knowledge base structure.

Generating accuracy reports

During the analysis of your knowledge base, the system generates test results that you can view using various reports, graphs, and tables. You can use this analysis data to evaluate and improve the accuracy of your knowledge base. Certain reports and graphs provide you with a view of the knowledge base’s overall accuracy, while other reports and graphs allow you to see the performance of each category within the knowledge base.

Take the following steps to generate the reports of your choice:

1. Click View Reports on the toolbar to open the View Reports window.
2. A good starting point to understand your knowledge base’s overall accuracy is to view the Knowledge Base Data Sheet, the Cumulative Success summary reports, and the Total Precision vs. Recall graph.

   In Figure 3-18 on page 77, we request these reports and graph.
Interpreting the reports

View and interpret the reports to check the accuracy of a knowledge base.

**Knowledge Base Data Sheet**

Figure 3-19 on page 78 is the report of the Knowledge Base Data Sheet for HR, showing the results of *total cumulative success*. The total cumulative success indicates how well the knowledge base suggests the correct categories. In this example, the report shows that Classification Module selects the correct category as the first choice 97.44% of the time and selects the correct category as the first two choices 100% of the time. It indicates that the HR knowledge base has good overall accuracy.
To understand how well each category performs, you can run the *Cumulative Success* summary report for each category, as shown in Figure 3-20 on page 79.
The **Total Precision vs. Recall** graph provides an immediate, visual sense of the knowledge base’s overall accuracy. The graph in Figure 3-21 on page 80 indicates that the HR knowledge base performs well - the curve is in the upper-right portion of the graph. When the curve is in the lower-left portion of the graph and the number of categories is small, the knowledge base’s overall performance is poor. If the results are poor, you might want to view the **Precision vs. Recall** graph for each category to see if individual categories are particularly problematic.
Reanalyzing and recreating knowledge base based on existing knowledge base structure

If required, you can improve the knowledge base accuracy by reanalyzing and recreating the knowledge base based on existing knowledge base data:

1. Click **Create, Analyze and Learn Wizard** on the toolbar to launch the wizard.

2. In the Options window, select **Create knowledge base using active view** and the option of **Create using even, analyze using odd**.
3. In the Specify options for the selected process window, select **Create knowledge base, using existing knowledge base structure**, as shown in Figure 3-22.

![Create Knowledge base, using existing knowledge base structure](image)

Figure 3-22  Create Knowledge base, using existing knowledge base structure

4. In the Match Fields window, accept the default match field settings, and check **Learn during analysis**, as shown in Figure 3-23 on page 82.

The **Learn during analysis** option enables the knowledge base to learn from categorized content items as the items are analyzed. All the content items that you designate for analysis are also applied to the knowledge base for learning. The knowledge base learns each text after it is analyzed. For each text learned by the knowledge base, the knowledge base adds the new knowledge information pertaining to the text into the knowledge base. Therefore, the knowledge base is constantly receiving feedback.

In this example, we have the HR content set containing 1000 items. We create the knowledge base using the even-numbered items and analyze the knowledge base using the odd-numbered items. In the process of learning during analysis, the feedback from each of the odd-numbered items during the analysis phase gets processed by the Classification Module server and is added to the knowledge base to enhance its accuracy. In the meantime,
each time that an item is analyzed, all previous feedback is utilized in the analysis.

![Create, Analyze and Learn Wizard](image)

**Figure 3-23  Learn during analysis**

5. Click **Finish** to continue. Wait for the system to analyze and process.

6. Click **View Reports** on the toolbar to open the View Reports window.

7. Check the following report and graph:
   - Cumulative Success summary report
   - Total Precision vs. Recall graph

8. View the reports and compare them against the previous versions of the reports.

### 3.3 Creating and analyzing a decision plan

Classification Module uses either knowledge bases or decision plans to perform the classification task. Unlike a knowledge base that you build by supplying a sample set of pre-categorized content items, you build a *decision plan* by configuring one or more rules, determining the rule order, and grouping rules
together if necessary. Decision plan rules can refer to one or more knowledge bases for a combination of rule-based and content-based classification.

This section introduces working with a decision plan in two typical scenarios:
1. Create a decision plan.
2. Analyze a decision plan.

3.3.1 Creating a decision plan

The following steps summarize creating a decision plan by using Classification Workbench:
1. Create a new decision plan project.
2. Add knowledge bases.
3. Create rules with a trigger and actions.

Creating a new decision plan project
Create a new project with a project type of decision plan:
1. In Classification Workbench, click Project → New to display the New Project window.
2. In the New Project window, as shown in Figure 3-24 on page 84, perform the following steps:
   a. Enter the project name in the Name field. In this example, we enter HR_DP.
   b. Select the project type in the Type field. In this example, we choose the project type of Decision Plan.
3. Select one of the following options to create the project:
   – Create a project by importing a content set.
   – Create a project by importing an existing decision plan file.
   – Create an empty project.

In this example, we choose **Create an empty project**, as shown in Figure 3-25 on page 85.
4. Figure 3-26 shows the project overview of the newly created empty decision plan project. The subsequent steps of adding knowledge bases and configuring rules are required to complete the decision plan creation.
Adding knowledge bases

Adding knowledge bases to your decision plan project enables the decision plan to make decisions based on knowledge base classification.

Perform the following steps to add knowledge bases to your decision plan project:

1. In the Decision Plan view, right-click Referenced Projects and select Add Project to add a knowledge base project, as shown in Figure 3-27.

**Note:** Knowledge bases to be used in a decision plan are created prior to building your decision plan project.

![Figure 3-27](image)

Figure 3-27  Add an existing knowledge base project as a referenced project

2. In the Add Project window, select the knowledge base project that will be used in the decision plan. In this example, we select the HR project from the project list and click Add, as shown in Figure 3-28.

![Figure 3-28](image)

Figure 3-28  Add a knowledge base project to your decision plan
Creating rules
The decision plan in this example is to identify and file the Health Care documents among all incoming documents in a designated IBM FileNet P8 folder for review. You can fulfill this task using the following two rules:

- Rule 1: Identify the most relevant top category of each incoming document.
- Rule 2: If the document belongs to the Health Care category, file it in a designated FileNet P8 folder for review.

Perform the following steps to create these two rules:

1. Define a rule group. In this example, we define a rule group by renaming the default New Group.

   Right-click New Group and select Rename to define its name, as shown in Figure 3-29.

![Figure 3-29 Defining a rule group](image)

2. A rule consists of properties, a trigger, and actions. Perform the following steps to create Rule 1:

   a. Right-click the newly defined rule group and select New Rule to create the first new rule, as shown in Figure 3-30 on page 88.
b. In the New Rule window, click the **Properties** tab. It has the following three settings:

- **Name**: The name of the rule.
- **When triggered**: Refers to the rule’s behavior when it is triggered. When you have more than one rule in a group, you can specify how you want to affect group processing or decision plan processing based on the trigger. There are three options available for the trigger: Continue, Stop group processing, and Stop all processing.
- **Enabled**: Specify this option to enable the trigger. If the check box is cleared, the rule is skipped.

Figure 3.31 on page 89 shows the settings of Rule 1’s properties.
c. In the New Rule window, select the **Trigger** tab. You can click **Add** to open a series of guided menus to define the trigger, or you can edit the trigger text directly. Click **Undo** to undo your previous action, and click **Clear** to start over. After building your trigger, click **Validate** to ensure that your trigger is written correctly.

In this example, Rule 1 is used to process all incoming documents. So, we set the trigger to `true`, as shown in Figure 3-32.

![Figure 3-31  Rule 1 Properties tab](image)

![Figure 3-32  Rule 1 Trigger tab](image)

d. In the New Rule window, select the **Actions** tab. You can access a set of predefined action scenarios by clicking **Add Actions**. Most action scenarios are designed for using IBM FileNet P8 integration. For example, you can use action scenario to file and unfile documents in IBM FileNet P8 folders, set document classes, and declare documents as records in IBM FileNet Records Manager. In addition to selecting and configuring action
scenarios, you can build your own scenarios by selecting one or more advanced actions.

In this example, because Rule 1 is used to identify the category to which each incoming HR document belongs, its action is defined in the following manner.

Analyze the document by Classification Module based on the HR knowledge base, and then, assign its most relevant top category name to a user defined content field, as shown in Figure 3-33. This user defined content field will be used in Rule 2 to take further actions.

![Add Actions](image)

**Figure 3-33  Rule 1 Add Actions**

3. You can use an approach similar to the approach that is described in step 2 to create Rule 2:
   a. Define the Rule 2 properties, as shown in Figure 3-34 on page 91.
b. Define the Rule 2 trigger, as shown in Figure 3-35.

![Figure 3-35  Rule 2 Trigger tab](image)

```
$DocCategory : ~Health Care~
```

Figure 3-35  Rule 2 Trigger tab

c. Define the Rule 2 action as shown in Figure 3-36 on page 92:

i. Select **File document into a specific in IBM FileNet P8.**

ii. In the Folder Name field, type the name of the IBM FileNet P8 folder where we want to file the documents. In our example, this name is `Content_OS/icm_integration/HealthCareReview`.

**Note:** To ensure a correct classification from Classification Module into the IBM FileNet P8 repository, the P8 folder name must be the complete path:

```
<object_store_name>/<folder_name>/<subfolder_name_if_any>
```
iii. In the Preview Decision Plan Actions pane, you can see the action syntax, which is `add_to_content_field 'P8:File' 'Content_OS/icm_integration/HealthCareReview'` in our example.

iv. Click **OK**.

Now, you have created a decision plan with two rules and one knowledge base, as shown in Figure 3-37 on page 93.
3.3.2 Analyzing a decision plan

You can perform the following steps to analyze a decision plan and view its analysis reports:

1. Import a testing content set into Classification Workbench.
2. Assign the content type value to content fields to be analyzed.
3. Analyze a decision plan.
4. View and interpret analysis reports.

**Importing a testing content set**

Import a testing content set into Classification Workbench to analyze your decision plan:

1. Click **Project → Import** to launch the Import wizard.
2. In the Import Content Set window, determine which type of content set you want to import. In this example, we select the **CSV** external format, as shown in Figure 3-38 on page 94.
3. In the Import Content Set window, click **Browse**, and locate the CSV file, as shown in Figure 3-39. And, click **Finish** to import the content set into Classification Workbench.
Assigning content type

In the Testing Content Set view, assign at least one content type value to a content field that will be used to analyze the content set. A different content set comes with different fields, which will be used for the content field. In our example, we use the Message_body field (of the Plain Text content type) to be used as the content field to be analyzed.

To assign the content type value, follow these steps:

1. In the Project Details pane on the right, double-click **Message_body** to display its Field Properties window, as shown in Figure 3-40.

![Figure 3-40 Editing property of the Message_body field](image)

2. Open the Content type pull-down menu and select **Plain Text**, as shown in Figure 3-41 on page 96.
Analyzing a decision plan
To analyze a decision plan with the imported content set:

1. In the Project Details pane on the right, double-click Initial View (1) on the Views tab to display the testing content set view.

2. Click Analysis → Analyze Decision Plan to launch the wizard, as shown in Figure 3-42.

3. In the Specify options window, select all the fields that will be displayed in the decision plan report. See Figure 3-43 on page 97.
Viewing and interpreting analysis results

View and interpret the following two analysis results:

- Decision plan summary report
- Analyzed content set

**Decision plan summary report**

A decision plan summary report contains the following statistical information:

- The *Rule statistics* section lists all rules in the decision plan and for each rule, whether it was triggered or skipped, the IDs of the content items on which each rule took action, and the percentage of these items relative to the testing set.

- The *Impact on content fields* section provides an overall breakdown of what happened to each content field when the decision plan was run against the specified content set.

- The *Modified content fields* section shows each content field in the content set that was modified by the decision plan.

- The *Deleted content fields* section lists content fields that were deleted from one or more content items by the decision plan.
The Content field values section shows all content fields in the content set with the values that they contain after running the decision plan.

For example, Figure 3-44 shows the rule statistics information after testing our decision plan against the content set. It indicates that the rule of Identify HealthCare Doc has been triggered for all 156 items in the content set, while the rule of File to P8 Review has been triggered 42 times, and has not been triggered for the rest of the content items.

![Figure 3-44 Decision plan summary report]

**Analyzed content set**

The Analyzed Content Set view shows the analyzed content set after running the content set through the decision plan. In addition to the content fields that are inherited from the testing content set, Classification Workbench can optionally display the user-defined content fields or system-defined content fields utilized in your decision plan. For example, Figure 3-45 on page 99 displays the DocCategory and P8: File fields that are added to the testing content set after running it through our decision plan of HealthCare Review. Per the rule definitions, the DocCategory field contains the category to which each item belongs, and the P8: File field includes the FileNet P8 folder path where the
Health Care document will be filed. Moreover, the *Analyzed Content Set* view provides additional analysis information for each content item, such as the list of suggested categories by Classification Module and their associated relevancy scores, its changed content field, and fired rules.

Figure 3-45  Analyzed content set

We have shown you how to create and analyze a decision plan with two rules and one knowledge base. Based on the requirements of your applications, you can create decision plans with rules only, or rules along with multiple knowledge bases, by using the techniques that are described in this section.

### 3.4 Building a knowledge base advanced topics

We recommend that you build a knowledge base from a categorized content set, as described in 3.2, “Creating, training, and analyzing a knowledge base” on page 60.

This section introduces you to advanced cases of building a knowledge base:

- Using an uncategorized or partially categorized content set
- Using keywords
- Enabling the knowledge base with feedback processing
3.4.1 Using an uncategorized or partially categorized content set

If you do not have a fully categorized content set, you can use the Taxonomy Proposer tool to help you with the categorization process before you build a knowledge base. The Taxonomy Proposer tool is installed with Classification Workbench and runs in a Windows environment.

To start the Taxonomy Proposer tool, double-click the TaxonomyProposer86U.exe file that is located in the ICM_home\Classification Workbench\Program Files directory.

The Taxonomy Proposer displays the Workflow Assistant, which guides you through typical scenarios, as shown in Figure 3-46.
It is beyond the scope of this book to go into detail about Taxonomy Proposer. Follow the Taxonomy Proposer Workflow Assistant to learn to use the tool.

### 3.4.2 Using keywords

When a content set is not available, you can use *keywords* to build the initial knowledge base and enhance its performance by learning from feedback. Keywords are words or phrases that you expect will appear in text that is classified by Classification Module. When the system identifies keywords in a text, the category associated with the keywords is more likely to be returned.

For example, assume that you expect to receive questions about your company’s health care coverage. You might want to choose keywords, such as *health*, *medical*, and *dental* for this category.

**Tip:** It is possible to repeat the same keyword up to three times for a category in order to increase the weight of this keyword.

You can enter keywords manually, or you can gather data offline and import it.

To prepare data prior to import, you can enter keywords in a spreadsheet (such as Microsoft Excel) and save the file in the comma-separated value (CSV) format. Figure 3-47 on page 102 shows the example of a keyword CSV file for a HR knowledge base. It contains the following columns:

- Proposed categories
- Keywords associated with each category

You can import the CSV file into Classification Workbench and create an initial knowledge base by using the *Create, Analyze and Learn Wizard* based on the data provided.
3.4.3 Enabling the knowledge base with feedback processing

Classification Module can adapt to the nature of your business by learning to identify separate categories from examples that you provide to it. A feedback is one example that you present to Classification Module to confirm or correct its current categorization. A positive categorization example confirms that a text belongs to a specific category. Submitting accurate feedback constantly improves Classification Module’s ability to classify texts. In this way, your knowledge base’s accuracy keeps pace with your business, rapidly adjusting to changes as they occur.

**Benefit:** Learning is a unique capability of Classification Module to continually process feedback and update a knowledge base.

As you provide feedback on the knowledge base accuracy, you can either adjust the knowledge base in real time or delay feedback processing until a later time.

Three options are available for the Classification Module system to process feedback on a knowledge base:

- **Process as accumulated:** Tracks and applies feedback to the knowledge base as it is accumulated in real time.

- **Defer processing:** Tracks the feedback events. Users can extract and analyze the feedback, and then, they can apply it to a knowledge base at a later time.

- **Do not process:** Does not track or apply feedback to a knowledge base.
If your application is designed to work with deferred feedback, these steps show a typical workflow of maintaining the accuracy of your knowledge base:

1. Defer the processing of feedback for the relevant knowledge base.
2. Extract the saved analysis data containing the postponed feedback.
3. Use the saved analysis data, and add learning to your knowledge base offline.

The accuracy of feedback to a knowledge base has a direct impact on the knowledge base's performance. We recommend using the defer processing option to have the knowledgeable users review and analyze feedback before it is applied to their knowledge bases.

**Note:** Depending on which suggested functions were used, the analysis data might not be saved to the server. In general, if you plan to analyze knowledge base performance or to defer feedback, you must turn on Save analysis data in the global properties of the server before the events are sent to the server. Otherwise, no data will be saved unless special suggest and feedback functions are used.

Follow these steps to enable the knowledge base with the defer processing feedback option:

1. Launch Management Console by clicking **Start → Programs → IBM Classification Module 8.6 → Management Console**.

2. Enter the server listener URL, as shown in Figure 3-48, which is defined during the Classification Module product installation. In this example, it is `http://localhost:18087`.

   ![Connect to the Management Console](image)

   *Figure 3-48  Management Console logon*

3. In the Management Console window, select **Knowledge bases** on the left pane.

4. Right-click your knowledge base on the right pane, and select its **Properties** menu entry, as shown in Figure 3-49 on page 104.
5. In the Properties window, set the Feedback field to **Defer processing**, as shown in Figure 3-50 on page 105.
6. Restart the knowledge base so that the new setting takes effect, as shown in Figure 3-51 on page 106.

The system now starts to track the analysis data for this knowledge base. The analysis data that can be useful for knowledge base analysis includes *feedback* (the confirmation or correction of how a content item was classified) and *matches* (the categories suggested for a particular content item and the associated relevance scores).
Extracting the saved analysis data

A ready to use, predefined conversion tool is available for you to extract the analysis data. On Windows, it is the `bnsExtractTexts86.exe` command. On AIX®, Linux®, or Solaris™, it is the `./bnsRun extract` command.

Perform the following steps to extract the saved analysis data:

1. Edit a configuration file that is required by the conversion tool. We provide a sample configuration file for your reference:

   ```
   C:\IBM\ClassificationModule\Bin\ExtractTextsExampleConfigFile.txt
   ```

   In this example, we intend to extract the analysis data of the postponed feedback events for the HR knowledge base during a specified time period. The output XML files are created in a user-defined folder. This folder needs to be empty before the command is run.

   Our configuration file looks similar to the file that is shown in Example 3-1.

   **Example 3-1   Sample configuration file for feedback data extraction**

   ```
   # The folder where the output XML files are to be created
   XmlDir = C:\Data\Xml

   # The time period for the event data that you want to import
   StartTime = 2008/08/18 00:03:00.000   # Events starting at this time
   EndTime =   2008/11/20 23:59:59.000   # Events ending at this time

   # The type of data that you want to extract
   # Options are TextOnly, KB, or DP.
   ExtractType = KB
   ```
# The name of the knowledge base project (.kb) that you want to
# train or analyze.
KBName = HR

# Extracts all feedback or match events during the specified time
# period, or only the last feedback or match event.
# Options are All or Recent.
Scope = All

# The type of feedback event data that you want to extract.
# Options are 0 to not extract feedback, or any combination
# of Feedback, FeedbackPostpone.
FeedbackEvents = FeedbackPostpone

# The type of suggest event data that you want to extract.
# Options are 0 to not extract matches, or any combination
# of Suggest, SuggestFromDecide, and SuggestDocument.
# 1. Feedback is used together with suggest to analyze how well
# the KB performed in the past.
# 2. Just Postponed Feedback is used to add learning (offline
# feedback) after auditing.
SuggestEvents = 0

# The action to take if both feedback and suggest data is extracted.
# Yes: If both FeedbackEvents and SuggestEvents are configured, a
# text will be skipped if just one event type is found for the text.
# No: Extracts data for texts that have one event type.
Correspond = No

2. Run the bnsExtractTexts86.exe command to extract the stored analysis data
in XML output files. For example, on Windows:

cd C:\IBM\ClassificationModule\Bin
bnsExtractTexts86.exe extractHRConfig.txt
3.4.4 Working with offline (deferred) feedback

You can add learning to your knowledge base with the offline feedback and then run a new analysis with new testing content.

Add learning to your knowledge base with the offline feedback:

1. In Classification Workbench, create a new project of the knowledge base type by importing the knowledge base from the live knowledge base running on the Classification Module server.

2. Import the XML files containing feedback into the same project. Make sure to assign the feedback category for the imported analysis data, because the classification fields and the text field need a content type. At this point, the user can review and audit the feedback.

3. Have the knowledge base learn with all the feedback in active view by using the Create, Analyze and Learn Wizard. In particular, choose Learn using active view, as shown in Figure 3-52.

![Create, Analyze and Learn Wizard](image)

Figure 3-52  Learn using active view

After adding learning to your knowledge base with offline feedback, you can run a new analysis with a testing content set:

1. Import a testing content set of your choice. Make sure that the data contains a text field with a content type and a classification field for analysis.

2. Analyze your newly learned knowledge base with the testing content set in active view by using the Create, Analyze and Learn Wizard. In particular, choose the option Analyze knowledge base using active view, as shown in Figure 3-53 on page 109.
3. View the reports to analyze the knowledge base accuracy.
4. If you are satisfied with the accuracy of your newly learned knowledge base, you can deploy it to your production system.

### 3.4.5 Handling overlapping categories

Even though approaches to tuning knowledge bases vary from accuracy problem types, they all use the same procedure:

1. Analyzing the knowledge base with a representative testing content set
2. Generating and viewing the knowledge base reports
3. Taking proper actions based on the reports

This section introduces the process to tune the performance of a knowledge base with the overlapping categories problem. We created the HR knowledge base in Classification Workbench. Analyze it in the following manner:

1. Click **View Reports** on the toolbar, and select the following reports, as shown in Figure 3-54 on page 110:
   - Summary reports:
     - Cumulative Success
     - Knowledge Base Data Sheet
   - Graphs:
     - Cumulative Success
     - Total Precision vs. Recall

![Figure 3-53 Analyze knowledge base using active view](image-url)
2. Review the Knowledge Base Data Sheet report:
   a. The Total cumulative success table, as shown in Figure 3-55 on page 111, indicates the low cumulative success.
Figure 3-55 Knowledge Base Data Sheet report

b. The *Pairs of Categories with overlapping intents* section lists the pairs of categories with overlapping intents, as shown in Figure 3-56 on page 112.
c. As an example, we further examine the following pair of categories, as shown in Figure 3-57, with additional reports.

\[
\begin{align*}
\text{Gift Certificates (9) } & \quad \text{Gift Wrap (6)} \\
\end{align*}
\]

Figure 3-57  Example of overlapping categories: Gift Certificates and Gift Wrap

d. Click View Reports on the toolbar. On the Category Graph and Tables tab, generate reports for the Gift Certificates category, as shown in Figure 3-58 on page 113.

Note: It is important to choose Classification Workbench View so that you are able to open and view each content item directly by using the Content Item Scoring report.
3. We reach the following conclusion: The *Stealing/Stolen Table* report in Figure 3-59 shows that the Gift Wrap and Gift Certificates categories are highly overlapped.

![Category graphs and tables](image)

*Figure 3-58  Category graphs and tables*

<table>
<thead>
<tr>
<th>Category</th>
<th># of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>APO_FPO</td>
<td>10</td>
</tr>
<tr>
<td>Account Info</td>
<td>10</td>
</tr>
<tr>
<td>Address Change</td>
<td>15</td>
</tr>
<tr>
<td>Batteries</td>
<td>11</td>
</tr>
<tr>
<td>Bridal Registry</td>
<td>11</td>
</tr>
<tr>
<td>Catalogs</td>
<td>15</td>
</tr>
<tr>
<td>Certificate Expired</td>
<td>2</td>
</tr>
<tr>
<td>Company</td>
<td>7</td>
</tr>
<tr>
<td>Credit Declined</td>
<td>4</td>
</tr>
<tr>
<td>CreditCards</td>
<td>8</td>
</tr>
<tr>
<td>Dresses</td>
<td>8</td>
</tr>
<tr>
<td>Gift Certificates</td>
<td>9</td>
</tr>
<tr>
<td>Gift Orders</td>
<td>3</td>
</tr>
<tr>
<td>Gift Wrap</td>
<td>6</td>
</tr>
<tr>
<td>Hours</td>
<td>5</td>
</tr>
<tr>
<td>International Order</td>
<td>4</td>
</tr>
<tr>
<td>International Shipping</td>
<td>11</td>
</tr>
</tbody>
</table>

Open selected graphs as: Classification Workbench View

![Stealing/Stolen Table](image)

*Figure 3-59  Stealing/stolen table*
4. The *Content Item Scoring graph* in Figure 3-60 shows the scoring for each content item for the selected Gift Certificates category.

   The light colored (blue) points represent content items that belong to the selected category, while the darker colored (maroon) points represent content items that do not belong to the selected category.

   The content items with low scores are at the bottom, and those content items that received high scores are in the upper part of the graph.

   You can click a point and open the document to read it and to decide whether its classification is correct.

5. A common solution to overlapped categories is to combine those categories. In this example, we decide to combine the Gift Wrap and Gift Certificates categories to one Gift Certificates category.

   Combine the categories using these steps:

   a. In the Project Details pane, on the Category tab, right-click **Gift Wrap**, and select **Show Items**, as shown in Figure 3-61 on page 115.
Figure 3-61  Show Items of the Gift Wrap category

b. Select all content items belonging to the Gift Wrap category, as shown in Figure 3-62.

Figure 3-62  Select all items belonging to the Gift Wrap category
c. Right-click the selected items, and select Categorize Highlighted As, as shown in Figure 3-63.

![Image of Categorize highlighted items]

Figure 3-63  Categorize highlighted items

d. Select the Gift Certificates category on the left pane, click «», and click Apply, as shown in Figure 3-64.

![Image of Select Categories]

Figure 3-64  Select the Gift Certificates category to apply to highlighted items
e. In the Project Details pane, delete the **Match n** fields on the **Fields** tab to remove the previous classification result, as shown in Figure 3-65.

![Project Details](image)

*Figure 3-65  Deleting the Match n fields*

6. Use the Create, Analyze and Learn Wizard to create a new knowledge base after removing the **Gift Wrap** category:

   a. Select the right content set view in Classification Workbench.

   b. Click the **Create, Analyze and Learn Wizard** on the toolbar to launch the wizard.

   c. Select **Create and analyze knowledge base using active view**, and **Create using even, analyze using odd**.

   d. Select **Create new knowledge base, deleting any existing knowledge base in this project**.

   e. Accept the default analysis options, and click **Finish**.

   f. Wait until the Create and Analyze operation finishes, and click **Close**.

7. Click **View Reports** on the toolbar to generate the *Knowledge Base Data Sheet* report. Figure 3-66 on page 118 shows that the cumulative success has improved.
Figure 3-66  Knowledge Base Data Sheet with improved cumulative success
Chapter 3, “Working with knowledge bases and decision plans” on page 59 provides the mechanics of working with knowledge bases and decision plans using IBM Classification Module (Classification Module). It includes the basic techniques for creating, training, and analyzing knowledge bases. This chapter reexamines the major implementation steps and focuses on understanding these steps and the critical factors behind them, in order to help you create the best knowledge base for your system.

We discuss the following topics:

- Use cases
- Working with Classification Module
- Defining categories
- Preparing the content set
- Building and training the knowledge base
- Analyzing knowledge base performance: Identifying the problems
- Fine-tuning your knowledge base: Fixing the problems
- Maintaining your knowledge base over time: Using feedback
- Analyzing the knowledge base in production
- Application design considerations
- Summary
4.1 Use cases

In each step of the knowledge base development, it is useful to consider the nature of the problem that Classification Module is trying to solve. These use cases can fall into a number of categories:

- Routing e-mail and automatic reply systems
- ECM e-mail archiving, document storage, content management, and records management
- Self-help applications
- Customer Relationship Management applications
- Search applications
- Document filtering

**Routing e-mail and automatic reply systems**
Classification Module can help route e-mails to human agents and supply a mechanism for automated replies.

**ECM e-mail archiving, document storage, content management, and records management**
Classification Module can assist enterprise content management (ECM) systems in tasks, such as e-mail archiving, records declaration, document management, and eDiscovery readiness, by automating the decision making process. E-mail content (from a mail server) can be extracted by IBM Content Collector and sent for archiving. The ECM application, empowered by Classification Module, performs tasks, such as choosing among the various storage options (using a decision plan), moving or copying a file to a new location, removing documents, and archiving. Typically, a category is mapped to a storage location, such as a directory or archive. ECM can also add additional attributes to the document for the purpose of easy document retrieval.

**Self-help applications**
Self-help applications, empowered by Classification Module, automate e-mail handling in customer service business processes. Self-help applications can automatically classify customer correspondence, identify the customer’s problem quickly, and often deliver an answer back immediately based on the classification, which removes the requirement for human intervention.
Customer Relationship Management applications

Customer Relationship Management (CRM) systems, empowered by Classification Module, are similar to self-help applications, but they usually involve intervention from a human agent. In CRM systems, factors, such as high confidence of the suggested category, determine whether an automated answer is sent or whether suggested answers are presented to the agent.

Search applications

Typically, categorization is added to search and retrieval systems as an overall organizational mechanism. The category tree can also be used as a navigational device to help users find possible areas of interest. Within the traditional search, categorization adds to the document description and can be used within traditional search techniques (queries). It can add to the knowledge about a document and help in cases where the search query returns a large result set containing too many irrelevant documents. Typically, this large result set with too many irrelevant documents happens when identified keywords do not reflect the user’s intent.

Document filtering

You can use Classification Module for document filtering in the following areas:

- Custom spam
  The spam filtering system categorizes all incoming e-mails as either spam or non-spam, according to the internal conventions of the company (rather than on universal spam concepts). The user decides how rigid a filter to use, depending on the margin of error that the user will accept.

- Personal/Business
  Often, workplaces want to archive only business correspondence and documents on a file system. The knowledge base needs to learn to differentiate between personal and business-related documents and e-mails.

- Pinpointing relevant categories and documents
  Another filtering use case involves finding relevant documents among a large collection of mostly irrelevant documents.

4.2 Working with Classification Module

Before we describe building, training, and fine-tuning a knowledge base in detail, we examine a typical life cycle of a knowledge base, Classification Module’s Natural Language Processing engine processing the content set, the most
important aspects of a successful Classification Module implementation, and the major implementation steps.

### 4.2.1 Typical life cycle of a knowledge base

When using Classification Module, the knowledge base that is initially created undergoes changes over time in order to adapt to new content and changing business needs:

1. The knowledge base is first created based on a set of categorized texts. After it is satisfactorily analyzed, it is installed on the Classification Module server.

2. An application sends a text to the Classification Module server.

   **Note:** Separate types of applications send separate types of texts to Classification Module for classification. For example, e-mail applications send messages, and self-help applications send queries and questions.

3. The Classification Module server sends the text to the relevant knowledge base or decision plan. The decision plan might send the text to a knowledge base.

4. Classification Module returns the most relevant categories together with their associated scores (the match results).

5. The application uses the match results to take action (for example, to send an automatic response to an e-mail message, to suggest responses to an agent, or to classify the corresponding document into a records management system).

6. The application might receive feedback from the user about the correct categorization of the text and pass it to the Classification Module. Classification Module can use this information to adapt the knowledge base by using its online learning capabilities. For example, this learning occurs when an agent selects a response.

7. Optionally, you can analyze how well the knowledge base is categorizing your text (using Classification Workbench). Based on these analysis results, you can decide to modify and retrain your knowledge base using Classification Workbench. After making these changes, you export the new knowledge base to the Classification Module server.

8. Depending on changing business needs, you can decide to gather or create separate sets of training data to create a new version of the knowledge base.

These steps are iterative, allowing the knowledge base to be both dynamic and adaptive.
4.2.2 Natural Language Processing

The Natural Language Processing (NLP) engine enables Classification Module to understand the intent of the text, rather than just treating the text as a collection of randomly ordered strings. It adds significant accuracy to Classification Module’s analysis capabilities.

The NLP engine is responsible for the extraction of concepts from incoming texts. Every text consists of a set of fields known as Name Value Pairs (NVPs). In e-mail applications, each field contains a separate component of a message, such as To, From, Body, or something about the message, such as its source.

NLP consists of a number of operations, including:

- Language identification and encoding conversions: The process of identifying the language of a specific text and the ability of converting ASCII (code page dependent) texts into Unicode
- Tokenization: The process of identifying words, phrases, e-mail addresses, and other elements, within long texts
- Text cleanup: The process of eliminating parts of text that are not relevant to classification
- Spelling and other error corrections
- Morphological analysis: Used to derive the base forms of words
- Other types of proprietary linguistic analysis
- Keywords and concept extraction

Classification Module analyzes input text against categories in the knowledge base and returns relevancy scores. These scores represent the Classification Module’s confidence level that the text falls within that category.

Classification Module returns the category names and scores to the host application (Classification Module-based application). The host application can then use this information to automate actions, such as routing, filtering, sorting, and responding.

Classification Module stores all of the data that is required for online learning and returning scores in the knowledge base. In addition, the knowledge base might have rules that indicate how to route texts to specific categories, based on content, or other properties of the text, such as whether the message is received over a secured channel or not.
4.2.3 Classification Module implementation success

To ensure a successful Classification Module implementation, building, training, and fine-tuning the knowledge base that is used in the implementation is extremely important.

The knowledge base is a map of the inputs to the system and the outputs that you want to receive from it. It reflects your specific environment or domain, for example:

- If we want to classify animals, we can build a list of animals and teach the system how to recognize the animals in the list.
- If we want to classify customer e-mails, we can build a list of the topics that customers discuss through e-mail and teach the system how to recognize each of these topics.

The success of Classification Module implementation depends on these factors:

- Identifying and defining the right categories
- Using and preparing the right content set
- Building and training the knowledge base with these categories and this content set
- Properly fine-tuning and maintaining the knowledge base

4.2.4 Classification Module implementation summary

The process of building, training, and fine-tuning a knowledge base consists of the following major steps:

1. Defining categories.
   Understand the customer’s problems and issues. Identify the correct and distinct categories that are used to build a knowledge base.

2. Preparing content and creating a knowledge base.
   Gather and prepare a suitable content set (corpus) for building the knowledge base. Split the content set into a training set and an analysis set. Create and train a new knowledge base, using the training segment of the content set.

   Test the knowledge base and generate the analysis reports, using a testing segment of the content set. Assess the knowledge base performance by viewing the reports.
4. Fine-tuning your knowledge base.
   Improve the knowledge base performance by fine-tuning the content and retraining the knowledge base.

5. Deploying knowledge bases.
   Optionally, create a decision plan to modify how the requests are sent to one or more knowledge bases.

6. Maintaining your knowledge base.
   Ensure that any changes in the type of texts sent to the server are reflected in the knowledge base, such as removing obsolete categories, adding new categories, or retraining existing categories.

In the following sections, we explore each step (except the actual deployment) in detail in order to help you implement a successful Classification Module solution.

4.3 Defining categories

Before building your knowledge base, you must perform these required tasks:

1. Get to know the business problem that you need to solve, and familiarize yourself with the data.
2. Choose and define the relevant categories.
3. Get pertinent sample content for each category.
4. Organize your data into categories for later processing.

During this manual preparation, analyze the data and determine the best representation of your data.

4.3.1 Become familiar with the problem and your data

Investigate the nature of the problems that the company wants to solve. Do not be limited to only the company's view of the solution. This concept is an extremely important concept. The company might not be familiar with how Classification Module works and might have incorrect assumptions and expectations. Concentrate on the existing problems and explore the possible solutions.

Provide common examples of the problems that companies want to solve:

- E-mail: How does the company want the e-mails to be classified? Does the company want to classify all e-mails under specific categories? What does
the company want to do with unclassified e-mails (classify later manually, delete, or other actions).

- ECM: Does the company want to simply categorize the documents, rearrange them within the file system, or ingest them into a full records management application?
- Content assessment: Investigate the data that you have gathered and stored for years before you start the eDiscovery readiness preparation, compliancy, or document management organization.

### 4.3.2 Choosing categories

It is not always clear to the company what its categorization needs are. The company might have an informal system of categorization that is well understood within the company. Or, you might need to analyze the content to discover the informal system of categorization. Classification Module can assist with both of these situations.

Choosing categories correctly is probably the most important step in building an initial knowledge base. The following tips help you to choose the appropriate categories:

- Choose categories that represent intents.
  
  This step involves creating a list of category names. Each category must have a clear intent. For example, if the intent is to categorize levels of customer satisfaction, the text “I am unhappy with your customer service” is given the “Customer Service Complaints” category. In each text, the reason for the dissatisfaction can differ, although the intent of the text is the same in context of the knowledge base design.

- Identify and define clearly the intent of each category.
  
  It is necessary to create a definition of the intent for each category that you choose. We highly recommend that you define the intent for each category for situations where multiple people create and maintain your documents (corpora) and knowledge base. The purpose of the definition is create a single view of the data for the defined task. It can also be helpful to note specific common misunderstandings or ambiguities (for instance, for the “Bicycles” category, “Do not use for plastic tricycles, use Toys instead.”).

- All texts with similar intent need to belong to the same category.
  
  After the intent of each category is clearly defined, the team will be able to categorize the content consistently, in spite of possible alternate interpretations of the data. Even if the intent is not fully agreed with, the team must agreed to it.
Categorization must not be based on external factors that are not reflected in the texts themselves.

For example, several issues might easily be assigned to a “Requires Research” category, although they might actually have little in common. Therefore, “Requires Research” is typically a bad choice for a category.

If there is a valid need to have an alternate view of the data, more than one knowledge base can be created for the same data set, representing different facets (for instance, “Country of origin” and “Subject”). However, if any category is based on external information, this factor needs to be dealt with externally to the knowledge base, by either a custom application or by writing rules in a decision plan.

Categories must be distinct from each other.

If, for example, you create a category to categorize texts related to spectator sports, your category overlaps with a category for the topic of football. If you use Classification Module for an e-mail classification application, overlapping categories will naturally confuse the agent that provides the feedback to Classification Module. As a result, Classification Module will also find it difficult to differentiate between these overlapping categories.

Texts with multiple intents, however, are handled extremely well by Classification Module. A single message can belong to more than one category.

Categories must reflect the business practices that you want to address by using Classification Module-based applications.

For practical reasons, sometimes the suggestion is made to train categories by the answer that will be sent, and not by the intent. However, training categories using this method might not always provide optimal performance, because the answer texts differ in nature from the question texts. If this method is chosen, special care must be taken to analyze and tune the knowledge base, after it has been created and trained. The knowledge base needs to learn to associate the category with the actual texts that it receives, either by feedback or retraining.

**Level of granularity: Number of categories**

According to the specific problem that you are trying to solve, the optimal number of categories will vary. These categories can be extremely general or extremely specific, depending on the variety of tasks that the application and agents need to perform. For instance, it is best to avoid multiple categories mapping to a single task, such as moving files to a single file location. These situations indicate that the categories are too specific and might be difficult to maintain, whereas a single, general category serves the purpose more efficiently.
Importance of category names
We recommend that you use category names that are descriptive of the category’s intents. Using meaningful category names helps the users of the Classification Module-based applications to identify categories correctly (for example, agents who choose categories to provide feedback to Classification Module).

4.3.3 Getting good sample content for each category

After you have a list of agreed-upon categories, the next task is to work with the potential business users to get sample documents for each category. Several techniques exist for organizing and importing this content into Classification Workbench. An easy way to build a knowledge base using sample content is to create a folder per category on a file system. Alternatively, you can create structured files, such as comma separated value (CSV) files, or XML, that contain all of the field data (structured and nonstructured) for creating the content set.

Make sure that your sample content contains relevant, unbiased, reliable data if there is no agreed-upon list of categories. These categories will need to be discovered from the customer content.

4.3.4 Discover the categories from your data

In cases where you do not have a clear picture of the data structure, Classification Module Taxonomy Proposer is one possible tool that can suggest possible categories by analyzing the customer’s content. Taxonomy Proposer provides you with groupings that can suggest categorization. This process requires that you review items in each grouping and try to determine what makes them similar. For example, if many items in a grouping relate to the topic “Windows® 2000 installation issues,” creating a category for this topic might be warranted.

4.3.5 Separate content and categories for separate uses

At the beginning of this chapter, 4.1, “Use cases” on page 120, we listed a number of use cases. Here, we mention a number of aspects of content selection and knowledge base building specific to each type.
Routing e-mail and automatic reply systems
In order to determine the categories that are needed, examine sample e-mails and the agents' replies. Typically, the agents' answers can be reduced to a number of standard answers, each of these “canned” answers corresponds to a single category.

E-mail archiving, document storage, content management, and records management
Classification Module can assist ECM systems in choosing among the various storage options: moving or copying a file to a new location, removing documents, and archiving. The knowledge base structure needs to strongly resemble the storage structure and include special categories, such as “Delete” and “Send to archiving.” Classification Module provides the Extractor tool that assists the user in extracting data from the ECM repository in order to build the taxonomy.

Self-help applications
Typically, users try to find texts relevant to their problems. These answer texts need to be identified (which might involve dividing a large document into topics) and mapped to categories. For each category, a number of sample user requests must be identified and incorporated into the training or learning data.

CRM
Similar to Self-Help applications, the agents' answers need to be mapped to possible questions. CRM systems can differ in that routing to agents might require another level of organization within the knowledge base. For example, each agent can deal with a separate area of expertise, and this area becomes a category. The subcategories correspond to each possible answer that is sent by that agent. In this scenario, you can build the knowledge base using example e-mails and documents that have been manually categorized by agents.

Search applications and text analytics
Within search applications, it is important to receive an organizational overview of the material from subject experts or experienced users. A representative sample of the entire collection needs to be categorized by these same users and experts.

Classification Module also provides a platform for text analytics that gives the users tools to extract information from the content that can be further used as additional facets. This functionality is mainly available through the decision plan.
4.3.6 A detailed example of the Self-Help/CRM scenario

Your company's help desk receives e-mails and phone calls about products and services. Agents respond by manually pasting canned text responses into e-mail answers and reading canned text responses over the phone to handle telephone inquiries. Your goal is to build a system that automates the text responses.

You begin by analyzing a recent sampling of customer inquiries to determine which categories are appropriate. Categories must match the intents of your customer inquiries. In certain applications, each category can be matched to one or more canned text responses, to satisfy a high percentage of the inquiries.

After considering which categories are appropriate, identify the unique responses that are contained in your outgoing e-mail (e-mail replies, including questions and answers). You filter these e-mails by finding textual patterns that will generally correspond to the canned texts that your agents use in their responses. This process will give you a way to achieve an initial categorization.

Next, search through the texts to see if there are other hidden intents and meanings in the texts. For example, if you initially determine that there is a category about poor service, you might discover that this category actually contains complaints about long wait times and discourteous service. Because the responses differ for each type of complaint, it is worth splitting the category into two separate, more specific categories.

Now, check to see if there are any redundant categories, that is, multiple categories that describe the same intent. For example, you might have a category for customers requesting to learn the location of your business and a category for customers requesting driving directions to your business. Because the intents of the requests differ only slightly, you might want to create one category (with an associated canned response text) for these related intents. For example, many of those customers requesting the location are doing so in order to figure out the driving directions.

After categories have been assigned to all texts and you are reasonably certain that the categories are correct, it is time to build and train the knowledge base.

After you build and train a knowledge base, you analyze your knowledge base to verify that the accuracy is reasonable in the most common categories and that texts in the testing corpus have been correctly categorized. You discover several texts that were incorrectly categorized, return to the corpus, categorize these texts again, and retrain the knowledge base.

When you are reasonably satisfied with the knowledge base's performance, you begin to use it in a production environment. The knowledge base's performance
is improved automatically over time through Classification Module’s feedback and learning mechanism.

4.4 Preparing the content set

The simplest method for building and training a knowledge base is to prepare the corpus and to use the categories of the corpus to build and train the knowledge base. In many cases, you have access to a categorized or partially categorized corpus. If this is not the case, you need to gather texts and build a corpus.

Using a proper corpus is crucial in building the correct knowledge base and training it to function accurately and efficiently.

4.4.1 Data sources for knowledge base building

A corpus that is used for building a knowledge base is a set of texts supplied by your organization or customers that have been (ideally) pre-classified into appropriate categories.

When building a corpus, consider the following factors:

- Which texts to choose
- Number of required texts
- Length of texts

Which texts to choose

The texts that you choose for your corpus need to be representative of the types of texts that will be categorized using Classification Module. For example, for an e-mail response application, use texts that are typical e-mail messages that contain questions that are similar to the type of questions you expect to receive.

These examples of content are good sources of training data:

- Data imported from a file system (into Classification Workbench)
- Data extracted from an ECM system (such as IBM FileNet P8)
- E-mail content, or other examples of call center agents’ responses to customers, exported into CSV or XML format
- Manually developed training samples
**Number of required texts**
The more texts that you have in the training corpus, the more accurate the categorization results will be. Depending on the type of data your application handles, the categorization results will improve per category as the number of texts increases. It is important for the number of texts to be distributed evenly among the categories. Also, for shorter texts, a larger number might be needed to provide ample training material.

**Length of texts**
We recommend that the texts be similar in length to the actual texts that the Classification Module is likely to encounter in your system.

4.4.2 **The nature of the content**

There are various types of text. In the simplest case, a single medium size text (25 - 300 words approximately) is sent to the Classification Module, and a set of matched categories is returned.

However, Classification Module can handle a wide range of texts.

Many Classification Module-based applications (for example, e-mail response and self-help applications) present two distinct types of texts: texts containing inquiries or questions and responses to these inquiries.

Within a text retrieval system, the nature of the user's query can differ significantly from the more formal language appearing in the stored texts. The knowledge base must be trained to recognize both types of text, because the knowledge base is used to add the category to each document at indexing time. And at run time, the knowledge base must assign the correct category to the user's query. If there is a real gap between the types of language used, you can maintain two separate knowledge bases.

When Classification Module is used to classify structured content (such as content from a database), each field can differ significantly in the type of knowledge that it contains. In this case, you can use a decision plan to send each field to a separate knowledge base, depending on the business needs.

For applications that work with both question and answer texts, it is important to understand the differences between them. Question texts are often short and usually express a writer's lack of knowledge about a particular subject. Conversely, answer texts are often lengthy, contain more details, and are generally written by experts.
4.4.3 Cleaning corpus texts

Classification Module can differentiate between *noise content* (noise content or noisy data is content that does not provide any useful meaning to anyone’s system) and informative content. However, cleanup might still improve the categorization results. Noise is most harmful when it occupies a large percentage of the text (for example, when an e-mail contains Multipurpose Internet Mail Extensions (MIME) headers and the sender's signature, or a template of the form, while the actual content consists of one short sentence or even a single word).

Other candidates for cleanup are any identifiable templates that are not exactly identical in all texts. It is less crucial to clean a repeated pattern that is identical in all texts. However, a pattern that characterizes only part of the texts needs to be cleaned up, because its existence can be informative for Classification Module (and can change over time).

While it might be tempting to clean certain repeating pattern strings, such as “Get your own free e-mail at mail.com,” which can appear in many of your corpus texts, we generally advise against this cleaning. Classification Module is well-suited for analysis of noisy data. To the extent that the noise is just as likely to occur in the actual texts to be processed by Classification Module, it is better to leave the noise, because it can actually help in the categorization process.

Classification Workbench gives you the tools to identify and clean up the noisy data. When necessary, you can use the corpus editing tools, for example, Find Patterns and Find/Replace words.

**If you decide to clean texts, remember:** If you clean up a fixed pattern or other noise from the training corpus, this noise also needs to be filtered out of the incoming real-world texts (that is, through a custom preprocessing script) before the texts are categorized by Classification Module.

**Example one**

Your organization plans to use a Classification Module-based application to categorize news articles. You build a corpus by collecting and importing a number of sample news articles from the Internet. Along with the main body text, the Web pages include extra, seemingly unnecessary text (for example, copyright information) that is unrelated to the news articles’ content.

If this extra text will not be included in actual news articles that you plan to categorize using Classification Module, you need to remove it. If, however, you expect to categorize news articles with similar unnecessary text, leave it in the sample news articles to maximize knowledge base training and performance.
Example two
You create a corpus by gathering and importing archived e-mail messages. Other people in your company forward messages to you for inclusion in the corpus. By forwarding these messages, your own company’s signature is included at the bottom of each message.

Because this footer text will not appear in the messages that you expect to receive and categorize using Classification Module, you search for and remove all occurrences of your company’s footer. You might want to leave other footer text (besides your company’s footer) in the e-mails, if you expect to receive messages with these types of footers.

Handling signatures at the bottom of e-mails
E-mail signatures are difficult to clean, because they do not follow any fixed pattern and people are often creative when it comes to their signatures. However, if the typical relevant content of corpus texts is not extremely short, Classification Module will successfully filter out most of the noise that the signature creates, automatically.

Defining fields
By defining fields, you determine how you want to process the texts in your corpus. You must define fields and their content types based on the text elements that you expect Classification Module to analyze and act upon.

Important: The name of an field must not begin with an underscore (_)

character.

4.4.4 Various types of texts: Their use and content types

Because of the wide range of text sources handled by Classification Module, there are many options to help you work with each situation.

Table 4-1 on page 135 summarizes the types of texts that Classification Module applications process and learn from. For example, you can use the body text of an e-mail message for online learning, matching, and training. However, a canned answer can be used for initialization only and not for matching or learning.

This table also indicates whether special text processing (NLP) or DocFilterManager processing is required for each type of text. For example, no special processing exists for self-help inquiries. In e-mail messages, the NLP automatically removes the original message. In all cases, texts are processed according to their language.
When Classification Module receives a text directly from a document, this text is first processed by the DocFilterManager. This process extracts metadata and turns binary content into text. The extracted metadata is transformed into fields, which can be handled by a decision plan and either serve as the basis for decisions or be selectively sent to the knowledge base for matching.

Table 4-1 Application text type for Classification Module process and learn

<table>
<thead>
<tr>
<th>Type of application text</th>
<th>Receives special NLP processing?</th>
<th>Suggested content type</th>
<th>Used in Classification Module function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-help inquiry</td>
<td>No</td>
<td>PlainText</td>
<td>Feedback Matching Training</td>
</tr>
<tr>
<td>Forwarded e-mail message body</td>
<td>Removal of original message (for forwarded e-mails and replies)</td>
<td>Body</td>
<td>Feedback Matching Training</td>
</tr>
<tr>
<td>E-mail message subject</td>
<td>Removal of “RE:” and “FW:”</td>
<td>Subject</td>
<td></td>
</tr>
<tr>
<td>E-mail message sender</td>
<td>E-mail address parsing</td>
<td>Sender</td>
<td></td>
</tr>
<tr>
<td>Canned answer standard response</td>
<td>No</td>
<td>Not applicable</td>
<td>Knowledge base Initialization</td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
<td></td>
<td>Initialization of a new category</td>
</tr>
<tr>
<td>Category name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer’s title</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fields from structured content (database)</td>
<td>No</td>
<td>PlainText</td>
<td>Decision plan can be used to differentiate the use of fields with differing semantic content</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Selected fields can be used for feedback, matching, and training of multiple knowledge bases</td>
</tr>
<tr>
<td>Document (in document retrieval or classification applications)</td>
<td>Metadata is extracted by DocFilterManager</td>
<td>Document</td>
<td>Feedback Matching Training</td>
</tr>
</tbody>
</table>
Categorizing texts

As mentioned earlier, the best way to build an effective knowledge base is by manually categorizing the training corpus. A well-categorized corpus maximizes the initial knowledge base performance. If your corpus is not fully categorized, your data might contain structural information that can be used to categorize the texts, for example:

- Newsgroup name if you are categorizing Newsgroup
- Tags in HTML data
- XML tags in XML extracts

If your data includes this structured information, it can be assigned directly as the category for each sample text, as we have shown in 3.2.2, “Building a knowledge base” on page 62.

Classification Workbench categorization

Classification Workbench offers several methods to assist you in the categorization process:

- Clustering: Helps you identify initial groupings (that is, clusters) of texts that are textually similar in an uncategorized (or partially categorized) corpus. Clustering can also help you to identify new categories in a collection of categorized corpus texts.

When using this clustering function, you do not control how Classification Module clusters the corpus texts. However, you control the number of clusters that Classification Module produces. After you use the clustering process, you must analyze what each cluster represents. For example, you can look at the dominant concepts that are highlighted in each text to assist you in this analysis.
To understand this functionality, consider the task of categorizing the following texts: pen, water glasses, coffee cup, laptop, and writing tablet.

If you are asked to make two clusters, you can find objects for writing and objects for drinking. If you are asked to make four clusters, you can find laptops, writing instruments, water glasses, and coffee cups.

Clustering is also available in the Taxonomy Proposer. Taxonomy Proposer also assists the user to iteratively build a taxonomy from the suggestions of the Clustering algorithm.

**Tip:** Follow the Taxonomy Proposer Workflow Assistant.

- **Classify by response:** If your sample texts include standard responses, you can use Classification Workbench to automatically classify corpus texts according to the particular canned responses that they contain. The system identifies common responses even if they have been modified slightly based on the specific response, such as adding the customer’s name and account number.

- **Finding patterns:** If your text does not contain standard responses, but it contains repeating patterns that can be used to categorize the data, you can use Classification Workbench to find these patterns. The resulting list of items can be examined as a possible category.

- **Manual categorization:** This method is the most labor-intensive method, but it is sometimes the only option available. If your test system does not have any systematically identifiable categories, you can use Classification Workbench to assign a category to each text.

Your use of these techniques depends on the state of your texts. In most cases, corpus texts require categorization and cleanup. Depending on how the resulting knowledge base performs, you can use the categorization and cleanup techniques to fine-tune the corpus and improve knowledge base performance. Optimal training results are achieved when your corpus contains texts that are as close as possible in content and structure to the real-life texts that Classification Module will categorize.

### 4.4.5 Overcoming the lack of data: Using initialization data

You can use initialization data to populate a new knowledge base or a new category within an existing knowledge base. Initialization data can be used instead of, or in conjunction with, corpus-based knowledge base creation. This method can be especially useful when a corpus of sample texts is not available.
In spite of their limitations (minimal length and content), within the question and answer types of applications, it is best to train Classification Module using pre-categorized question texts, because these question texts are the texts that the system will be required to classify. However, when sample questions are not available, you can use Classification Workbench’s initialization functionality by providing answer texts, a manually created list of keywords, or even the category names to train the Classification Module. This initialization functionality is useful both when initializing an entire knowledge base and when adding a new category to an active knowledge base.

Initialization data

Initialization data consists of keywords and texts that are associated with individual categories. Keywords are words or phrases that you expect will appear in texts classified by Classification Module. When the system identifies keywords in a text, the category associated with the keywords is more likely to be returned.

For example, you expect to receive questions about your company’s exchange policy. You want to choose keywords, such as “exchange” and “return” for this category.

Note: You can attach the same keyword to more than one category.

In addition to keywords, you can associate one or more texts with categories in the knowledge base. For example, for a Classification Module-based e-mail classification system, an appropriate text is a “canned answer” sent in response to e-mail inquiries. Classification Workbench analyzes the text and uses this information to classify incoming texts appropriately. You can also add shorter texts, referred to as titles, to a category (for example, the subject of the canned answer).

You can choose to enter initialization data manually or to gather data offline and import it.
When using the initialization functionality, the following predefined content types are available:

- **Title**: Used for short descriptions
- **Canned answer**: Used for predefined texts that the application returns for a category
- **Keywords**: Words or phrases associated with categories

You cannot change these predefined types.

**Overcoming the lack of categories: Taxonomy Proposer**

The Taxonomy Proposer has clustering capabilities similar to those of Classification Workbench. In addition, it provides an interface for manipulating and organizing the clusters into categories.

Taxonomy proposer assists in the situation where, although there is sufficient data, the customer is unable to define categories. Taxonomy Proposer helps to discover possible categories among the content set.

### 4.5 Building and training the knowledge base

The crucial stages of building and training a knowledge base involve defining the correct categories and preparing the content set. After you have your categories and corpus, you can build and train a knowledge base, as described in Chapter 3, “Working with knowledge bases and decision plans” on page 59.

In this section, instead of focusing on the mechanics of performing the tasks (which is covered in Chapter 3), we focus on the two methods of adding learning to a knowledge base and the qualitative differences of the two approaches. The two methods are creating and training a knowledge base with existing data and giving feedback online to an existing knowledge base.

#### 4.5.1 Building a knowledge base with existing categories and data

You can build knowledge bases with existing data:

1. Split the existing data into two parts.
2. Use the first part to train the initial knowledge base.
3. Use the second part to analyze the initial knowledge base.
In the next sections, we discuss using precision and recall with thresholds to analyze your knowledge base.

4.5.2 Building a knowledge base when the categories are unclear

You can use Taxonomy Proposer to create clusters of data that serve as the basis for proposed categories. Using this tool, you can accept, reject, and combine clusters in order to discover categories within your data.

4.5.3 Training as opposed to online learning (feedback)

An Classification Module-enabled application accepts feedback and passes it to Classification Module. Classification Module, in turn, uses this feedback to modify the knowledge base during an incremental process known as online learning (or just feedback).

Online learning is most suitable for live maintenance of a knowledge base. Because Classification Module is designed to learn from new data, it considers new feedback as more relevant than old feedback. We discuss the feedback process later in this chapter.

If you need to rewrite your application from the beginning, import the knowledge base into Classification Workbench, collect all relevant feedback in a content set, and run the Create, Analyze and Learn Wizard using the “Create and analyze knowledge base using active view” option. You might decide to keep the old knowledge base structure or to take the structure from the content set data. This process of “teaching the system from scratch” is referred to as training. Use the training to get started, or use it after making significant changes to the knowledge base.

Tip: We recommend that you keep a small to medium-sized corpus to be used as a benchmark for checking after retraining or periodically after feedback. This corpus needs to contain a broad range of the most common types of queries. The purpose of this corpus is to test basic functionality after fine-tuning more specific areas. Often feedback is biased toward a few newer and popular categories.
4.5.4 Discover categories from your data using Taxonomy Proposer

In cases where you do not have a clear picture of the data structure, Classification Module Taxonomy Proposer is one possible tool that can suggest categories by analyzing the existing content. Taxonomy Proposer provides you with groupings that might suggest categorization and guides the user to iteratively build the taxonomy.

4.6 Analyzing knowledge base performance: Identifying the problems

Whether you are examining a new knowledge base or analyzing the performance of an existing knowledge base from your live application environment, the same best practices apply.

4.6.1 Analyzing your knowledge base

You need to run analysis:

- Before launching a knowledge base in a live system. Only by running analysis can you feel confident that your knowledge base will meet your requirements in the application. Analysis before a live launch is critical when Classification Module’s suggestions are presented to external users. For example, a self-help application on your company’s Web site must provide accurate results immediately. However, an internal help-desk application can tolerate reduced accuracy at the beginning, because the users are internal employees and can wait for the accuracy of the system to improve with their use and feedback.

- On a regular basis, in order to check the effect of feedback over time and in accordance with the business requirements. If an organization requires a monthly report of system performance, analysis must be run monthly. For this purpose, it is helpful to maintain a benchmark corpus that includes the most relevant and frequent types of texts.

- When you suspect degradation in performance. Systems that accept live feedback from various sources can have unexpected changes in knowledge base performance.

- When significant changes are made to the knowledge base (for example, due to changes in business needs).
This section describes the analysis of a knowledge base after it is created, but before it has been deployed on a live server (the first option listed). We discuss the analysis of live performance later in this chapter.

To analyze your knowledge base’s performance, use a set of texts (also referred to as an analysis content set), consisting of texts that have been categorized (by a subject expert or other business user) into the appropriate categories. For each text, the categories and their suggested scores that were suggested by Classification Module are compared with the categories that were chosen by a person. This information enables you to understand how your knowledge base aligns with the accepted knowledge base design and also how it will perform on live data. Using this process, you can gain insight into potential areas for improvement.

Specifically, analyzing the knowledge base’s performance allows you to gain the following benefits:

- Identify how accurately the knowledge base performs.
- Analyze the use of thresholds for process automation.
- Identify categories that require improvement.
- Identify possible new categories to add to the knowledge base or obsolete categories that need to be removed.

The analysis process consists of understanding your analysis data, viewing and understanding performance, and identifying categories for improvement.

Figure 4-1 on page 143 shows the stages of knowledge base analysis and the activities that you perform during each stage.
4.6.2 Sources of analysis data

There are typically two types of data: exported analysis data (from a live server) and data created within Classification Workbench for investigating specific aspects of a knowledge base, using a copy of the knowledge base that you want to analyze.

**Exported data (analysis of live performance)**

You can set up Classification Module to save all the matches (and scores) that are returned by the server, along with the feedback given by the agent. This data is in XML format and can be imported into Classification Workbench. Analysis of this data reflects the performance of the Classification Module over time (history). Within integration or custom applications, you can collect live data independently for analysis purposes. For instance, the Classification Module installation provides a tool for extracting analysis data from the integrated IBM FileNet P8 server.
**Benchmark or specific case data**
You can collect this data manually from various sources and save it as a csv file to import into Classification Workbench. Or, Classification Workbench can import this data from a file system. It is helpful to save benchmark data to perform verification checks on the knowledge base after periods of feedback.

**Data gathered for creating the knowledge base**
In order to create a new knowledge base, you need to create a test corpus, which contains the categorization of texts. Divide this data randomly, but distribute it evenly among the categories: use one half to create the knowledge base and one half for analysis. Use this method to avoid the bias that results when a knowledge base tests itself on texts that it has already *learned*.

This section focuses on analyzing a newly created knowledge base (the last option listed) before it has been in production. Later in the chapter, we discuss the analysis of a knowledge base that has been in production.

**4.6.3 Scores used for analysis**

The analysis process is performed on the scores that each text received and its correct categorization. Two possible sources for these scores exist:

- The real scores returned by Classification Module (stored in the Classification Module server or extracted from the ECM repository). The analysis data includes the relevancy scores of each suggested category (or “match”). These scores are included in all Classification Module exported data, and they must be included in any custom mechanism for saving analysis data.

- When analysis data is created within Classification Workbench, these scores are automatically calculated and saved with the matches.

**4.6.4 Understanding your analysis data**

To analyze your knowledge base, you present it with a sampling of texts that are representative of the texts received by the live system, such as e-mails or documents. The accuracy of the analysis depends on the quality of your analysis or test data.

**Data quality**
Your data must contain the correct categories for each text, which are obtained through manual verification, such as feedback or any other reliable mechanism. These categories represent your expectations and provide the basis for the calculation of success. In addition, your test data must be relevant, unbiased, and reliable. Working with feedback describes the ideal data for feedback.
Similar concepts apply to ideal analysis data, because the category field represents the ideal categorization of the text.

**Category distribution**
The distribution of your analysis data must be representative of the distribution of the actual data that Classification Module will classify using your knowledge base. For example, if you expect that Category A will account for 20% of your classifications, the analysis data needs to reflect that, as well.

In general, focus on the most frequently used categories during the analysis process. In certain cases, extremely small categories can be highly important. Make a note of these categories, so that you can give them special attention during your analysis.

For specific case analysis, use selected categories.

**Languages**
The language of the analysis data needs to agree with the categories that are being analyzed.

### 4.6.5 Running analysis in Classification Workbench

In order to analyze your application’s data in Classification Workbench, you must gather the data and export the following elements in a format that is recognized by Classification Workbench:

- **Knowledge base**
- **A corpus of texts sorted in the order of their arrival.** The corpus can be in a format, such as CSV, XML, or a Classification Module corpus. The corpus can also be imported from a file system. In addition, it is preferable for the corpus to contain the original scores, if this data has been extracted from a Classification Module system.

When you import XML data that is extracted from a running server, all of the necessary analysis data is included; therefore, you can run the reports directly after importing the data.

In order to run analysis from a classified corpus within Classification Workbench, you need to run the “Analyze Knowledge Base” option in the Create, Analyze and Learn Wizard.
4.6.6 Measures of accuracy

The goal of analyzing a knowledge base is to determine if it will perform well in your application. How you define good performance depends on your application and your expectations. In each type of application, Classification Module solves your categorization requirements differently. Similarly, interpreting analysis results is based on how you measure your application’s success.

Classification Workbench provides a wide range of reports. Several of the reports and graphs provide you with a view of the knowledge base's overall accuracy, while other reports and graphs allow you to see how each category performs within the knowledge base.

The most important measures of your knowledge base’s overall accuracy are the Cumulative Success Summary report and the Precision vs. Recall graph. Cumulative success represents how often the top categories suggested by Classification Module are correct. Precision is the percentage of text that the application correctly identifies as relevant to a category based on Classification Module scores. Recall is the percentage of text that is relevant to a category that the application identified as relevant based on Classification Module scores. Precision and recall are especially relevant measures for automatic response and automatic routing systems. There are separate percentages for matches received in the top category, in the two top categories, and in other categories. The percentages for matches measurement is generally the most relevant measurement for self-service and agent support applications. Next, we provide more details about interpreting these reports and the meaning of the various scores.

Select the analysis measures that are appropriate for your application. If your application will present the top three categories that Classification Module chose for a specific text, you need to know how well Classification Module places the correct answer in the top three categories. If your application allows you to set a threshold for an automatic action, you will need to examine various auto-classification thresholds and the precision/recall that is offered by each threshold, specifically:

- Automatic e-mail reply systems: Your application receives user queries and immediately presents the three most likely answers in response. Your users

Tip: We recommend to store the ID, name, or other identifier of the person who provided the original feedback in the system. This information can be useful when running reports and performing analysis based on specific users or agents.
are happy to receive the correct answer as one of the top few suggestions, while unanswered questions can be directed to an e-mail or a phone number.

Your application measures success based on cumulative success or how often the Classification Module identifies the correct answer as one of the top three answers.

- Custom spam filtering systems or Spam Filtering: Your application categorizes incoming messages as spam or non-spam. Your users cannot afford to lose more than one non-spam e-mail in 100,000, but they are willing to live with a small amount of spam that remains unfiltered.

A successful knowledge base can assure that an extremely high percentage (0.001%) of texts that belong to non-spam are assigned correctly (high precision), while a fairly high percentage of spam messages are correctly assigned (reasonably high recall). Classification Workbench allows you to set thresholds per category to adjust the precision/recall levels.

- For self-help applications, the user is expected to be able to choose the most appropriate answer from a list of suggested results. You will be satisfied with your knowledge base as long as it identifies most of the interested users (moderate precision), while you will excuse false positives in this category (moderate recall).

### 4.6.7 The process of analysis

*Analysis* is the process of testing your knowledge base's performance against a content set of representative texts, e-mails, or messages.

Typically, users will collect tens or hundreds of recent text items that the system has recently classified per category. Each text will also contain an agent-assigned category. Each of these texts is scored by Classification Module using the current knowledge base. Analysis of the knowledge base is based solely on the comparison between Classification Module's categorization and the correct categorization of the test set.

### 4.6.8 Automation and accuracy

One of the goals of analysis is to find a level of automation that provides an appropriate balance between the level of accuracy and the coverage for your needs. For example, you might determine that by dropping the accuracy levels of auto-responses from 90% to 80%, you can reduce your staff by 25%. Alternatively, you can describe this scenario as a 33% increase in staffing in order to increase accuracy from 80% to 90%. Is this price worth paying? This decision is an executive business decision that must be made by business managers.
As a result of analysis, you might determine that your current knowledge base does not give you the levels of automation and accuracy (that is, return on investment (ROI)) that you require. If so, you must determine the causes and attempt to correct them.

If your application relies on precision and recall ratings, examining a graph of the average precision and recall for all categories might give you sense of the accuracy that you can expect for given levels of automation. You will probably want to examine individual categories and determine the correct level of automation for each category.

### 4.6.9 The number of categories returned by matching

Classification Module allows you to designate how many top-scoring categories will be returned by Classification Module. Consider the following factors when making this decision:

- An increase in the number of categories returned will increase processing time and the memory required.
- If your application makes decisions based on the thresholds of specific categories, rather than the top-scoring categories, you might want to return all the categories that exceed a specified threshold.
- When performing analysis, Classification Module needs to return the same number of categories that are returned in the application. Otherwise, the measured accuracy can differ from the real accuracy.

**Note:** Classification Module does not return categories that receive scores of exactly zero.

### 4.6.10 Typical analysis process (report interpretation)

Classification Module provides various reports to help you to determine the overall performance of your knowledge base.

A typical performance analysis of your knowledge base consists of the following steps:

1. Check the summary report.
   
   Look for good performance and poor performance categories. Analyze data as to certain categories are poor categories:
   
   - Is it because the data is bad?
   - Do we have duplicate categories?
Are there insufficient items for specific categories? Typically, the performance of a category decreases if it has fewer items than other categories.

2. Look for categories with lots of content but poor performance.
   Possible solutions include:
   - Merge overlapping categories.
   - Split categories with multiple intents.
   - Create new categories.
   - Use precision to eliminate false positives.

3. Look at other reports for problems:
   - Steal and stolen report: Do you need to merge the categories or retrain with more accurate texts?
   - Precision vs. recall report: Review the report, explain the report’s meaning to the client, and understand what the client wants to do with this information. Does the client want more precision or more recall? You can adjust accordingly. Understand False Positive (false yes) and False Negative (false no).
   - Threshold information: Check the threshold setting and determine how it can be set for better precision or recall scores.
   - Deflection report information: Analyze the report and learn how you can use the information for better performance.

In the following section, we cover Cumulative Success and Precision/Recall reports. Also, you can review other reports to analyze the performance and improve your knowledge base.

4.6.11 Reviewing Cumulative Success scores

*Cumulative Success scores* refer to the percentage of correctly matched categories among a specified number of the top scoring categories. Usually, cumulative success scores come as a set of values, and the values can be presented in a table or graph, as shown in Table 4-2 on page 150.
Table 4-2  Cumulative Success scores

<table>
<thead>
<tr>
<th>Number of categories returned</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of items correctly classified</td>
<td>81</td>
<td>85</td>
<td>90</td>
<td>92</td>
<td>94</td>
<td>95</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

This table indicates for this knowledge base that Classification Module selects the correct category as the first choice 81% of the time. It also indicates that if the application presents the top seven scoring categories for an incoming text, the correct answer will be included 99% of the time.

Higher percentages reflect better performance. Make sure that you check the cumulative success when running analysis on any newly improved knowledge base.

4.6.12 Understanding precision and recall

Another tool that you can use to determine the overall performance of your knowledge base is Precision/Recall analysis. You can use a Precision vs. Recall graph for all categories in the knowledge base. The sample graph, as shown in Figure 4-2 on page 151, shows a knowledge base that performs well - the curve is in the upper-right portion of the graph. When the curve is in the lower-left portion of the graph and the number of categories is small, this graph indicates that the knowledge base’s overall performance is poor. If the results are poor, you might want to view the Precision vs. Recall graph for each category to see if individual categories are particularly problematic. You can then proceed to examine the individual categories' performance.
4.7 Fine-tuning your knowledge base: Fixing the problems

When the overall knowledge base performance does not meet your application's requirements, you can use the Classification Module reports and graphs that you use for analysis to identify categories that are candidates for accuracy improvement.

The following steps are your guidelines for fine-tuning your knowledge base:

- Correct the knowledge base structure, such as an improper hierarchy between nodes, and errors in the definition of rule nodes.
- Make changes at the business level, such as adding, dividing, or combining categories. In addition, you might need to reorganize your knowledge base to support business changes.
- Identify gaps between the actual use of the categories and the intent of the knowledge base designer. If you identify gaps, you might need to reclassify problematic texts and determine if this reclassification improves the accuracy of the system. If so, communicate these changes to the users who provide feedback.
To further help you fine-tune your knowledge base, we examine the following issues and topics:

- Overlapping categories (categories stealing from each other)
- Low scores in all categories
- Category does not represent intent of a message
- Possible multiple intent categories
- Human error (poor manual classification)
- Poor sampling (randomization issues)
- Identifying hidden subcategories
- Poorly performing category with too few examples
- Identifying obsolete categories

### 4.7.1 Overlapping categories (categories stealing from each other)

One of the issues that can be detrimental to the performance of a category relates to the presence of overlapping categories. To determine if a poorly performing category has a problem with overlapping categories, look for a set of test texts in a specific category that receives higher scores in other categories.

You might notice that a category steals from several other categories and that other categories steal from it. Maximum scores are low, and the Corpus Scoring graph shows a score limit at a 50 - 60% confidence level. A good place to look for this problem is in the largest categories. Generally, a category with multiple intents becomes very popular and gets a lot of feedback.

When Classification Module returns Category A as the highest scoring match for texts in Category B, Category A is said to be *stealing* from Category B. This problem indicates that texts that should have received top scores in Category B received top scores in Category A.

When examining possible overlapping categories, it is important to note that the problem can be compounded when the corpus text scores high in both a correct category and an incorrect category. In this case, examine the text to see if it actually contains multiple intents. If your application will return more than one of the top scoring categories, you might be satisfied when one of the top matches returned by Classification Module is correct.

Another situation where you have a high stealing and stolen percentage between categories is when two categories are created because they handle separate business needs, but they represent the same message intent as in the following situations:

- Situation one: Two applications both have login issues, but they require different responses.
Situation two: Two categories are created (Mortgage Interest and Checking Interest). Mortgage Interest supports interest rate questions for mortgages. Checking Interest supports interest rate questions for checking accounts.

Situation three: Out of Stock category is similar to Inventory category (one category is a subset of the other category).

Resolving overlapping categories
Divide or restructure the categories so that all the texts contained in a given category are similar and distinguishable from the texts of other categories.

Depending on your assessment of the cause, use one of the following solutions:

- Situation: Categories are closely related, and they might be treated as one category.
  Solution: Merge the weaker category into the stronger, better performing category, and retrain the category based on all available texts in the category.

- Situation: Categories are similar, but external criteria determines to which category a particular text belongs.
  Solution: Create a rule in the decision plan to identify and route texts to separate categories according to external criteria. If necessary, add a function in your application to help you with this routing. For example, if categories depend on the membership level of the writer and this information can be determined by the application, make this information known to Classification Module so that a rule can be applied.

- Situation: Categories are unrelated, although there might be a similar tie in the customer intents.
  Solution: Decide on clear criteria for inclusion of a text in each category and make these criteria known to everyone who will be manually classifying texts in your system. Go back through texts in both categories, correct any misclassified texts and use the learning or training mechanism, as required.

4.7.2 Low scores in all categories

When analysis for a specific text shows low scores for all categories in the knowledge base, typically, the text does not fit into any of the categories in the current knowledge base.

Possible explanations for this situation include these reasons:

- There is no problem with the categories. These texts need to receive low scores. However, you might want to add a “Not in Knowledge Base” category at the top level of your knowledge base and provide this text as feedback to this new category.
There is a problem. However, this kind of text is so rare that it is not worth providing a solution.

The text is currently classified to a generic category. If we divide this generic category and allocate a separate category for this kind of text, Classification Module will be able to identify this type of text in the future.

There is a technical problem, such as an incorrect rule or a problem with language identification.

The text represents a new topic that requires a new category because of a predicted need.

4.7.3 Category does not represent intent of a message

The name of the category usually indicates its nature (for example, Escalation). A category steals from several other categories. And, other categories steal from it. The cumulative success is extremely low. The maximum confidence level is low.

This issue happens with one of the following situations:

- Situation one: This situation is the result of a “Special offer” and is unrelated to the incoming messages.
- Situation two: Separate business units have different responses for the same categories, such as Firewall messages as compared to Firewall Pro.
- Situation three: There is a generic answer when the incoming message is missing important data (for example, the account number is missing and it is necessary to answer the question).
- Situation four: There is a generic answer when multiple issues have the same response (for example, “No future plans to release products”).

4.7.4 Possible multiple intent categories

When analysis for a specific text shows high scores in multiple categories, the following reasons might apply:

- The text contains multiple intents. In this case, it is correct for the text to receive a high score in both categories. You will want to make sure that your application correctly interprets these multiple high scores and acts accordingly.
- An incorrect category is attempting to steal the text from the correct category. In this case, see 4.7.1, “Overlapping categories (categories stealing from each other)” on page 152.
Errors in the manual classification that you used to train the system have caused categories to be confused (with multiple intentions). You must always define clear objective criteria for inclusion in a category.

If you determine that two categories are often coupled, you might decide to merge them. That is, the multiple intentions might be extremely close in meaning and impossible to separate in practice.

4.7.5 Human error (poor manual classification)

Assuming that most of the texts in your training and testing corpora were classified properly by your organization, you will often find that the reason that texts score poorly in their category is due to manual classification errors. Texts scored poorly, because they actually do not belong to the category that was manually selected.

Classification Module learns from your examples, and it is crucial that your examples are classified correctly. When you analyze this problem and determine that texts have been misclassified, you can reclassify them and check whether this problem is solved. In addition, if you detect a problem with the feedback that your users submitted, you must clarify the intents of all categories within the team. Classification Module can learn fairly well, even in the presence of a few categorization errors. However, the analysis process (using manual classification) is strictly based on the categorization of the texts, and any error in the texts that are used for analysis is reflected in the analysis results.

Detecting manual categorization errors

You can identify manual categorization errors by examining texts that received unexpected scores in Classification Workbench’s Corpus Scoring charts or other tools.

Another mechanism is using Classification Module’s clustering algorithm to cluster texts in a corpus. For example, if you have a corpus containing 30 categories, it might be helpful to generate 30 clusters to see how Classification Module groups similar texts together. By reviewing all of the texts in a specific category and then looking at the texts that Classification Module clustered together, you can see which texts are most similar. Texts that were not clustered together might have been misclassified.
4.7.6 Poor sampling (randomization issues)

An extreme case is when certain categories appear in the test set, but these categories are not included in the knowledge base or training set. This situation can happen if you train on the first half of the corpus and test on the second half of the corpus, and a newly defined category only appears in the second half of the corpus. Another possibility is if the corpus is sorted by categories. Always make sure that you test the knowledge base on relevant categories.

If you discover a sampling error in your test set, consider reorganizing the order of the texts so that the distribution of categories among the test and training sets is relatively even. Alternatively, you might want to obtain a more representative sample.

Sometimes, performance for a specific category can suffer due to sampling errors. It is important that the training and testing sets each contain similar proportions of texts in a specific category.

If your training set is made up of 50% of texts related to the “Unsubscribe” category, a similar proportion must be expected in the test set.

4.7.7 Identifying hidden subcategories

Sometimes, your knowledge base seems to perform well, but you suspect that it is not sufficiently granular. Using clustering, you can divide all of the texts assigned to a specific category into several clusters. When all of the texts assigned to a specific category are clustered, you can examine texts in each cluster to try to discern differences in the clusters.

To understand what makes a cluster unique, you can treat the clusters as categories:

- Create a knowledge base based on these categories/clusters.
- Examine individual messages in each cluster, viewing “significant concepts”. These concepts are often the keys to understanding the subtle differences between clusters.

If you discover a subcategory for which it makes business sense to create a new category, you can choose to delete the existing category and to create two new categories. Alternatively, you might want to create just one new category and clearly define the criteria for manual categorization in each category.
4.7.8 Poorly performing category with too few examples

If the results of your analysis indicate that a category is performing poorly, the reason might be that the category has too few examples. For optimal Classification Module performance, the more examples per category, the better. If you have identified a corpus with too few examples for the training and testing corpora, try to collect additional representative examples.

If collecting a large set of examples is problematic, one solution is to use the same examples for analysis and knowledge base building, paying careful attention to the order of their use:

1. Start building your business taxonomy and train an initial knowledge base using one of the following techniques:
   - Use the Taxonomy Proposer to start the clustering process with a small sample of documents. Build your business taxonomy iteratively using the hints provided by the Taxonomy Proposer.
   - If you have knowledge of your data, start with a partially trained knowledge base. Use initialization techniques, such as keywords per category.

2. After the initial knowledge base is created, you can add learning using data from additional sources.

3. Import the knowledge base into Classification Workbench.

4. Run analysis, on the additional data, for instance data exported from your server.

5. Depending on whether you have feedback, there are two possible scenarios:
   - With feedback: We advise that you examine the documents one at a time by order of arrival. The order that they appear in the corpus has impact on the learning: the later the item is listed, the more impact it has on the knowledge base. You need to review the feedback (classification category) to make sure that the category choices are valid.
   - Without feedback: After the scores are received for a given text, supply feedback for that text, by assigning it a category. This way, you enrich and maintain your knowledge base.

6. Run the Analysis wizard, choosing the “Learn during analysis” option.

7. This analysis can be an iterative process, using small sets of data to test incremental changes within the knowledge base. Each time, use the most “educated” knowledge base (that is, with the most up-to-date learning) for each subsequent text analysis. Use this process if you work continuously in one project.

8. When the knowledge base behaves as desired, return it to the live server.
As the analysis progresses, the knowledge base that is used to test each text will always contain the maximum amount of knowledge. In this way, every text used in analysis is also learned, creating a knowledge base much like one in a live system.

4.7.9 Identifying obsolete categories

Over the course of time, you likely will add new categories to your knowledge base. Just as new categories are added, old categories become obsolete. By looking at a representative sampling of recent system feedback, determine if certain categories are no longer in use in your system. Frequently, the fact that a particular category is no longer in use will be valuable business feedback. It might mean that customers no longer report a certain kind of issue or that a product feature has become obsolete.

4.7.10 Automation thresholds

Automation involves actions taken by an application based on Classification Module scores, without any user intervention. Automatic action depends on the nature of the Classification Module-enabled application and might only be appropriate for a limited set of categories. For example, your e-mail response application might automatically unsubscribe users who sent e-mails with “unsubscribe” in their subject lines. You can easily implement this automatic action with a Classification Module rule.

Typically, automation is based upon the thresholds that are defined within each knowledge base node or that are defined within the decision plan. Within the decision plan, you can set the threshold for a category dynamically, such as when a specific rule fires (for instance, you can lower the threshold in the case where certain key terms appear in the text, increasing the threshold’s chance of being matched for that category). You can also set the threshold of whole groups of categories according to specific threshold files.

You can select your threshold based on one of the following areas:

- Capacity
  
  Set your threshold based on the number of texts that you can handle in your organization. For example, if you have personnel available to support 100 messages per day for a specific category, you need to set your threshold so that only 100 messages will exceed this threshold. Based on your previous experience or using Classification Workbench, you might know that if you set your threshold for a category to 80%, you will receive about 100 messages per day for that category. Alternatively, your application can also use a queue to hold the top scoring texts for each category.
Recall
Determine the percentage of texts that you want to catch for a specific category. Realize that recall and precision are inversely related and that there are costs associated with catching a higher percentage of texts. For example, a filtering application in a security environment might opt to set the recall value for the Terrorism category to 90%.

Recall = The percentage of items that are actually relevant to the category, which are recognized as such by Classification Module (the rest are “false negatives”).

High Recall means that you do not have many “false negatives”; you do not “miss” many items; and you “catch” almost all of the items that belong to the category.

Precision
Determine the percentage of texts that are caught correctly for a given category. For example, you might want 90% of the automatic responses to be correct, regardless of the associated cost.

Precision = The percentage of items that Classification Module identifies as relevant to a category, which are actually relevant to the category (the rest are “false positives”). High Precision means that you do not have many “false positives”; you do not claim many items to belong to the category when in fact they do not belong.

Note: Selecting a threshold based on recall or on precision is possible using Classification Workbench.

Testing your knowledge base: Controlling precision/recall with thresholds
Your company wants its Classification Module-enabled application to automate responses to e-mails in the category “New Product Release Dates”, but only if it can be assured that Classification Module is 80% confident that this answer is correct:

Question: If you set your threshold for automatic responses for a category to 80%, and if this threshold yields a precision of 85% and a recall of 70%, what does that result tell you?

Answer: The decision to send an automatic response will be correct for 85% of the automatic responses that are sent for the “New Product Release Dates” category. In addition, the system detected 70% of the messages that belong to this category (that is, it missed 30% of the messages that belong to this category). In addition, an automatic response will be sent when Classification Module is at least 80% confident that this message belongs to this category.
Question: If you want to be accurate 85% of the time (that is, the precision), what do you do?

Answer: Make sure that your scores are calibrated and then set the threshold to 80%. For the Web Self-Service page on your company’s Web site, or within the ECM or records management automation, the system is completely unmanned, and your goal is to provide the three most likely responses to every question asked.

Question: When you provide the three top scoring categories (and their associated answer texts) to the user, how do you know your level of accuracy?

Answer: Look at the Cumulative Success value “3”. This value will tell you the percentage of times that the correct answer was included in the top three Classification Module matches.

Question: If you want to automatically archive e-mails into an ECM Repository (for example, IBM FileNet P8) with a precision of 90% (that is, precision), what do you do?

Answer: The e-mail archiving scenario using IBM Content Collector to inject e-mails into IBM FileNet P8 or IBM Content Manager CM8, can be performed automatically by using Classification Module to decide the location of the e-mails inside the repository. For this purpose, we calculate using Classification Workbench the thresholds setting for Precision 90% and use them in the decision process (for example, you can set a decision plan that will use a knowledge base and its associated thresholds set for precision 90).

The same automation process based on thresholds can be used for e-mails inside the ECM repository in the reclassification process.

4.8 Maintaining your knowledge base over time: Using feedback

In the previous sections, we identified a number of problems typically encountered during the knowledge base creation phase, before deployment.

After a well-trained knowledge base is deployed into the production environment, the biggest challenge is how to maintain its accuracy over time. Because the knowledge base can receive feedback over time, its nature is dynamic and cannot always be predicted. For this reason, you need to monitor knowledge base behavior.
4.8.1 Working with feedback: Avoiding the problems

The knowledge base adapts itself to new expectations by learning from user input. If this user input is inconsistent among the users, the impact is negative.

There must be a clear definition and use of categories in the knowledge base. The design rationale and the use of feedback must be clearly presented to all of the users who will provide feedback.

Deferring feedback
Depending on the type of user that supplies the feedback (internal personnel or random users), it might be advisable to defer the feedback, saving it on the server but not sending it to the knowledge base for learning. You can set up this option, prior to giving feedback, using the Management Console. There is a feedback option in the Properties dialog for each knowledge base. Choose “Defer processing”. When you want to process the feedback, run bnsExtractText, exporting only the deferred feedback. This data can be imported into Classification Workbench and audited before running the “Learn using active view” option in the Create, Analyze and Learn Wizard.

Ideal feedback data
Classification Module learns from examples. It learns best from examples that provide a good representation of the real-world environment, both the texts themselves, and the order in which they are received. Under optimal conditions, the data from which Classification Module learns needs to have these characteristics:

► Relevant: The feedback text sent to Classification Module must contain relevant data only.
► Unbiased: Texts sent as feedback to the Classification Module must be an accurate representation of the content (and the quantities) of real-world data.
► Reliable: Although Classification Module will not be dramatically affected by occasional feedback errors, generally feedback must be submitted by knowledgeable users.
► Order of feedback: Feedback must be supplied to the system in a natural order, that is, by arrival time, and must not be artificially reordered.
► It might not always be possible to operate under optimal conditions. Bending the rules occasionally will not necessarily interfere with the proper operation of Classification Module, but you must always strive to comply with these recommendations for optimal Classification Module behavior.
**Relevant data**
Preparation text prior to sending it to Classification Module will help ensure that the system works with the best possible data. This recommendation applies despite the fact that Classification Module is designed to handle imperfect data and automatically ignores nonrelevant texts.

**Note:** The same text preparation needs to be applied to all texts sent to the Classification Module for training, learning, and matching.

**Unbiased data**
Under optimal conditions, it is best to provide accurate and unbiased data to Classification Module as feedback. When unbiased data is not available, you must carefully select a random sampling of this data and use it as feedback.

The following situations are typical examples of bias in the training set/feedback:

- **Self-help applications:** When feedback is received from users who click a feedback button (for example, “Did this answer help you?”), it is possible that only satisfied customers (or only unsatisfied customers) will submit feedback. This data is both biased and unreliable. Therefore, it is preferable to manually audit this feedback. An objective auditor might use a separate “Feedback Tool” to audit (for example, confirm or modify) the user’s feedback, which provides a more reliable and less biased source of input to the Classification Module.

- **E-mail auto-response system:** Bias can occur when agents only provide feedback on manually handled messages, while messages that receive auto-responses do not receive feedback. In this situation, Classification Module will not learn the characteristics of messages with high scores (in this example, the auto-response messages), and the results will be biased.

This type of bias might not always be undesirable. We recommend that you designate a random percentage of auto-responses for manual audit and quality control.

Alternatively, consider this scenario: the agent can only view and provide feedback to the five high-scoring categories suggested by Classification Module and cannot access the other (low-scoring) categories in the knowledge base. This undesirable situation will create the opposite bias described in the previous example, because only the high-scoring categories receive feedback. Therefore, agents must always have access to the entire knowledge base.

- **Certain users of Classification Module-enabled applications do not provide positive feedback when Classification Module suggests the correct category and answer with a high confidence level. Often, low-scoring texts (which are more likely to be selected for feedback submission) are not good**
representatives of the category and might not be the best training material. Feedback must be implicit (triggered by user's actions) whenever possible. For example, the selection of a canned response needs to automatically trigger a feedback.

In general, Classification Module takes into account the popularity of categories in the example set and gives more weight to larger categories. As a result, any large deviation from the real world can introduce bias. For example, a category that is not visible to users and receives fewer feedbacks than it deserves will receive lower Classification Module scores in the future.

One possible implementation for handling biased feedback is to use the random sampling of texts or messages for feedback. Your application can use the results of a random number generator to select a certain percentage of the messages that will be sampled. For example, if you want to sample 10% of your messages, your application can assign a random number (between 0 and 1) to a message and select it for sampling when the random number value is below 0.1. The application must ensure that all these sampled messages receive feedback (through user feedback or auditing). The other 90% of the messages must not provide any feedback to Classification Module (regardless of the feedback that they might get in the application).

In cases where feedback can be provided to all messages, there is no need to use a sampling mechanism. This approach is specifically true for e-mail response applications that never send automatic responses.

**Reliable feedback**

Unreliable data can originate from feedback submitted by a person who is not familiar with the entire category set. For example, a user who has a problem with the mobile computer display might think that the Laptop-Problems category provided a satisfactory answer, although there is a more specific and appropriate category.

When determining how and from whom your system will receive feedback, consider the trade-off between the amount and the reliability of the feedback. Depending on the environment that you support, it is generally preferable to opt for more reliable feedback, as long as the amount of feedback is sufficient to keep your system running accurately. It is important to note that Classification Module does expect a portion of the feedback to be incorrect and, therefore, has mechanisms to avoid problems resulting from incorrect feedback.

**Feedback sent in proper order**

There is no significance to the order of texts in the corpus. However, in consecutive calls to the feedback function, there is significance to the order in which the items in the corpus are submitted. For example, the items submitted to
the most recent feedback bear more weight than the corpus presented to the feedback in previous calls.

**Repeated feedback to a single category (bulk feedback)**

It is a bad practice to give multiple feedbacks to a single category, because each separate feedback gives negative feedback to all of the other sibling categories. The application programming interface (API) allows a single feedback to contain a number of texts; therefore, we recommend this practice.

### 4.9 Analyzing the knowledge base in production

We have described the process of analyzing a knowledge base before deployment. Now, we discuss the analysis of a live server using data that is saved on the server for this purpose. Although the sources of the data differ, the principles that govern the analysis are similar.

Several options exist to evaluate the system behavior in a production environment:

- **Option one: IBM FileNet P8 integration: Export the analysis data directly from the FileNet P8 server:**
  a. Run the batch classification within Classification Center.
  b. Go through the Review process, which provides feedback to the Classification Module server.
  c. Extract the XML from the ECM repository (for example, IBM FileNet P8 repository) using the utility in the ECMTools directory (startExtractP8.bat).
  d. Import the XML analysis data into Classification Workbench. From this data, you can run reports directly without running the Classification Workbench analysis. The reports will reflect the performance of the Classification Module server over the period that is represented within the extracted data.

- **Option two: Export the analysis data directly from the Classification Module server.** To configure the server so that analysis data is saved, go to the Management Console, in Global Properties, check the Save analysis data option (all subsequent requests will be saved as analysis data). You can export the analysis data and import it into Classification Workbench. This option exports all match and feedback events, but it does not connect them in a way that can be used by Classification Workbench for full analysis of the server performance.

- **Option three: Use the Mapping Function, which is part of the specific API functions that connect the Match and feedback events.** From the external
application, you can use special functions to create a text that is stored on the server. This text is assigned an ID and subsequent functions (suggestWithID, decideWithID, and feedbackWithID) will save the data on the server in such a way that the user feedback is connected to the matched categories.

Using the command-line tool bnsExtractText, you can export XML data from the server and import it into Classification Workbench. The resulting data can be used to analyze the performance on the Classification Module server over time.

We recommend that you work in this mode if you want full control over the analysis of past knowledge base performance.

The disadvantage of this approach is that a lot of data is stored on the server. If the special ID functions are not used, the exported data will contain separate items for the Suggest/Decide and Feedback events. These items can be imported into Classification Workbench and examined manually, or alternatively, the data can be manipulated externally.

► Option four: Write an external application that stores all of the requests that are sent to the Classification Module server and that is capable of writing this information to XML files that can be imported into Classification Workbench. The disadvantage to this approach is that you must verify that the requests were actually received by the server and that no errors occurred.

For analysis options within Classification Workbench, see 4.6.5, “Running analysis in Classification Workbench” on page 145. For solutions to knowledge base problems, see 4.7, “Fine-tuning your knowledge base: Fixing the problems” on page 151.

4.9.1 Adding and removing categories

In 4.7, “Fine-tuning your knowledge base: Fixing the problems” on page 151, there are solutions to various knowledge base problems that require adding and removing categories. Before the knowledge base is in production, retraining the knowledge base using the original data is a rather simple procedure.

When a knowledge base is in production, retraining the knowledge base using the original data is not an easy task if the knowledge base has accumulated knowledge through feedback.

In production, when you remove a category, its effect on other categories fades out as the system receives new feedback. When adding a new category, make sure that the rules for this category do not conflict with any rules that were established prior to this rule. For example, if your knowledge base supports several languages, do not place an English-specific category inside a French subtree.
When a new category is added to the system, it will not be active until it receives at least one feedback. When you create the new category in Classification Workbench, you can initialize just one category using initialization data. Monitor the performance of a knowledge base with a new category, making sure that it receives appropriate feedback.

4.10 Application design considerations

The following distinct processes occur in a typical Classification Module application:

- **Matching/Suggest**: Classification Module matches an incoming text against all categories and returns match information.
- **Decide**: Classification Module works on an incoming text and extracts information based on statistical analysis against one or more knowledge bases and by applying rules. Classification Module returns a series of fields containing the text analysis information, as well as the statistical analysis matching results.
- **Learning (using the Feedback function)**: Classification Module is supplied with example texts from which it will learn how to classify messages. Typically, after a user receives suggested matches for the user’s text, the user confirms one or more categories to be the best match.

In this section, we provide you with application design considerations for each of these processes. In addition, we describe several topics for advanced users.

4.10.1 Matching

Classification is based on the scores that are returned by Classification Module’s Match function and can be integrated at one or more points in your application. The following examples are a few typical scenarios:

- **E-mail response application**

  The application can perform matching when the e-mail is received. In this case, the agent can view the match results at a later time. Or, the application can perform matching each time that the agent views the message. In this case, the classification will be based on a more up-to-date knowledge base, which can include relevant feedback that was processed. This approach will generally improve the accuracy of the system.
Self-help/search/text retrieval applications

The application calls the matching function each time that the user submits a question/query. This data can be saved and used later for analysis. If feedback is included, this feedback can be used for later retraining. Or, if feedback is deferred (not submitted to the knowledge base), it can be audited and used for offline learning.

ECM E-mail Archiving, Documents Management, and Records Management scenario

In the ECM integration between Classification Module and FileNet P8 or IBM Content Manager CM8, the classification information is used to automatically place an e-mail into the appropriate archiving folder, using IBM Content Collector to extract the e-mail content and store it in the ECM system. Or, the classification information is used to assign the appropriate document class, according to the results of a decision plan. Or, the classification information is used to assign the appropriate document class or define the needed record. You can review this information using the Classification Center process that will accumulate feedback and enrich the respective knowledge bases with new information.

4.10.2 Feedback

The sections that follow describe issues that you need to consider when working with the feedback function.

Submitting feedback

When Classification Module receives feedback, it learns from the submitted text and adds learning to the submitted category. Only learning categories are affected by feedback. Depending on the structure of the knowledge base, a node can be affected by feedback to its parent node.

The knowledge base updates itself based on texts that are supplied as feedback. The text has a positive effect on categories that receive positive feedback and a negative effect on the other categories.

Feedback for multiple intents

Certain texts contain more than one topic or question and are known as multiple-intent texts. These texts are generally associated with multiple categories. When providing feedback on multiple-intent texts, it is important to provide the feedback to Classification Module for all the categories at one time (that is, in a single call). It is not advisable to send the feedback for each category separately, because this approach provides conflicting information to Classification Module, as shown in the following example.
Consider the case of categories X, Y, and Z. If a multiple-intent message arrives that applies to categories X and Y, we must provide feedback to both X and Y in a single text with two associated categories. If we provide this feedback as two separate texts, you get the following result:

- We first supply positive feedback to category X. This action implicitly provides negative feedback to categories Y and Z (that is, the system tells categories Y and Z that the text does not belong to their categories).
- Then, we supply positive feedback to Y. This action implicitly provides negative feedback to categories X and Z.

Due to this incorrect use of feedback, Classification Module is now faced with contradictory information. When a new message arrives that is similar to this message, Classification Module will know that Z is not the correct category, but it will be confused by the conflicting feedback on categories X and Y. Therefore, follow the recommended method of providing feedback as a single text. This method provides positive feedback to categories X and Y and negative feedback to category Z with one call.

**Processing indirect feedback**

Your application can gather direct feedback from explicit user actions, during which users are aware that they are providing feedback. However, your application can gather indirect feedback from actions where users are not committed to providing correct feedback to the system. Consider an example in the self-help domain where the application is using indirect feedback. The application gathers indirect positive feedback every time that the user selects the “more information” link for one of the answers (that is, by assuming this action indicates that the suggested answer is the correct category).

How does a Web-based self-help application provide this indirect feedback to Classification Module? Classification Module must receive all of the user’s feedbacks to a given query (that is, all of the links that the user selected) as one consolidated feedback to multiple categories. If a separate user asks the same question, the application needs to send a single text, containing all the links of this user, to Classification Module as new feedback (that is, the feedback from each user must be sent separately).

**Negative feedback**

Negative feedback is information that is provided by a user indicating that the answer received is incorrect. An application can use this information, but the application must not send it directly to Classification Module for handling.
If you use an auditing mechanism to review and validate user feedback, the auditor needs to convert this information into positive feedback before it is sent to Classification Module, using one of the following methods:

- The auditor, who is familiar with the entire knowledge base, specifies the correct category for the text.
- If there is no clearly correct category in the knowledge base, but the application needs to simulate this situation by adding a “Not relevant to the knowledge base” category to handle this situation and supply this text as feedback to that category. This type of category, when added to the knowledge base, is called a background category. The background category might or might not return a score, depending on how you decide to set it up.

**Note:** Extensive use of the “Not relevant to the knowledge base” category will lower the score of the other categories in the knowledge base and might reduce Classification Module’s accuracy for the categories that are of interest. Use this category with caution.

### 4.10.3 Using Classification Module to route texts

Consider the following example.

An e-mail application uses the following Classification Module workflow:

- Every message is routed to an agent who is an expert in this area.
- Classification Module provides the agent with the suggested answers to send to the user.

How can you design the knowledge base and categories in order to best handle this scenario?

The following options are available:

- Build a knowledge base with one category for each suggested answer. The Classification Module-enabled application must handle how messages are routed to the appropriate agent (based on the suggested category). We recommend this option.

  **Tip:** You can include the agent’s name as part of the category name (for example, `agentID: category description`).

- Build a knowledge base with one category for each agent. This kind of knowledge base handles the routing of messages to the appropriate agents, but it does not differentiate between the suggested answers. This option is
not recommended, because the categories for each agent are too widely defined and are associated with too broad of a range of suggested answers.

### 4.10.4 Gathering feedback in an application

When using a custom application, you can gather feedback without submitting it to the knowledge base. In this way, the feedback can be audited and otherwise manipulated.

The time that it takes Classification Module to process (that is, learn from) a corpus of texts is the sum of the processing time per text and the processing time per category.

**Note:**
- If the size of the corpus (that is, the number of texts) is small, the processing time is proportional to the number of categories in the knowledge base.
- If the number of categories is small, the processing time is proportional to the size of the corpus.

Use the following options for collecting and handling feedback in an external application:
- The feedback can be collected and consolidated, so that groups of texts relevant to a single category are sent in a single feedback request.
- Feedback can be saved and added to a previous corpus/content set/data, allowing a full retrain of the knowledge base.
- The feedback can be collected, audited offline, and added to the knowledge base using Classification Workbench.
- The feedback can be saved on the server (and not submitted to the knowledge base), exported, audited, and then submitted to the knowledge base offline using Classification Workbench.

### 4.10.5 Using read-only and read-write knowledge bases

Certain applications require immediate match processing, such as:
- A Web-based self-help application
- A text search/retrieval system
- A help-desk Questions and Answers (Q&A) application
Other applications do not require immediate matching, such as:

- E-mail response applications in which the sender does not expect an immediate reply
- Single user desktop applications, where the user does not expect matching to work while the user is working on the knowledge base (for example, reorganizing it and training it)

Applications that require immediate matching must always keep at least one instance of the Classification Module knowledge base running. In this case, it is desirable to dedicate one instance of Classification Module to run all the learning and knowledge base management tasks and to dedicate several other servers to run instances of the knowledge base to read-only functions (each instance on its own machine).

Classification Module allows you to create a read-only instance of a knowledge base, which is significantly smaller, takes fewer resources, and loads more quickly. We recommend that the learning instance of the knowledge base run on one Classification Module server and that it generates a read-only instance for matching. The learning instance is responsible for updating the read-only instances at a regular interval.

4.10.6 Retraining a knowledge base after a major reorganization using Classification Workbench

In a typical day-to-day scenario, most feedbacks are given to existing categories. Categories are added or deleted infrequently. However, if major changes are made to the knowledge base and corpus (for example, when completely reorganizing mail folders), it makes sense to completely retrain the knowledge base, using texts that are relevant to the newly created category. In this case, the retraining process gives more weight to more recent (that is, more relevant) texts:

1. Select all of the texts from the last period of time (for example, last month), so that the average category has at least 50 - 100 texts.
2. Import the corpus into Classification Workbench, making sure that it includes classified texts for the new category.
3. Create and analyze a new knowledge base with the Create, Analyze and Learn Wizard.
4. Export the knowledge base to the Classification Module server, replacing the old knowledge base.
4.10.7 Design considerations for filtering systems

In filtering systems, only an extremely small portion of the entire data is of interest to the users. For example, in certain filtering systems, only 0.1% of the data is interesting; the other 99.9% is considered irrelevant. It is possible to design an application that places thresholds on Classification Module scores or implements a queuing mechanism in which only 0.1% of the data exceeds the threshold.

If the interesting data (the 0.1%) is actually subdivided into categories, it is generally better to divide the data into categories and to not treat it as one category. In practice, all categories will have extremely high thresholds.

Proper use of background categories

Background categories are categories that are irrelevant to the users of an application. In the previous example, 99.9% of the data is irrelevant and needs to be associated with the background category. Although the actual scores in the irrelevant category/categories are not of interest, it is important for Classification Module to obtain feedback for several of the relevant texts.

Feedback bias

Because only a small number of the texts in a filtering system are read by users (that is, the messages that receive high scores in one of the relevant categories), the overwhelming majority of messages do not get feedback. This situation creates two kinds of bias:

- Bias in the feedback to the background category
- Bias in the feedback to the relevant categories

**Note:** In general, the bias in the feedback to the relevant categories is generally more problematic than the bias in the feedback to the background categories.

For example, a drugs category might focus on domestic drug dealers and will return high scores for messages that refer to this organization. Users will not receive suggestions nor provide feedback for messages about foreign drug dealers due to differences such as the names of people, their locations, and language. This situation creates a problem with biased data.
We suggest the following methods for handling this bias in the feedback to relevant categories:

- Use an external mechanism to look for relevant messages, and feed them as feedback to new (or existing) categories in Classification Module. Such mechanisms can be a combination of:
  - External knowledge, such as newspapers, and relevant intelligence
  - A keyword-based search engine
- Audit and give feedback to a certain percentage of the messages that were filtered out.
- It might be desirable to create two separate categories (for domestic and foreign drugs) due to the differences in the content of their texts.

The method for overcoming the bias in the background category is to audit a certain percentage of the texts that were filtered out.

Provide all the texts that received low scores on all the relevant categories as feedback to the background category, which assumes that the overwhelming majority of these texts belong to the background category “Accuracy of the filtering system”.

When analyzing the accuracy of a filtering system that makes use of queues, consider the following measurements:

- The percentage of texts that are routed to the correct queue.
- The percentage of texts that are routed to the incorrect queue (that is, they are handled, but they are placed to wrong queue).
- The percentage of texts that entered the queues but should have been filtered out.

These measurements cannot be obtained immediately by using Classification Workbench, because Classification Workbench cannot simulate the queuing mechanism of the application.

### 4.10.8 User interface for application feedback

The user interface for obtaining feedback must not make use of “yes” and “no” questions, such as “Is this answer helpful?”. The user interface must allow the user to select the correct answers and categories (from a list of the top matches provided by the system), which is much more informative. It is often easier to learn from this kind of feedback, and it is much better for analyzing the Classification Module’s accuracy.
4.11 Summary

This chapter presented the typical life cycle of a knowledge base and showed how to optimize its behavior, in the context of a number of common scenarios.

Within Classification Module, the knowledge base plays a crucial role. Even though the decision plan was later added to enhance the manipulation of one or more knowledge bases, using rule-based mechanisms, only the knowledge base has the ability to adapt dynamically to the content that it receives and to learn new user needs (through feedback). However, as a result of this powerful aspect of knowledge base behavior, it is essential that you closely observe and direct its performance. This chapter showed a wide range of concepts; you need to analyze your own unique requirements to apply and adapt the techniques that are described in the use cases.
Integration details

In this part, we describe the integration between IBM Classification Module and other products and solutions, including IBM FileNet P8 systems and IBM Content Collector for both E-mail and File Systems. We describe the steps to enable the integrations and step-by-step integration and configuration instructions with use case scenarios.
Integration with IBM FileNet P8

This chapter describes the integration of IBM Classification Module (Classification Module) V8.6 with the Enterprise Content Management (ECM) repository product IBM FileNet P8 V4.0.

Starting from a use case, the chapter guides you through logical tasks to address all the use case requirements with detailed step-by-step instructions to configure and run this integration.

This chapter contains the following major topics:

- Solution introduction
- Integration steps overview
- Enabling the integration
- Training Classification Module with existing IBM FileNet P8 content
- Configuring and performing classification
- Reviewing and fine-tuning classification
5.1 Solution introduction

IBM FileNet P8 offers enterprise scalable repositories that provide companies with the right combination of content management and business process management to activate content, transform processes, and grow through time.

Integrating Classification Module with IBM FileNet P8 repository brings out the value of auto-classifying content and archiving it into IBM FileNet P8 repositories or declaring mission critical content as records under the control of the IBM FileNet Records Manager system.

In this section, we describe the following topics:

- Integration architecture
- Use case

5.1.1 Integration architecture

Classification Module provides automatic classification services for IBM FileNet P8. The integration for IBM FileNet P8 requires the following components:

- Classification Center
- Content Extractor

**Classification Center**

The Classification Center is a Web application that provides the following functions:

- Configure classification settings (filters folders and documents to be classified).
- Run automatic classification of chosen content.
- Monitor classification activity and errors.
- Review documents according to various filter settings (such as those documents that did not exceed the configured confidence threshold for an automatic classification action), and, if necessary, reclassify documents.

**Content Extractor**

This is a command-line tool that extracts content from an IBM FileNet P8 repository to create and train knowledge bases.

We explain more about their configuration and usage in the later sections of this chapter.
Figure 5-1 illustrates the Classification Module and IBM FileNet P8 integration architecture with two major workflows: Setup and Production.

**Setup workflow**

The setup workflow consists of the following steps (the numbers match the numbers in Figure 5-1).

1. **Step 1: Extract data from the repository**: Content Extractor extracts the documents from the IBM FileNet P8 repository to one or more XML files. The attachments referenced in the extracted documents are filed in a *bin* subfolder. We discuss detailed information about configuring and using Content Extractor in the next section of this chapter.

2. **Step 2: Training and analysis in Classification Workbench**: Use Classification Workbench to import extracted files (in XML format) together with their attachments and use them to build knowledge bases. A decision plan is created.
**Production workflow**

The production workflow consists of the following steps (the numbers match the numbers in Figure 5-1 on page 179).

Step 3: The knowledge base and decision plan are deployed to the Classification Module server, which will perform the classification activities.

Classification and review setup with the Classification Center: Configure the Classification Center to classify the document stored in the IBM FileNet P8 object store. It passes the document content to the IBM Classification server, where each document is evaluated by the decision plan. According to the decision plan, the system processes each document (either moves or files the document) based on the category classified by one or more knowledge bases. Certain documents might receive no categories or actions and can remain unclassified. These documents are highlighted for review.

You can use the Classification Center to review documents that match specific conditions (a filter). For example, you can review documents that do not exceed a preset confidence threshold for an automatic classification action.

**Clarification:** The threshold is processed by the decision plan rules. Thresholds are not processed in the Classification Center filters. However, you can go to the filters to view the documents that do not exceed a certain threshold.

Step 4: Classification Module automatic classification: If necessary, Classification Module reclassifies these documents.

Using the Content Extractor, the metadata of classified documents can be exported to XML to evaluate the knowledge base accuracy. This analysis is based on the categories that are suggested by the decision plan and those categories that are chosen by the user in the review process. This data can be imported into Classification Workbench where you can create a large range of reports.

**5.1.2 Use case**

To provide easy-to-understand practical information about what it takes to integrate Classification Module with IBM FileNet P8, we use a case study approach to provide a use case integration scenario and step-by-step procedures for the integration.
Our sample company must classify content in an Enterprise Content Management system and ensure that document retention and disposition policies are enforced.

The present enterprise corporate repository is built on IBM FileNet P8. Initially, the content in the repository is not well organized, and it does not comply with records management policies. To address this issue, the IT specialist responsible for the repository wants to ensure that all of the documents in the IBM FileNet P8 content store are organized into a consistent set of folders (or document classes). The business analyst wants to ensure that data in the repository is organized according to a corporate taxonomy. To achieve this goal, the analyst defines a new corporate taxonomy for assigning document properties and classifying content into folders. The analyst also defines records management policies (rules for the retention and disposition of documents) per the company’s records manager specialist’s instruction.

The IT specialist’s task is to organize content in the company repository by using the new taxonomy policies. The IT specialist works closely with the business analyst to reclassify content that already exists in the repository by applying this new taxonomy. They must work together to eventually ensure that the correct documents are declared as records so that the records can be managed according to the records management policies.

To automate this task, they decide to use Classification Module. They need to configure a set of rules that uses knowledge bases and assign each document to one or more categories in the corporate taxonomy. The rules need to be easy to configure but powerful enough to allow documents to be classified on the basis of both metadata (document properties) and content. During the reclassification phase, information about the documents, such as the document properties and the target folder or document class, might require updating, and certain documents might need to be declared as records. The IT specialist extracts sample content from IBM FileNet P8 for training and testing purposes, and then, the rules can be tested with part of the extracted content to see how the system arrives at its classification decisions.

The business analyst plans to use Classification Module to review documents. If the analyst disagrees with a classification decision, the analyst can reclassify a document by applying other classification criteria. By reviewing documents and either confirming the classification decision or reclassifying the content, the analyst helps to fine-tune the system and improve accuracy over time.

This use case integration solution requires the following software stack:

- IBM FileNet P8 (Either IBM FileNet Content Manager or IBM FileNet Business Process Manager)

- Classification Module

Chapter 5. Integration with IBM FileNet P8
5.2 Integration steps overview

Follow these summary steps to enable, configure, and use Classification Module with IBM FileNet P8:

1. Enabling the integration:
   a. Installing integration component for IBM FileNet P8
   b. Configuring IBM FileNet P8 for integration:
      i. Importing the AddOn or icm_prps_addon.xml file
      ii. Installing the AddOn file
      iii. Add Classification Module properties to each document class
   c. Configuring connectivity between Classification Module and IBM FileNet P8

2. Training Classification Module with existing IBM FileNet P8 content:
   a. Using Content Extractor
   b. Creating a knowledge base
   c. Creating a decision plan
   d. Analyzing a decision plan and its related knowledge base
   e. Exporting knowledge bases and decision plans from Classification Workbench
   f. Deploying the knowledge base and decision plan using Management Console

3. Configuring and performing classification:
   a. Configuring Classification Center
   b. Verifying IBM FileNet P8 structures
   c. Performing classification

4. Reviewing and fine-tuning classification:
   a. Classification Center review function overview
   b. Providing feedback to the knowledge base
   c. Configuring the Classification Center for review
   d. Reclassifying a document
5.3 Enabling the integration

The Classification Module and IBM FileNet P8 integration requires the integration component for IBM FileNet P8. This section assumes that you have a fully operational IBM FileNet P8 V4.0 system running and that you have an administrator ID and password.

5.3.1 Installing integration component for IBM FileNet P8

To enable the integration between Classification Module and IBM FileNet P8, you must install the Classification Module V8.6 integration component for IBM FileNet P8. If you did not install it during the Classification Module installation, install it now before continuing.

To check if you have the integration component installed, from your server, go to Start → Programs → IBM Classification Module 8.6. If you have the Classification Center menu, as shown in Figure 5-2, you have the integration component installed and you can skip the remainder of this section.

![Classification Center Integration Components menu](Image)

Figure 5-2   Classification Center Integration Components menu

If the Classification Center menu is not shown, launch the Classification Module installation wizard as described in the IBM InfoSphere Classification Module Version 8.7 Information Center:

http://publib.boulder.ibm.com/infocenter/classify/v8r7/

Look for the chapter “Deploying on IBM FileNet P8”, section “Getting Started with IBM FileNet P8 integration”, and paragraph “Installing the integration components for IBM FileNet P8”.
During the installation procedure, you will see the window in Figure 5-3, where you select the **Custom** radio button, and then, click **Next**.

![IBM Classification Module 8.6 - InstallShield Wizard](image)

*Figure 5-3  Basic or Custom installation*

When you see the dialog that is shown in Figure 5-4 on page 185, select **IBM FileNet P8 Integration**.
Complete the remainder of the installation. At the end, after rebooting, verify that the integration components are installed, as explained at the beginning of this section.

5.3.2 Configuring IBM FileNet P8 for integration

To work with IBM FileNet P8 content (in Setup or Production mode), the FileNet P8 object store, folder structure, and document classes must match exactly the category names in the proposed knowledge base. Therefore, the first stage of the knowledge base design is to organize the IBM FileNet P8 content within the folder structure or, alternatively, to assign document classes.

In our use case, we use IBM FileNet P8 folder structure to automatically classify documents. Figure 5-5 on page 186 shows the snapshot of the IBM FileNet P8 Workplace client browsing the Content_OS object store, where the folder structure is created under the icm_integration folder.
The Classification Module knowledge base structure that we use to classify IBM FileNet P8’s unclassified content must match the existing folder structure. As a reference for now, Figure 5-6 shows a Classification Workbench - Knowledge Base Editor window, with the knowledge base category structure that will be created in 5.4.2, “Creating a knowledge base” on page 215 of this chapter.

Classification Module classifies documents based on categories that are defined in the knowledge base. If these categories relate to FileNet P8 folders or FileNet P8 document classes to which documents are to be assigned in the FileNet P8 repository, these folders and document classes must already exist in FileNet P8. Classification Module server and the Classification Center are designed to not create these folders and classes automatically.

To configure IBM FileNet P8 for the integration:

1. Importing the AddOn or icm_prps_addon.xml file.
2. Installing the AddOn file.
3. Add Classification Module properties to each document class.

**Importing the AddOn or icm_prps_addon.xml file**
You use IBM FileNet Enterprise Manager, an administration application for IBM FileNet P8, to import, install, and add properties that are required by Classification Module. You must add these properties to all of the document classes that are involved with Classification Module automatic classification.

Before you begin, ensure that IBM FileNet P8 is running.

To configure IBM FileNet P8 parameters, follow these steps:

1. From the Classification Module installation directory, copy the Classification Module properties file to any directory of the IBM FileNet P8 system server, for example, to the \Temp directory. In our environment, we copied the C:\IBM\ClassificationModule\ECMTools\icm_prps_addon.xml file to the IBM FileNet P8 system \Temp directory.

2. In the IBM FileNet P8 system, start the IBM FileNet Enterprise Manager:
   a. Select it from the Start menu, as shown in Figure 5-7.

   ![Figure 5-7 Start IBM FileNet Enterprise Manager](image)

   b. Select the IBM FileNet P8 server to integrate with Classification Module, and click **Connect**. See Figure 5-8 on page 188.
Figure 5-8  Select IBM FileNet P8 server to integrate with Classification Module

3. Right-click the **AddOn** folder (see Figure 5-9) in the right pane, and select **New → AddOn**.

Figure 5-9  Create a new AddOn

4. The window that is shown in Figure 5-10 on page 189 opens. Type a name for the new AddOn, for example, Classification Module. Then, click **Browse** to find the directory where the file `icm_prps_addon.xml` is copied (in our example, the C:\Temp directory).
5. Select the `icm_prps_addon.xml` file, and click **Open**.
6. Verify that the correct file path and file name show in the Import File field, as shown in Figure 5-12 on page 191. Then, click **OK**.
Figure 5-12 Verify the correct import file name and path for the new AddOn

Figure 5-13 on page 192 shows the result of the AddOns Import operation.
Installing the AddOn file

After importing the AddOn file, you install the AddOn file in the IBM FileNet P8 V4.0 object store that will be used with Classification Module:

1. In IBM FileNet Enterprise Manager, right-click the object store of your choice. In our example, we installed it in the **Content_OS** object store, as shown in Figure 5-14 on page 193. Then, select **All Tasks → Install AddOn.**
2. The Add On Installer window appears, as shown in Figure 5-15 on page 194. In the Select AddOns to Install section, select the **IBM Classification Module AddOn** check box, and then, click **Install**.
3. The AddOn Installation Status window appears confirming the successful installation, as shown in Figure 5-16. Click OK and the window disappears.

4. To see the imported properties in the object store, from the menu bar, select Action → Refresh, as shown in Figure 5-17 on page 195.
Add Classification Module properties to each document class

After the AddOn is installed, add the new Classification Module properties to each document class that is involved in the Classification Module automatic classification:

1. Using IBM FileNet Enterprise Manager, right-click **Document Class** (the Document base class for that object store), and click Add **Properties to class**, as shown in Figure 5-18 on page 196.
2. The Document Class properties window that is shown in Figure 5-19 on page 197 appears. Select the **Property Definitions** tab, and click **Add/Remove**.
3. The window in Figure 5-20 on page 198 showing the list of the available properties appears. Scroll down the list of the available properties to find properties with the ICM_ prefix. These properties were added in the previous step 4. Select them all from the left pane and add them to the right pane by clicking **Propagate** (Figure 5-19), and then, click **OK**.
4. In the Document Class Properties window, click **Apply**, and then **OK**.

Repeat the previous steps for each document class.

### 5.3.3 Configuring connectivity between Classification Module and IBM FileNet P8

After the configuration of IBM FileNet P8 is completed, we need to configure and set up Classification Module to allow it to connect to the repository. To set up the integration between Classification Module and IBM FileNet P8, we need to update several configuration files and copy several files from IBM FileNet P8 to Classification Module.

To set up connectivity between Classification Module and IBM FileNet P8, follow these steps:

1. From the Classification Module `ICM_home/ECMTools/lib` directory, copy the following files; if they already exist in the directory, overwrite or rename them before copying:
   - `FileNet_home/ContentEngine/lib/Jace.jar`
   - `FileNet_home/ContentEngine/lib2/javaapi.jar`
We used these files in our installation:
- C:\Program Files\FileNet\ContentEngine\lib\Jace.jar
- C:\Program Files\FileNet\ContentEngine\lib2\javaapi.jar

2. If Classification Module and IBM FileNet P8 are on separate servers, copy the entire FileNet_home/ContentEngine/wsi directory from the IBM FileNet P8 server to the Classification Module server, in any place on the hard disk. In our example, this step is not required, because we installed the two products on the same server.

3. Edit the ICM_home/ECMTools/setupCommandLine executable file and change the value of the WASP_HOME property to the path where the wsi directory resides, or to the path where you copied it in the previous step. In our example, Figure 5-21 shows where the file is located.

4. Edit the following executable files and change, if needed, the value of the P8_VERSION property to 4.0, as shown in Figure 5-22 on page 200. There are .bat files for Windows and .sh files for UNIX operating systems:
- ICM_home/ECMTools/startConnectTest
- ICM_home/ECMTools/startClassificationCenter
- ICM_home/ECMTools/stopClassificationCenter
- ICM_home/ECMTools/startExtractor
We used these files in our example:
- C:\IBM\ClassificationModule\ECMTools\startConnectTest.bat
- C:\IBM\ClassificationModule\ECMTools\startClassificationCenter
- C:\IBM\ClassificationModule\ECMTools\stopClassificationCenter
- C:\IBM\ClassificationModule\ECMTools\startExtractor

```
startConnectTest.bat - Notepad
Set PB_VERSION=4.0
Setlocal
Call setupCommandLine.bat
Set CLASSPATH=%CLASSPATH%;lib/connectTest.jar
@Rem USER SHOULD SET THE VERSION, PROTOCOL AND PLATFORM OPTIONS.
UserID, PASSWORD AND CONFIG WILL BE SET ON THE COMMAND LINE
@Rem EX:  startConnectTest -user myUserName -password myPassword
        -config conf/WcmApiConfig40.properties
If %PB_VERSION%%==3.5 Goto PB_VERSION3.5
If %PB_VERSION%%==4.0 Goto PB_VERSION4.0

:PB_VERSION3.5
..\jre\bin\java %JAVA_OPTIONS% ConnectTest.ConnectTest %*
    -version 3.5 -protocol SOAP -platform WAS
Goto END

:PB_VERSION4.0
..\jre\bin\java -Xmx128M %JAVA_OPTIONS% ConnectTest.ConnectTest %*
    -version 4.0 -protocol SOAP -platform WAS
Goto END

:END
```

*Figure 5-22  The startConnectTest.bat file*

5. Edit the ICM_home/ECMTools/conf/wcmConfig.properties file to change the value of the wcmConfig property to WcmApiConfig40.properties.
   
   In our example, Figure 5-23 on page 201 shows where the file located.
6. If necessary, change the value of the \texttt{wcmConfig} property to \texttt{WcmApiConfig40.properties}, as shown in Figure 5-24.

7. Edit the \textit{ICM\_home/ECMTools/conf/WcmApiConfig40.properties} file to identify the host name or IP address and the port for the IBM FileNet P8 server. See the example in Figure 5-25 on page 202.
8. Verify that Classification Module can connect to the IBM FileNet P8 server:
   a. Start the IBM FileNet P8 server if it is not running already.
   b. Open a command-line window. Go to the $ICM_home/ECMTools$ directory, which is $C:\IBM\ClassificationModule\ECMTools$ in our example.
   c. Locate the $startConnectTest.bat$ file. In order to verify the connection with the IBM FileNet P8 system, you must launch this file with the following parameters:
      - IBM FileNet administrator ID and password
      - Configuration file
   We launched this file in In our example:
   
   ```
   startConnectTest.bat -user administrator -password filenet
   -config conf\WcmApiConfig40.properties
   ```
   d. The test procedure runs for a while. At the end, you see a message on the command-line window similar to the message that is shown in Figure 5-26 on page 203.
If the test connection fails, follow this step:

- For detailed information about the test results, open the html file that was created by the startConnectTest.bat file, that is, the ConnectTest.html file. In our installation, the ConnectTest.html file is located in the C:\IBM\ClassificationModule\ECMTools directory. The file appears similar to Figure 5-27 on page 204 when the test is successful. Ignore the warnings about possible .jar file server version mismatches.

- Verify that all of the previous steps have executed correctly.
5.4 Training Classification Module with existing IBM FileNet P8 content

Classification Module works with a knowledge base that reflects the type of data that the system is expected to handle.

Before the knowledge base can analyze texts, it must be trained with a sufficient number of sample texts (collectively called a content set) that are properly classified into categories. One way to provide a content set for knowledge base
training is to use existing documents that are stored in the FileNet P8 repository. These documents are extracted using the Content Extractor command-line tool.

If documents are already organized in folders and document classes in the FileNet P8 object store, you can use the same structure and exact folder/document class names to create the initial knowledge base categories.

### 5.4.1 Using Content Extractor

Content Extractor is a command-line tool that extracts data from the IBM FileNet P8 repository and generates XML output based on the document metadata. The documents are extracted according to options that you specify in a properties file. For example, you specify the source IBM FileNet P8 object store and which folders, subfolders, and document classes contain content to be extracted. You can also limit which documents are extracted by specifying constraints, such as extracting only documents that have certain document properties or that were modified after a specific date.

To create a knowledge base, import the XML output and extracted files into Classification Workbench and configure categories that reflect the structure of folders and document classes in IBM FileNet P8. After you associate the knowledge base with a decision plan, use the Classification Center to run statistical and rule-based classification and automatically classify the documents into IBM FileNet P8 folders or document classes.

When you run Content Extractor, the content of any attachment of the processed documents is exported to a binary output directory that is created under the XML output directory. The extracted documents are given unique file names upon extraction, for use by the Classification Module import function to Classification Workbench (and do not reflect the names that are used by the content stored in IBM FileNet P8).

#### Setting the Content Extractor properties

You use the Content Extractor tool to extract a set of sample documents from FileNet P8 object store.

Before you begin, ensure that IBM FileNet P8 is running.

Configure Content Extractor by editing the **Extractor.properties** file, which is located in `IbmClassificationModule_installation_path\ECMTools\conf`. In our example, the file path is `C:\IBM\ClassificationModule\ECMTools\conf`, as shown in Figure 5-28 on page 206.
Edit the properties file and set, at a minimum, the following properties before you run Content Extractor:

- If you extract content from an IBM FileNet P8 Version 4.0 server, ensure that the `filenetWcmApiConfigPath` property is set to `conf/WcmApiConfig40.properties` or to the name of the `WcmApiConfig.properties` file that you are using, according to the setup in 5.3.3, “Configuring connectivity between Classification Module and IBM FileNet P8” on page 198. Refer to Figure 5-29 on page 207.
Ensure that the XML output directory that you specify in the XmlDirectory property exists and is empty. The default value is extractorOutput. Ensure also that the binaryOutputDirectory folder exists under the XML output directory and that it, too, is empty. The default value is binaryOutput as you can see in Figure 5-29. The default location for these two directories, which were created during the Classification Module installation, is C:\IBM\ClassificationModule\ECMTools, as shown in Figure 5-30 on page 208.
Ensure that the Path_n properties specify the paths of IBM FileNet P8 folders or document classes from which you want to extract content. The format is Path_x = objectStoreName, which means to extract everything from this object store, or Path_x = objectStoreName/FolderName/FolderName, which means to extract from this folder and its subfolders. In our example, we extracted documents from the review subfolder, which is contained in the icm_integration folder resident in the Content_OS object store. Therefore, the property is Path_1 - Content_OS/icm_integration/review, as shown in Figure 5-31. You can use more than one Path_x parameter for a single extraction operation.
The Path_n parameters are mandatory for running Content Extractor. You can configure additional parameters to extract documents from IBM FileNet P8, according to various filters. Refer to the Extractor.properties file or the product documentation for further details.

- **With_x = key[=value]**: Positive Document constraint: Extract the document only after you check whether the key exists, for example:
  - With_1 = DocumentTitle = restaurant_wine_list.doc
    Extracts documents with specific title.
  - With_2 = DocumentClass= Document
    Extracts all documents with document class equal to Document.
  - With_4 = DocumentBaseClass= BaseDocument
    Extracts all documents with BaseDocument class and all its subclasses.

- **Date**: Date constraint: Extract the document only if it was modified after this date; use this format for dates: *dd-MMM-yyyy*, for example:
  - Date = 13-Jul-2008
    Extract only documents modified after 13 July 2008.

- **FolderMax**: Maximum number of documents to extract from each folder; for example:
  - FolderMax = 10

- **FolderMin**: Minimum number of documents to extract from each folder; for example:
  - FolderMax = 1

**Testing the Content Extractor**

Run Content Extractor on the command line to extract documents from an IBM FileNet P8 object store.

To run Content Extractor:

1. Open a command-line window.
2. Move to the directory where the content extractor is located, that is, *IbmClassificationModule_installation_path\ECMTools*. In our example installation, it is *C:\IBM\ClassificationModule\ECMTools*.
3. Launch the command **startExtractor.bat** without any parameters. In this way, you test if the extractor is correctly configured. If everything is set correctly and you input the correct command, the help window is returned, as shown in Figure 5-32 on page 210.
Running the Content Extractor

Beginning with Classification Module Version 8.6, you can use the following options to obtain information about the IBM FileNet P8 object store without extracting any documents:

- **-m folders**
  
  Returns a list of IBM FileNet P8 folders for the specified object store

- **-m docclasses**
  
  Returns a list of IBM FileNet P8 document classes for the specified object store

- **-m metadata**
  
  Returns a list of available metadata properties for the specified object store

In our example installation, to know the document classes that are available in the object store previously configured in the C:\IBM\ClassificationModule\ECMTools\conf\Extractor.properties file, launch the command:

C:\IBM\ClassificationModule\ECMTools>startExtractor.bat -u Administrator -p filenet -f C:\IBM\ClassificationModule\ECMTools\conf\Extractor.properties -m docclasses

In this example, Administrator is the filenet system administration ID, and filenet is the password.
Figure 5-33 shows the Content Extractor console output for the previous command.

![Console Output]

Figure 5-33   Output of running Content Extractor with the -m parameter

When the -m parameter is used, the Content Extractor provides the output in the ExtractorOutput.txt text file, which is located in the directory C:\IBM\ClassificationModule\ECMTools. In our example installation, it is C:\IBM\ClassificationModule\ECMTools, as shown in Figure 5-34 on page 212.
To easily read the file, we suggest that you use WordPad. The file will appear, similar to Figure 5-35, listing all of the document classes that are available in the object store that was previously set in the Extractor.properties file.
To extract documents from IBM FileNet P8 repository:

1. Open a command-line window.

2. Move to the directory where the Content Extractor is located, that is, \_IbmClassificationModule_installation_pathECMTools. In our example installation, it is C:\IBM\ClassificationModule\ECMTools.

3. Launch the command:

   startExtractor.bat -u IBM_Filenet_P8_user_ID -p IBM_Filenet_P8_user_password -f Extractor.properties_file_with_complete_path

   In our example installation, the command is:

   C:\IBM\ClassificationModule\ECMTools>startExtractor.bat -u Administrator -p filenet -f C:\IBM\ClassificationModule\ECMTools\conf\Extractor.properties

   At the completion of the document extraction activity, the command-line console appears similar to Figure 5-36.

   ![Figure 5-36  After content extraction completes](image)

   The extracted documents’ metadata is stored in an XML file that is located in the C:\IBM\ClassificationModule\ECMTools\extractorOutput directory, as shown in Figure 5-37 on page 214.
The binary document contents are written in a separate file in the C:\IBM\ClassificationModule\ECMTools\extractorOutput\binaryOutput directory, as shown in Figure 5-38.
5.4.2 Creating a knowledge base

You create a knowledge base by using the Classification Workbench to analyze a collection of sample documents and to gather statistics.

Classification Module uses knowledge bases and decision plans to determine how to classify content. The knowledge base contains a set of categories that correspond to folders or document classes in IBM FileNet P8. Training the knowledge base with real-world examples enables the Classification Module to classify similar documents into the IBM FileNet P8 repository.

**Note:** If the planned classification deals with separate document classes, you must create separate knowledge bases, one for each document class.

Category names in the knowledge base must follow this convention:

- For classification into folders: `object_store_name/folder_path`, for example: `ObjectStore1/ParentFolder/SubFolder`
- For classification into document classes: `object_store_name/document_class_name`, for example: `ObjectStore1/DocumentClass`

In our user scenario, Classification Module works with a single document class. The documents are extracted from one IBM FileNet P8 folder, and the content set is used to build a knowledge base within Classification Workbench.

After you make the document set available to train a new knowledge base, you can import the content set into Classification Workbench and work with it to create a knowledge base.

5.4.3 Creating a decision plan

You create a decision plan by using Classification Workbench.

In this chapter, we do not explain how to create a knowledge base, which is explained in Chapter 3, “Working with knowledge bases and decision plans” on page 59. However, in this section, we provide the step-by-step procedures to build the decision plan to address the following use case:

- Work with IBM FileNet P8 documents stored in the `Content_OS` object class.
- Classify enterprise documents that are stored in the `Bulk` folder. The documents relate to three separate branches of the company, and several of them also relate to financial subjects.
Perform the classification using:

a. The statistical analysis provided by a knowledge base previously created, analyzed, and tested using Classification Workbench

b. The rules embedded in the decision plan that is created later in this section

A document with a score provided from the statistical analysis that is equal to or higher than 80% for one or more categories will be classified in the folders associated with those categories in the IBM FileNet P8 repository.

Documents with scores provided from the statistical analysis below 80% for all categories will be classified in the review folder for subsequent manual review.

Before starting this procedure, you need to have an available knowledge base. There are use cases where Classification Module and IBM FileNet P8 integration through the Classification Center works with only a decision plan, without the need of a knowledge base. However, in those cases, the classification is performed relying only on keyword rules, and the document classification will not benefit from the statistical analysis of the document content and attributes.

**Start the Classification Workbench**

To launch the Classification Workbench:

1. From the Windows server where Classification Module is installed, go to **Start → Programs → IBM Classification Module 8.6 → Classification Workbench**, as shown in Figure 5-39 on page 217.
2. The Workflow Assistant window appears in Figure 5-40. From the right pane, select **Create a decision plan project**.

3. The window in Figure 5-41 on page 218 appears.
Create a decision plan
The decision plan that we will create consists of two separate rules. The first rule is to classify documents in IBM FileNet P8 specific folders when the statistical analysis for the document content matches one or more categories with a score equal to or higher than 80%. The second rule is to classify documents in the IBM FileNet P8 review folder when the statistical analysis for the document content does not match any category with a score equal to or higher than 80%.

Create the first decision plan rule
To create the first decision plan (classify documents with scores greater than or equal to (>=) 80%):

1. Enter a decision plan name in the Name field. In our example, we use branches. Select Decision Plan, and click Next.

![Figure 5-41 Create a new project (decision plan or knowledge base)](image.png)

2. In the next window, as shown in Figure 5-42 on page 219, select Create an empty project, and click Finish.
3. The window in Figure 5-43 appears. Right-click the Referenced Projects tree in the lower left pane and add the project for the knowledge base that was created in 5.4.2, “Creating a knowledge base” on page 215. In our example, we select select_ranch.

4. Now, click New Group, go to the upper-right pane, right-click anywhere, and select New Rule.
5. The New Rule window appears as in Figure 5-44. Type a descriptive rule name in the Name field. Leave the When triggered field set to **Continue**, leave the **Enabled** check box selected, and then, click the **Trigger** tab.

![Figure 5-44 New Rule Properties tab](image1)

6. Referring to Figure 5-45, click the **condition** link, verify that the random field specifies 100% (which indicates that the rule will always be evaluated), click the check mark to apply the trigger, and then, click the **Actions** tab.

![Figure 5-45 Set the trigger for the new rule](image2)

7. The window in Figure 5-46 on page 221 appears. Click **Add actions**.
8. The window in Figure 5-47 on page 223 appears. Choose a scenario for your classification in IBM FileNet P8 from this list of available scenarios and their explanations:

- File the document in an IBM FileNet P8 folder.
  This scenario copies the document to one or more IBM FileNet P8 folders.
- Move the document into an IBM FileNet P8 folder.
  This scenario copies the document to one or more IBM FileNet P8 folders and removes (unfiles) the document from its source folder.
- Unfile the document from an IBM FileNet P8 folder.
  This scenario removes (unfiles) the document from one or all folders. The value of the content field contains a folder name or an asterisk (*), indicating that the document is removed from all folders. Note that documents are not deleted from the repository.
- Set the document’s IBM FileNet P8 document class.
  This scenario assigns a document class to documents.
- Declare the document as a record in IBM FileNet Records Manager.
  This scenario declares documents as records in IBM FileNet Records Manager.
- Set a metadata field for the document in IBM FileNet P8.
  This scenario sets a metadata field for the document in IBM FileNet P8. If the specified metadata field does not exist in the IBM FileNet P8 repository, it is added.
– Set a document content field.

This scenario sets the value of the specified content field to a value that is copied from another content field or the specified string, counter, or temporary variable. If you enter a new content field name, the content field is added to the content item.

In our use case, we perform the following actions:

1. From the Available scenarios pane, we select the action to File the document to an IBM FileNet P8 Folder.

2. In the Scenario settings pane:
   a. We select the File document into one or more folders, based on category scores in the selected knowledge base radio button.
   b. From the knowledge base name combination box, select the knowledge base that you have previously added as a referenced project. In our example, we use select_branch.
   c. You can select one of the three available options to file the document:
      - Top Category only: File the document in a single IBM FileNet P8 folder associated with the top score category.
      - All categories whose score is above this percentage: File the document in one or multiple IBM FileNet P8 folders, depending on the assigned categories, their relative scores, and the threshold that you set in the combination box.
      - All categories whose score is above the defined threshold in this file - File name: This option refers to comma-separated value (csv) files that were produced when setting thresholds during knowledge base tuning, when you work with Precision versus Recall values. For example, you might generate several threshold files based on different precision values, and then write a rule that refers to the different threshold files based on specific trigger conditions. Threshold files that are associated with the knowledge base are included when you add a knowledge base to a decision plan. Associated threshold files are listed under the knowledge base names in the Referenced Projects panel of the Decision Plan window. You can select or clear the check box next to each threshold file to enable or disable it in the decision plan. For details about this topic, refer to the Classification Module V8.6 product Information Center, chapter Configuring Classification - Knowledge Base Editor - Setting Thresholds:


In our example, we select the second choice, with a threshold of 80%.
3. After you complete the Add Action configuration, as shown in Figure 5-47, click **OK**. The New Rule window appears, which is shown in Figure 5-48 on page 224, with a summary of the configured actions. Make sure that the actions are correct, and click **OK**.
4. The Decision Plan window returns, as shown in Figure 5-49 on page 225, with all of the information of the new rule that was just configured.
Create the second decision plan rule

To create the second decision plan rule (if less than (<) 80%, go to the review folder):

1. Go to the upper-right pane of the decision plan, right-click anywhere and select **New Rule**.

2. The New Rule window appears. In the Properties tab, enter the name of the rule. In our example, we enter **review**, as shown in Figure 5-50 on page 226, leave the other default values as is, and click the **Trigger** tab.
3. In the Define a trigger for the rule window, open the combination box and select the \textit{less than} (\textless) symbol, as shown in Figure 5-51 on page 227.
4. Click the first instance of the word **number**, and click the option **specify match result**. In the next window that appears (see Figure 5-52 on page 228), select the knowledge base from the referenced project, which in our example is **select_branch**, and then, click **OK**.
Figure 5-52 Define Review rule based on score result from the select_branch knowledge base

5. Click the second instance of the word number, click the option constant number, and type the threshold value in the field. In our example, it is 0.8, which means a score of 80%. Click the check box to confirm. The final result of the above trigger setting is Trigger: $select_branch__scores[1] < 0.8 (see Figure 5-53 on page 229). Therefore, each document with a top category score (assigned by the knowledge base) of less than 80% will trigger the action that we will write in the following steps.
6. Click the **Action** tab, and the window in Figure 5-54 appears. Click **Add actions**.

**Figure 5-53** Define if score is less than 80% for Review rule

**Figure 5-54** Add action to the Review rule
7. In the Add Action window, as shown in Figure 5-55 on page 231, from the Available scenarios pane, select the action to **File the document to an IBM FileNet P8 Folder**.

8. In the Scenario settings pane:
   a. Select the radio button **File document into a specific in IBM FileNet P8**.
   b. In the Folder Name field, type the name of the IBM FileNet P8 folder where we want to file the documents with a top category score of less than 80%. In our example, this folder is `Content_OS/icm_integration/review`.
   
   **Note:** To ensure a correct classification from Classification Module into the IBM FileNet P8 repository, you must the complete path for the FileNet P8 folder name:

   `<object_store_name>/folder_name/subfolder_name_if_any>`

   c. In the Preview Decision Plan Actions pane in Figure 5-55 on page 231, you can see the action syntax, which is `add_to_content_field 'P8:File' 'Content_OS/icm_integration/review'` in our example.
   
   d. Click **OK**.
The New Rule window appears, as shown in Figure 5-56 on page 232, with a summary of the configured actions. Verify that the actions are correct, and click **OK**.
10. The main Decision Plan window returns again, as shown in Figure 5-57 on page 233, with both of the rules configured.
11. Save your decision plan (our example is named branches), as shown in Figure 5-58 on page 234. From the Classification Workbench's main menu, click **Project → Save**.
5.4.4 Analyzing a decision plan and its related knowledge base

Now, we have an available decision plan and the related knowledge base, which are named \texttt{branches} and \texttt{select\_branch} respectively in our example. Before we start to use them in the production environment by deploying them on the Classification Module server, it is extremely important to understand how they will perform, so we need to analyze them in the Classification Workbench.

To analyze the decision plan and the related knowledge base, we need perform the following steps:

1. Import a test content set into the decision plan project in Classification Workbench.
2. Assign the content type value to one or more content fields.
3. Analyze a decision plan and related knowledge base.
4. View and interpret analysis reports.

For the detailed steps to perform these operations, refer to 3.2.2, “Building a knowledge base” on page 62 and 3.2.3, “Analyzing and learning with a knowledge base” on page 76.
5.4.5 Exporting knowledge bases and decision plans from Classification Workbench

After creating the decision plan and knowledge bases, export them in order to make them available to be deployed in the Classification Module server.

Exporting knowledge bases
To export a knowledge base:

1. Launch Classification Workbench.
2. Select **Project → Open**, as shown in Figure 5-59.

![Figure 5-59 Open project in Classification Workbench](image)

3. Select the project for the knowledge base that you want to export, as shown in Figure 5-60 on page 236, and then, click **Open**.
4. After the Knowledge Base Project opens in Classification Workbench, start the Export Wizard by clicking the blue arrow icon that is shown in Figure 5-61. You can also use File \rightarrow Export to open the wizard.

5. The Export Wizard window appears, as shown in Figure 5-62 on page 237. Click Next.
6. Because we need to export the knowledge base to deploy to the Classification Module server, select **Knowledge base**, as shown in Figure 5-63 on page 238, and click **Next**.
7. It is possible to export a knowledge base in various formats, depending on how it will be used. For our purposes, select **Knowledge base version 6.x and later**, as shown in Figure 5-64 on page 239, and click **Next**.

*Figure 5-63  Classification Workbench Export Wizard: Select knowledge base for export*
Chapter 5. Integration with IBM FileNet P8

8. Specify a path and a file name for the exported knowledge base, as shown in Figure 5-65, and click **Finish**. The knowledge base is now exported.
Exporting decision plans
You must follow a similar procedure to export a decision plan:

1. Start Classification Workbench.

2. Select **Project → Open**, as shown in Figure 5-59 on page 235.

3. Select the decision plan project that you want to export, as shown in Figure 5-66, and click **Open**.

4. After the decision plan opens in Classification Workbench, start the Export Wizard by clicking the blue arrow icon, as shown in Figure 5-67 on page 241. You can also use **File → Export** to open the wizard.
5. When the Export Wizard window appears, click **Next** to continue.

6. Because we need to export the decision plan to deploy on the Classification Module server, select **Decision plan**, as shown in Figure 5-68, and click **Next**.

7. It is possible to export a decision plan in various formats, depending on how you will use it. For our purposes, select **Decision plan file (*.dpn)**, as shown in Figure 5-69 on page 242, and click **Next**.
8. Specify a path and a file name for the decision plan that you are exporting, as shown in Figure 5-70, and click **Finish**. The decision plan is now exported.

---

**5.4.6 Deploying the knowledge base and decision plan using Management Console**

To make the decision plan and knowledge base available to a Classification Module server, they have to be deployed to the server through an administrative application called the Management Console.
In our example, the decision plan has a referenced knowledge base. Therefore, in this section, we perform the deployment of the previously created decision plan and a knowledge base that was created by following the instructions that were provided in Chapter 3, “Working with knowledge bases and decision plans” on page 59.

**Deploying a decision plan**

Make sure that the Classification Module server is running. Deploy a decision plan to the Classification Module server:

1. Start the Classification Module Management Console, as shown in Figure 5-71.

![Figure 5-71  Launch Management Console](image)

2. When the window in Figure 5-72 appears, click OK.

![Figure 5-72  URL for the Classification Module listener component](image)

3. The Management Console main window appears, as shown in Figure 5-73. In the left pane, right-click **Decision plans** and select **Add decision plan**.
4. In the Add Decision Plan window, perform the following actions, as shown in the example in Figure 5-74 on page 245:
   a. Enter a name for the decision plan, which in our example is branches.
   b. Select the Import a decision plan radio button.
   c. Navigate to the DP_project_name.dpn file that you previously created with the Classification Workbench. The .dpn default file location is the Classification Workbench project directory, for example: 
      
      ICM_home\Classification Workbench\Projects_Unicode\branches.dpn
   d. In the Servers pane, click the Server combination box and select your Classification Module server, in our example, HQDEMO1.
   e. In the Supported Languages pane, because a decision plan does not really support languages explicitly, you must select a language. Select a language from the right pane, and move it to the right by double-clicking it or by using the move button (>).
   f. Leave all other default values, and click OK.
5. The warning message box in Figure 5-75 on page 246 appears, because the deployed decision plan refers to a knowledge base that is not yet available on the Classification Module server. We deploy this knowledge base in the next steps of this procedure. Click OK for now.
Deploy a knowledge base

Previously, we created the knowledge base select_branch and exported it from Classification Workbench to a directory. We now deploy it to the Classification Module server through the Management Console:

1. Start the Classification Module Management Console: Start → Programs → IBM Classification Module 8.6 → Management Console.

2. Right-click Knowledge bases, and from the menu, select Add knowledge base, as shown in Figure 5-76.

3. In the Add Knowledge Base window, perform the following actions, as shown in Figure 5-78 on page 248:
   a. Type the knowledge base name, which is select_branch in our example.

   Note: Regardless of the knowledge base file name, here, your knowledge base name must match the name used in the decision plan.

   b. Select the Import a knowledge base radio button.
c. Navigate to the \KB_project_name\.kb file that you previously created with the Classification Workbench. The .kb default file location is the Classification Workbench project directory, which is normally located in the folder ICM_home\Classification\Workbench\Projects_Unicode. In our example, the folder is the ICM_home\Classification\Workbench\Projects_Unicode\select_branch\ folder, and the knowledge base file is the select_branch.kb file.

d. In the Servers pane, click the Server combination box and select your Classification Module server, in our example, HQDEM01.

e. In the Supported Languages pane, select one or more languages from the left, and move them to the right by double-clicking the language or by using the move button (>). Your language choices depend on the languages present in the documents that your system expects and the languages that are used to build the (multilingual) knowledge base.

f. Leave all the other default values, and click **OK**.

4. The Management Console window reappears, as shown in Figure 5-77, with only the just deployed knowledge base.

![Figure 5-77](http://localhost:18087)

**Figure 5-77** Deployed knowledge base in Management Console

Figure 5-78 on page 248 shows the detailed information of the deployed knowledge base.
5. Start the decision plan and all related knowledge bases by clicking **Decision plans** in the console tree, right-clicking the decision plan name in the left pane, and selecting **Start decision plan and associated knowledge bases**.
Refer to the example in Figure 5-79. Another way to start the decision plan and all related knowledge bases is to click the green arrow icon.

![IBM Classification Module Management Console](image)

**Figure 5-79  Start decision plan and associated knowledge bases**

### 5.5 Configuring and performing classification

This section explains the configuration of the Classification Center and its use in our example scenario to automatically classify documents.

#### 5.5.1 Classification Center overview

Use the Classification Center to configure the content to be classified, monitor the progress of the classification activity, and review the classification results.

The Classification Center includes three main pages that you use to configure classification options, start and monitor the classification processes, and review classification decisions. See Figure 5-80 on page 250.
Configuration

Clicking Configuration on the main page enables you to launch the configuration tool and specify how Classification Module is to classify documents into folders or document classes in IBM FileNet P8.

General Settings

Specify information about the Classification Module server, the name of the decision plan that you want to use for rule-based classification, and the name of the IBM FileNet P8 object store where the classified content is to be stored. If your decision plan includes rules for declaring documents as records, you can create records from classified documents in a file plan object store and put the documents under the control of IBM FileNet Records Manager.
You can also structure content by specifying how IBM FileNet P8 document properties are to be mapped to fields in Classification Module. By mapping fields, you define the elements that you expect Classification Module to analyze and act upon.

**Content to Classify**
Select the folders or document classes that contain the content to be classified. You can refine the content set by excluding certain folders or including documents only if they meet certain criteria, such as documents that contain specific document properties or document property values, or if the document date falls in a specific time period.

**Runtime Settings**
You can control how the classification processes use system resources by limiting the number of documents to classify, specifying the number of seconds that the server is to wait before looking for new documents to classify, and increasing or decreasing the number of threads that the classification processes use to classify content.

You can also specify options for previewing the content of documents that you review. For example, you can prevent large documents from being displayed in the Classification Center. To be able to open documents for review in IBM FileNet P8, you can provide the URL for the IBM FileNet Workplace XT browser-based application.

**Dashboard tab**
Use the dashboard to start, stop, and monitor the progress of the classification processes. You can view process statistics, such as how many documents were classified and how many documents were flagged to be reviewed. You can also view summary information, such as which folders or document classes received the greatest number of documents and which rules returned the greatest number of matches.

**Event log**
You can view a history of all classification activity, such as when a document was classified and the name of the folder or document class in which it was filed.

**Error log**
If you need to troubleshoot a problem, you can view the error log. While viewing the log, you can select individual messages to see details about the error.

**Review tab**
Review the classification results and, if necessary, reclassify documents.
**Review documents**
Use the options on this page to view source documents and review the actions that were applied by the classification processes. You can either confirm that the document is correctly classified or select other categories and reclassify the document. By reviewing documents, you help to verify that the system is performing as expected and to ensure that the correct folders, document classes, and decision plan actions are applied during classification. In addition, when you review documents and manually select appropriate categories, the system learns from your selection, thereby improving future classification.

**Add document**
You can add new documents to the IBM FileNet P8 repository. When you add a document, Classification Module analyzes its content and suggests how to classify it. You can then review and confirm the actions or reclassify the document just as you reclassify any other document that is available for review.

### 5.5.2 Working with the Classification Center

Before you start the Classification Center:

- Ensure that the Classification Module server is running.
- Ensure that the IBM FileNet P8 server is running.

**Ensure Classification Module server is running**
To verify Classification Module server is running and the needed decision plan and related knowledge bases are started:

1. Launch the Classification Module Management Console by selecting `Program → IBM Classification Module 8.6 → Management Console`.
2. Verify that the server port, as shown in Figure 5-81, is correct for your setup. We keep the defaults and click **OK**.

![Figure 5-81 Verify server port is correct](image)

3. The Management Console main window appears. Click the decision plan that you will use and verify that it is started. Check the related knowledge bases
that you will use and verify that they are started by right-clicking each object. All running objects will show the Start option grayed out, as shown in Figure 5-82.

![IBM Classification Module Management Console](http://localhost:18087)

**Figure 5-82** Verify that the decision plan and knowledge bases are started

**Ensuring that IBM FileNet P8 server is running**

To verify that IBM FileNet P8 server is running:

1. Start an Internet Explorer® session and point to the address: `http://Filenet_hostname_or_ip_address:9080/Workplace`

2. If the login window appears, as shown in Figure 5-83 on page 254, IBM FileNet P8 is up and running. If it is not, you need to start it following a separate procedure, depending on which application server your IBM FileNet P8 installation relies.
You must have read/write permissions for all of the folders in the IBM FileNet P8 object store that are required for classification.

**Start the Classification Center**

To log on to the Classification Center:

1. Start the Classification Center on the server where the Classification Module integration components for IBM FileNet P8 have been installed by clicking `Start → Programs → IBM Classification Module 8.6 → Classification Center → Start Classification Center server.`

2. The Classification Center server takes a while to start. You can see the progress through the command-line window that is shown in Figure 5-84 on page 255. When the server is started, the following message appears:

   **Classification Center is started**

   **Note:** Do not close this command-line window; otherwise, you will not be able to connect to the Classification Center server. If the command line on your machine is in Quick Edit mode, do not touch inside the command-line console or else it will pause the server.
3. After the Classification Center server is started, click **Start → Programs → IBM Classification Module 8.6 → Classification Center → Classification Center** to open the Classification Center in a Web browser, as in Figure 5-85.

4. The login page in Figure 5-86 on page 256 appears. Type your IBM FileNet P8 administrator ID and password, and click **Log on**, as in our example.
The first time that the Classification Center starts, certain configuration options have yet to be set, resulting in the warning in Figure 5-87. We will address setting these configuration options next, so for now, click **Close**.
5.5.3 Configuring Classification Center

In this section, we perform the steps that are needed to configure the Classification Center to work with our use case, and we also provide examples of additional features that can be used.

Configuration: General setting

When you configure general settings, you specify system-wide options for how Classification Module classifies documents into IBM FileNet P8.

Classification Module settings

These settings specify information about the Classification Module system:

1. In Classification Center, select the Configuration page, and then, select General Settings.

2. Click Edit IBM Classification Module settings as in Figure 5-89 on page 259 to configure the Classification Module URL and the Classification Module decision plan.
3. When the window, which is shown in Figure 5-89 on page 259, appears, type these values:
   - The Classification Module URL, including the port number. In our example, it is http://hqdemo1:18087.
   - The Classification Module decision plan that was previously deployed. In our example, it is branches.

4. Click **Save**.
Figure 5-89  Classification Module configuration settings

IBM FileNet P8 settings

These settings specify information about the repositories to use for classified content. If content is classified into an object store with its record metadata in a file plan object store, the object store and file plan object store must be linked in the same IBM FileNet P8 server instance:

1. In the Classification Center, select the Configuration page, and then, select General Settings.

2. Scroll down the Classification Center - General Settings page to reach the section that is shown in Figure 5-90 on page 260, and click Edit IBM FileNet P8 Settings.
3. When the window that is shown in Figure 5-91 on page 262 appears, enter these values:
   - The IBM FileNet P8 object store that contains the content to classify and where to store the content after it is classified.
   - In the case of the decision plan that was previously selected, include rules for declaring documents as records:
     - Select the **Declare documents as records** check box.
     - Specify the name of the IBM FileNet Records Manager object store where records are to be created for documents that are declared as records when they are classified.

4. Click **Save**.
Field mapping

Classification Module analyzes document fields that are designated for natural language processing. To enable the classification processes to better classify documents, you need to map IBM FileNet P8 document properties to Classification Module fields.

The list of available IBM FileNet P8 document properties is determined by the object store that contains the content to be classified. To be able to view a list of the available document properties, ensure that the IBM FileNet P8 server is started.

Also, ensure that the Classification Module fields that you want to use for classifying content and to which you want to map document properties are defined in the Management Console application (field definitions).

If important information about the documents that you want to classify is stored in document property fields in IBM FileNet P8, you can map these document properties to Classification Module fields so that the information can be evaluated as part of the classification process. For example, if documents in your IBM FileNet P8 repository have a text field that contains comments about the document, you might want to evaluate this metadata in addition to the actual document content when the documents are classified.

**Note:** At least one mapping is required.
To map document properties to fields so that they can be used for classification:

1. In the Management Console, check which field definitions are available. A large number of default fields are available. Two common Classification Module fields are Document (data type “Document”), Title (data type “Text”), and FileName (data type “Text”). For the IBM FileNet P8 integration, at least one field of data type “Document” is required. If not, add the field, as shown in Figure 5-91.

![IBM Classification Module Management Console](image)

**Figure 5-91  Add fields if they do not already exist**

2. In the Management Console, restart all knowledge bases if you added new field definitions.

3. In the Classification Center, select **Configuration** → **General Settings** → **Edit Field Mappings**, as shown in Figure 5-92 on page 263.
4. The window in Figure 5-93 appears. You can see that the default field mapping has already been done by the system. In order to be used in the classification process, each IBM FileNet P8 document property needs to be mapped to an appropriate Classification Module field. If you need additional field mapping, click **Add Document Property**, and then, click **Browse**.
5. In the Document Properties window, expand the list of available document classes, select the class that contains the document property that you want to map, and then, select the document property.

6. From the list of available Classification Module fields, select the field to which you want to map the document property, and click **Save**. When the classification processes run, the content of the document property will be evaluated when determining which categories a document matches best and which decision plan actions to trigger.

**Configuration: Content to Classify**
When you configure the content to classify in the Classification Center, you specify the source of the documents that require classification. All content to be classified must be in the same IBM FileNet P8 object store.

**Classification filters**
When you configure the content to classify, you specify filters to include or exclude documents from the classification process.

Although you can specify the content that you want to classify by typing your preferences in fields, you can browse and select folders, classes, properties, and so on, only if the IBM FileNet P8 server is running.

To specify a classification filter:

1. In the Classification Center, select the **Configuration** page, and then, select **Content to Classify**. When the window in Figure 5-94 on page 265 appears, click **Edit Classification Filters**.
2. On the Content to Classify page that is shown in Figure 5-95 on page 266, expand the filter that you want to configure. You have two options:

- Enter the folder name into the field.
- Click **Browse**, connect to the IBM FileNet P8 object store, and select the folder that you need.
3. In our example, we have selected only the Start Folder and the Document class, choosing from the list obtained from the IBM FileNet P8 system, as shown in Figure 5-96 on page 267. After making your choices, click **Save**.
Chapter 5. Integration with IBM FileNet P8

267

Figure 5-96  Choose the folder and the document class

Configuration: Runtime Settings

Runtime settings control how the classification processes use system resources. You can also specify options for viewing source documents when you review documents in the Classification Center, such as the maximum size of documents that can be displayed. Two settings can be configured from this page:

- Runtime settings for classifying content
- Runtime settings for reviewing documents

Runtime settings for classifying content

When you configure the Classification Center, you can specify options to control how the classification processes use system resources:

1. In the Classification Center, select the Configuration page, and then, select Runtime Settings. The window in Figure 5-97 on page 268 appears. As you can see, the classification setting can be modified, but in our example, we maintain the default values. For explanations of each field, refer to the product documentation.
Configure options to control system behavior for classifying and reviewing documents. Learn more...

**Classification Settings**

You can specify options to control how the classification processes use system resources. Learn more...

- Maximum number of items to process: ALL
- Seconds to wait for new items: 5
- Number of threads: 1

**Review Settings**

You can specify options for viewing source documents and viewing the classification results in IBM FileNet P8.

- Amount of text to show when previewing document content: 128
- Maximum size of documents that can be viewed during review: 1000000

IBM FileNet P8 Workplace XT URL: http://hqdemo1:9080/Workplace/

**Figure 5-97** Edit runtime settings for classifying content
**Runtime settings for reviewing documents**

When you configure the Classification Center, you can specify options for how documents are to be displayed when you review classification decisions.

We modify the review setting in order to review documents after classification:

1. Referring to Figure 5-97 on page 268, click **Edit Review Settings**. The window in Figure 5-98 appears. Modify the default value [localhost]:7001 with the host name of your IBM FileNet P8 workplace values, and click **Save**.

![Figure 5-98 Configure runtime setting for reviewing documents](image)

2. When you are returned to the Classification Module - Configuration tab - Runtime Settings pane, click the **Dashboard** tab.

**5.5.4 Verifying IBM FileNet P8 structures**

Before we start classifying documents, you must perform a final verification on IBM FileNet P8 to check that all the required document classes, folders, and subfolders that Classification Module used to classify documents are available.

**Note:** If any document class, folder, or subfolder that is required by Classification Module to classify documents is missing in IBM FileNet P8, a specific error is shown in the error log pane of the dashboard, and the document assigned to that document class, folder, or subfolder will not be classified.

It is beyond the scope of this book to explain how to find the specific document classes and folders in IBM FileNet P8. To learn about IBM FileNet P8, refer to the product manuals and IBM Redbooks publications.
Remember that category names in the knowledge base must follow this convention:

- For classification into folders: `object_store_name/folder_path`, for example: `ObjectStore1/ParentFolder/SubFolder`
- For classification into document classes: `object_store_name/document_class_name`, for example: `ObjectStore1/DocumentClass`

To gather the document classes and folders that are needed in IBM FileNet P8:

1. Launch Classification Workbench.
2. Using the Classification Workbench Workflow Assistant, click **Open an existing project**.
3. Open the decision plan that you have configured in 5.4.3, “Creating a decision plan” on page 215 in the Classification Center, in our example, branches. As shown in Figure 5-99, select the decision plan, and then, click **Open**.

![Open Project](image)

*Figure 5-99  Open decision plan*

4. Look at the lower-left pane to see the knowledge bases that are referenced projects from the decision plan; in our use case, there is just one, `select_branch`. Double-click `select_branch` to open the knowledge base, as shown in Figure 5-100 on page 271.
5. Click the **Category** tab located in the left pane, and then, enlarge the pane to see the complete category name that must comply with the rule explained at the beginning of this section. See the example in Figure 5-101 on page 272.
If your results show that document classes or folders are missing, you have to refer to the IBM FileNet P8 administrator to address the problem, or make the Knowledge base match the IBM FileNet P8 document classes or folder structure.

**Note:** Decision plans can contain rules that move or file documents to specific IBM FileNet P8 document classes, folders, or subfolders. If you are not the person who created the decision plan, check with the author of the decision plan to understand if these types of actions are present in the decision plan. Alternatively, you can check by going through all the rules of the decision plan, looking for actions, such as the action that we put in the **branches** decision plan, in the **review** rule:

- **Rule Name:** review
- **Rule Status:** Enabled
- **When Triggered:** Continue
- **Trigger:** $select_branch__scores[1] < 0.8
- **Actions:**
  
  [on] add_to_content_field '$P8:File'
  'Content_OS/icm_integration/review'

**Figure 5-101** View categories for the knowledge base named select_branch
5.5.5 Performing classification

In the Classification Center, you can start the classification processes from the Configuration tab or from the Dashboard tab. It is more convenient to start the classification processes from the Dashboard tab, because the classification progress is displayed on the Dashboard tab.

Start classification

The Classification Module server and the IBM FileNet P8 server must be running in order to classify documents. The decision plan that is used to classify content and its associated knowledge bases must also be running. For detailed instructions, go to 5.5.2, “Working with the Classification Center” on page 252.

To run the classification processes:

1. If the Classification Center is not running on the Dashboard tab, follow these steps; otherwise, go to step 2:
   a. Start the Classification Center server by selecting Start → Programs → IBM Classification Module 8.6 → Classification Center → Start Classification Center server.
   b. Wait for the server to start, and then, select Start → Programs → IBM Classification Module 8.6 → Classification Center → Classification Center.
   c. Click the Dashboard tab, and when required, enter your IBM FileNet P8 administrator ID and password.

2. Start the classification by clicking Start classifying, as shown in Figure 5-102 on page 274.
3. The pane changes, as shown in Figure 5-103 on page 275. Notice the *Elapsed time* counter is running.
Chapter 5. Integration with IBM FileNet P8

Use the dashboard to view the progress of the classification activity or to open the event and error logs.

The classifier can be left running with the status "Waiting for documents", or it can be stopped by clicking Stop classifying.

Monitoring the classification process
Use the dashboard in the Classification Center to start the classification processes, to monitor the progress of the classification processes, and to stop them.

The dashboard provides the following information to help you monitor the progress and success of the classification processes. From the dashboard, you can also see details about all classification activity or details about classification errors by opening the event log or error log.

Event and error logs
You can open log files to learn about classification activities and errors that have occurred, by opening the Dashboard tab in the Classification Center.
5.6 Reviewing and fine-tuning classification

When reviewing documents in the Classification Center, you can confirm the classification actions or select other categories and actions for reclassifying the content.

5.6.1 Classification Center review function overview

When classification runs, the classification process automatically moves the majority of documents into the correct destination folders and document classes in IBM FileNet P8. Certain documents, such as documents that do not exceed the configured threshold for automatic classification, are flagged for manual review.

You can view source documents and review the actions that were taken to classify a document. You can either confirm that the document is correctly classified or reclassify the document by applying other categories or decision plan actions.

By reviewing documents, you help verify that the system performs as expected and ensure that the correct folders, document classes, and decision plan actions are applied during classification. In addition, when you review documents and manually select appropriate categories and actions, feedback is sent to the knowledge base through which the system learns, thus improving subsequent classification.

5.6.2 Providing feedback to the knowledge base

Classification Module learns from the feedback that is provided by your acceptance or modification of how a document is classified when you review documents in the Classification Center.

By reviewing documents in the Classification Center and specifying that other categories or decision plan actions are more appropriate for classifying a document, Classification Module modifies the statistical models in the knowledge base. In this way, the knowledge base is maintained and adapts to subtle changes in categories over time. There is no need to manually tune the system to account for changes in content. For example, the words that people use change over time, and learning from feedback ensures that documents are classified accurately despite changes in terminology.

Although you can review any document that was classified, you might want to configure a rule to set aside content into one or more specific folders. You can
then configure the document review filter in the Classification Center to review only the documents in those folders.

### 5.6.3 Configuring the Classification Center for review

When you review documents in the Classification Center, you can specify a filter to limit which documents you review. For example, you might want to review only those documents that contain document properties with values that match words that are relevant to a potential lawsuit, or only those documents that are flagged for review.

When you configure a filter for reviewing documents, the selections that you make are similar to the selections that you make when you configure the content to classify. The difference is that instead of selecting the documents to classify, you select the documents that are presented for review.

For example, you can select folders or document classes that contain the documents that you want to review, or you can specify that you want to exclude documents in specific folders from the review set. You can also limit the documents that you review according to the document dates or the existence of specific document properties or document property values.

To specify a document filter to limit which documents you review:

1. Start the Classification Center and click the **Review** icon on the page. See Figure 5-104 on page 278.
2. On the Review Documents page, click the **Filter Settings** button, as shown in Figure 5-105 on page 279.
3. Scroll down the window to find the **Edit Filter Settings** option, as shown in Figure 5-106 on page 280, and click it.
4. The Document Filter page appears, as shown in Figure 5-107 on page 282. Expand the filter that you want to configure, and specify criteria for the documents that you want to review. You can specify and configure one or more of the following criteria:

- **Start folders:** Specify any number of folders that contain documents that are to be reviewed. The review set will include documents in the folder that you specify and documents in subfolders of that folder. Click **Browse** to select a folder from the list of folders in the object store. Click **Add Folder** to include additional folders in the review set.

- **Skip folders:** Specify any number of folders that contain documents that are not to be reviewed. The review set will exclude documents in the folder that you specify and documents in subfolders of that folder. Click **Browse** to select a folder from the list of folders in the object store. Click **Add Folder** to exclude additional folders from the review set.
Document classes: Specify any number of document classes that contain documents to review. Click Browse to select a document class from the list of classes in the object store. Click Add Document Class to include additional document classes in the review set.

Document properties: Limit which documents to review by specifying that only those documents that have specific document properties are included in the review set. Click Browse to select the document class to which the property belongs, and then, select the document property. Click Add Property to specify additional document properties.

Document property values: Limit which documents to review by specifying that only those documents which have document properties with specific values are included in the review set. Click Browse to select the document class to which the property belongs, select the document property, and then, type the value that the property must contain. Click Add Property to specify additional document properties and values.

Date: Limit which documents to review by specifying that only documents with dates that match a date or date range are included in the review set. This value is the date that the document was last modified. You can specify whether the document date must occur before or after a specific date or specify that the document date must occur in a specific date range.

Classification status: Use a document's classification history to control which documents are included in the review set:

- Do not use the classification status to filter documents: To ignore the classification status when, for example, significant changes were made to the decision plan or a knowledge base after documents were first classified. This option ensures that all documents that match the other document filter criteria are presented for review.

- Include only documents that were previously classified by IBM Classification Module: To review documents that were previously classified if the decision plan or knowledge base significantly changed, and you want to ensure that the new classification actions are being applied correctly.

- Include only documents that were not previously classified by IBM Classification Module: To review only newly classified documents if the decision plan or knowledge base did not significantly change. You can limit the review set to documents that were added to the IBM FileNet P8 object store and classified for the first time.
In our example installation, we configured the start folder, as shown in Figure 5-107. The icm_integration/review folder is used to store documents with a classification score under the selected threshold.

5. You can select the **Save as the default filter** check box if the current document filtering options are to be automatically applied when you review documents in future Classification Center sessions. The default filter remains in effect until you save other options as a new default filter. If you change the document filter settings without saving them as the default filter, the settings remain in effect only for the duration of the current Classification Center session.

6. Click **Save** on the right side of the window to save your settings.
5.6.4 Reclassifying a document

When you review classification results in the Classification Center, you can select other categories and decision plan actions and classify the document again.

Documents that are marked for review are shown on the Review Documents page one at a time. After you review the classification decisions that were applied by Classification Module, you might decide to reclassify the document by specifying other categories. During reclassification, the decision plan evaluates rules according to the categories that you suggest and then suggests appropriate actions in return. You can apply the newly suggested actions or submit the document for reclassification again.

When you reclassify a document and manually select appropriate categories and actions, the system learns from your selection and uses that knowledge to improve future classification actions. You can also improve classification by selecting an option to save the document in XML format when you review the suggested reclassification actions. Later, you can import the saved XML data into Classification Workbench to use it to tune the decision plan. For details and guidance about how to set up this option, refer to 3.4.4, “Working with offline (deferred) feedback” on page 108.

To reclassify a document:

1. Start the Classification Center and click the Review icon.

2. On the left navigator pane, there are two options, as shown in Figure 5-108 on page 284:
   - Review Documents
   - Add Document

   Be sure you are on the Review Document page.

3. From this page, which is shown in Figure 5-108 on page 284, you can perform these tasks:
   - Navigate through the entire set of documents under review using the Previous and Next arrow buttons, maintaining the document in the review document set.
   - Ratify the classification assigned by the system using the actions that were applied when this document was classified, and remove the document from the review set, using the Confirm option.
   - Reclassify the document using the Reclassify option. For each knowledge base that is associated with the decision plan, the Reclassify window shows the top five classifications that Classification Module determined to be most appropriate for the document. The active decision plan is the
decision plan that was previously configured in 5.5.3, “Configuring Classification Center” on page 257.

– Launch IBM FileNet P8 Workplace XT using the appropriate option.
– Be aware of the amount of documents in the review set under revision.
– Look at and modify the Filter Setting through the appropriate button.
– Obtain information about the document under review, document type, and title. More information is available by clicking **Show details**.
– Look at the document content by clicking **View full document**.

*Figure 5-108   Review documents*

In our example installation, we use the branches decision plan, where a rule is designed to put all the documents with a top category score under 80% in the review folder (for details, see “Create the second decision plan rule” on page 225). Now, we review this document set and want to submit the document again to the Classification Module to use the feedback already received by the knowledge base in order to have a more accurate classification. It might be useful to look at the document content to understand it, using the **View full document** option.

4. In the Classification Module Review tab window that is shown in Figure 5-108, scroll the window down to find the Decision History section and look at the score previously assigned for each category, as shown in Figure 5-109 on page 285.
5. Click the **Reclassify** icon at the top of the window, and the window that is shown in Figure 5-110 appears.

6. You can reclassify a document using the category that is suggested by Classification Module, or make your own choice if you think it is more appropriate for an accurate classification. In our example, we selected the
Reclassify by selecting specific categories option. Here are several additional options that you can select:

- Open the combination box, as shown in Figure 5-110 on page 285, to choose a category from the top five categories.
- Click the points button (…) to choose a category out of the top five categories.
- Click the plus button (+) to add more categories to be assigned to the document, as shown in Figure 5-111.

![Figure 5-111 Reclassify document to additional categories](image)

7. After you click Submit, the Classification Rules and Actions window appears, as shown in Figure 5-112 on page 287. Review the rules that were triggered by the classification processes, the actions that the decision plan suggests, and the top scoring categories from each knowledge base associated with the decision plan.

8. If you want to save the document in XML format so that you can import the data into Classification Workbench and use it to tune the decision plan, click Save Document in XML Format. When you are prompted to open or save the file, select the option to save the file to disk.
Chapter 5. Integration with IBM FileNet P8

9. Click **Apply Actions**. The document is classified according to the selected actions and removed from the review set.

10. The document is classified in FileNet P8 as you have decided, and the next document in the queue in the review folder is presented to be reviewed; this process happens until there are no documents left in the queue to review.

**Adding documents to IBM FileNet P8**

You can add documents to the IBM FileNet P8 object store. When you add a document, Classification Module analyzes its content and suggests how to classify it.

When you upload a document, you can review and accept the suggested classification actions, or reclassify the document by specifying other categories and classification actions.

To classify a document that you add to IBM FileNet P8:

1. Start the Classification Center and click the **Review** icon on the page.

2. On the left navigation bar, there are two options, as shown in Figure 5-108 on page 284:
   - Review Documents
   - Add Document

   Click **Add Document**.
3. When the window in Figure 5-113 appears, click **Add Document**.

![Add Document](image)

*Figure 5-113  Add a document to IBM FileNet P8 for classification analysis*

4. On the Add Document window that is shown in Figure 5-114, click **Browse** to select the file that you want to upload or type its network path, and click **OK**.

![Add Document Window](image)

*Figure 5-114  Browse to the document location to add it for classification analysis*

5. Classification Module processes the document, and the decision history is displayed, as shown in Figure 5-115 on page 289. From the decision history, you can review the rules that were triggered for the document by the classification processes, the actions that the decision plan suggests, and the top scoring categories from each knowledge base associated with the decision plan. On this window, you can apply the classification actions that you want and clear the check boxes for actions that you do not want. Then, click **Apply Actions**, and the document is added to the IBM FileNet P8 folder or document class that the decision plan recommends.
6. If the suggested classification actions are not wanted, click **Reclassify** in order to specify other actions or categories for classifying the document. The window that is shown in Figure 5-110 on page 285 appears.

7. Follow the same steps described for reclassifying documents (step 7 to step 10 on page 286).

8. In the event that the classification of the document is no longer required, you can click **Discard Document**. The classification actions are discarded, and the document is not added to the IBM FileNet P8 object store.
Integration with IBM Content Collector for File Systems

In this chapter, we introduce the key aspects of the integration of IBM Content Collector for File Systems with IBM Classification Module (Classification Module) and the typical business use cases where the integration asset is used. In addition, we describe the steps to enable the integration and provide step-by-step integration and configuration instructions for a legal discovery and compliance use case.

We cover the following topics in the chapter:

- Solution introduction
- Integration steps overview
- Enabling the integration
- Validating the integration
- Configuring the integration system
- Performing the file system archiving
6.1 Solution introduction

IBM Content Collector is an enterprise content management (ECM)-based content collection and archiving solution. It is based on a modular, extensible architecture that enables organizations to take back control and unlock the business value of content, while enforcing compliance and operational policies.

IBM Content Collector working together with Classification Module brings out the value of auto-classifying content (for example, e-mails and file share content) and archiving the content into ECM repositories or declaring mission critical content as records under the control of a records management system.

In this section, we describe the following topics:
- Overview of IBM Content Collector for File Systems
- Integration architecture
- Use case description

6.1.1 Overview of IBM Content Collector for File Systems

Within the suite of IBM Content Collector products, IBM Content Collector for File Systems 2.1 is an enterprise solution for file system archiving and management, providing tools that help users to comply with corporate and regulatory policies. It uses rules and policies to monitor and analyze documents on network file shares in order to enforce document and records management policies.

You can create Content Collector task routes to automatically archive content on network drive shares to IBM ECM systems, such as IBM FileNet P8 or IBM Content Manager. Content Collector can also work with IBM FileNet Records Manager to automatically identify documents and declare them as records.

6.1.2 Integration architecture

IBM Content Collector for File Systems uses external characteristics (such as document name, source location, and custom metadata) to determine how to process a document. This information is sufficient for many migration and business process scenarios. However, when documents in a file share are not organized and only selected documents need to be captured into an ECM repository - as part of a legal discovery, for example, it is necessary to analyze the contents of documents to determine whether to capture the documents. Similarly, if a document is to be captured, the content of the document, such as target location, security, and record declaration, needs to be used to determine how to process the document. These capabilities are provided by Classification
Module, an application that enables Content Collector for File Systems to look inside documents to understand and classify their contents for use in a Content Collector task route.

Figure 6-1 shows the architecture overview of the integration between Content Collector for File Systems and Classification Module.

Within this integration, Classification Module uses a knowledge base to analyze documents to discover possible categories and their relevancy scores. You can use the top relevancy score and the top category name in IBM Content Collector task routes for any of the following typical scenarios:

- Assign the folder path for the archived documents in an IBM FileNet P8 or IBM Content Manager (CM8) repository.
- Identify documents that need to be declared as records.
- Identify documents that must be reviewed in the Classification Center to enhance the knowledge base accuracy (applicable to the IBM FileNet P8 repository type only when the Classification Module and FileNet P8 integration asset is installed and configured).
6.1.3 Use case description

To provide easy-to-understand practical information about what it takes to integrate Classification Module with Content Collector, we use a case study approach to provide a use case integration scenario and step-by-step procedures for the integration.

Our sample company needs to gather human resources documents from various file shares for an impending lawsuit. The company needs the relevant documents to be added to the repository based on the existing company taxonomy and to declare them as records for later review by company lawyers. In the meantime, the irrelevant documents need to be left out of the repository to reduce overall storage costs and the legal review costs.

This use case integration solution requires the following software stack:

- IBM Content Collector for File Systems
- Classification Module
- IBM FileNet Records Manager
- IBM FileNet P8 (either IBM FileNet Content Manager or IBM FileNet Business Process Manager)

The key concepts and information provided here are relevant for Content Collector for File Systems and Classification Module working with IBM Records Manager and IBM Content Manager Version 8 as well.

In order to implement this use case, a knowledge base is trained to identify human resources documents. In a Content Collector task route, as the Classification Module task analyzes each document passed to it, Classification Module generates a relevancy score (a rating between 0 and 1 of how closely a document matches the knowledge base criteria) and passes this value back to Content Collector for further actions. The Content Collector task route in our example is configured to perform the following actions:

- If the relevancy score for the analyzed document is greater than 90%, Content Collector for File Systems will file it in a suggested folder in an IBM FileNet P8 repository and declare it as a record under the management of IBM FileNet Records Manager. These documents will be available for the legal review.

- If the relevancy score is between 30% and 90%, Content Collector for File Systems will capture the document in an IBM FileNet P8 folder for further investigation.

- If the relevancy score is not greater than 30%, Content Collector for File Systems will not capture the document.
6.2 Integration steps overview

The following summary steps enable, configure, and use the Content Collector and Classification Module integrated system:

1. Enable the integration:
   a. Install the Classification Module client module.
   b. Register Classification Module with Content Collector as a utility connector.

2. Validate the integration between Content Collector and Classification Module.

3. Configure the integrated system:
   a. Define a knowledge base and its field definitions.
   b. Create a file system archiving task route with rules and records declaration.
   c. Activate the system for archiving.

4. Perform the file system archiving operation.

6.3 Enabling the integration

You need to enable the Content Collector and Classification Module integration through the following two tasks before the Classification Module task will be available to be placed in a Content Collector task route:

1. Install Classification Module client components on the Content Collector server.

2. Register Classification Module with Content Collector as a utility connector.

6.3.1 Installing Classification Module client components on the Content Collector server

This example assumes that you have installed the following software successfully:

- IBM Content Collector for File Systems
- Classification Module server components
- IBM FileNet Content Manager or IBM FileNet Business Process Manager
- IBM FileNet Records Manager
For IBM Content Collector to use the classification capabilities of Classification Module, you must install Classification Module client components on the Content Collector server by using the following procedure:

1. Run the Classification Module installation program on the IBM Content Collector server.
2. Select the option to install **Custom** components.
3. Select the **Classification Module Client only** check box, as shown in Figure 6-2, and then, finish the rest of the installation steps.

![Installing Classification Module client components on the Content Collector server]

**Figure 6-2** Installing Classification Module client components on the Content Collector server

### 6.3.2 Registering Classification Module with Content Collector

To use Classification Module as a utility connector, Content Collector for File Systems requires access to the following three Classification Module .dll files, which are located in the `/bin` directory of the Classification Module installation path:

- bnsClient86.dll
- Package01123.dll
- stlport_ban46.dll
Follow this process of registering Classification Module .dll files:

1. Copy the three .dll files to the ctms directory under the Content Collector for File Systems installation path (for our case study, it is the C:\Program Files\IBM\ContentCollector\ctms directory), and rename the bnsClient86.dll file to the bnsClient85.dll file.

2. Open a command window, and from the ctms directory, run the utility connector in registration mode by entering the following command:
   ```
   utilityconnector.exe -r
   ```
   If the command returns an error message saying, “Failed to create service: ibm.ctms.utilityconnector.UtilityConnector -- The specified service already exists”, most likely you registered the .dll files in the wrong location. To correct the operation, perform the following steps:
   a. Use the command below to first unregister the Classification Module .dll files in the location where you issued the initial registering command:
      ```
      utilityconnector.exe -u
      ```
   b. Ensure that you change to the ctms directory, and register the .dll files again using the following command:
      ```
      utilityconnector.exe -r
      ```

### 6.4 Validating the integration

To validate the success of the integration of Content Collector for File System with Classification Module, make sure that the following attributes are set up properly in the system:

- Windows registry entry of ibm.ctms.utilityconnector.ICMClassificationTask
- Classification Module task type in Content Collector Configuration Manager
- Classification Module system metadata properties in Content Collector Configuration Manager

#### 6.4.1 Windows registry entry

Confirm that the ibm.ctms.utilityconnector.ICMClassificationTask Windows registry entry is available by using the following steps:

1. Launch the Windows Registry Editor by issuing the following command in the command prompt:
   ```
   regedit
   ```
2. In the Registry Editor window, navigate to **My Computer** → **HKEY_LOCAL_MACHINE** → **SOFTWARE** → **IBM** → **EMRC** → **4.0** → **Service Declarations2** → **ibm.ctms.utilityconnector.UtilityConnector** and validate that the registry entry of **ibm.ctms.utilityconnector.ICMClassificationTask** is listed, as shown in Figure 6-3.

![Registry Editor](image)

Figure 6-3  Validating that the Windows registry entry of **ibm.ctms.utilityconnector.ICMClassificationTask** exists

### 6.4.2 Classification Module task in Content Collector Configuration Manager

Verify that the Classification Module task type is available in the Content Collector Configuration Manager:

1. Launch the Content Collector Configuration Manager by clicking **Start** → **Programs** → **IBM Content Collector** → **Configuration Manager**.

2. In the Content Collector Configuration Manager:
   b. In the Toolbox pane, ensure that the IBM Classification Module task type is listed under the Utility tasks, as shown in Figure 6-4 on page 299.
6.4.3 System metadata

The integration introduces five Classification Module system metadata properties that can be used in Content Collector task routes. See Table 6-1.

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Relevant Categories</td>
<td>List of top categories matched</td>
</tr>
<tr>
<td>All Relevant Categories and Scores</td>
<td>Combined list of categories and scores</td>
</tr>
<tr>
<td>All Relevant Scores</td>
<td>List of top scores</td>
</tr>
<tr>
<td>Most Relevant Category</td>
<td>Winning category</td>
</tr>
<tr>
<td>Most Relevant Score</td>
<td>Winning category score</td>
</tr>
</tbody>
</table>

The Classification Module task produces the metadata values during the run time. In turn, you can use them in Content Collector task routes to determine how documents are processed.
Use the following steps to ensure that the Classification Module system metadata properties are available in the Content Collector Configuration Manager:

1. Launch the Content Collector Configuration Manager by clicking Start → Programs → IBM Content Collector → Configuration Manager.

2. In the Content Collector Configuration Manager:
   a. Go to the Metadata and Lists box by clicking Metadata and Lists in the Navigation pane.
   b. In the Metadata and Lists box on the left pane, select the System Metadata type, and then, select the IBM Classification Manager system metadata in the middle System Metadata pane. Ensure that five Classification Module system metadata properties are displayed on the right pane. See Figure 6-5.

![Figure 6-5  Classification Module system metadata properties in Content Collector Configuration Manager](image)

### 6.5 Configuring the integration system

After you enable the integration between Content Collector for File Systems and Classification Module, you are required to configure the integration system.
before using it in file system archiving solutions. The following list summarizes
the configuration process to use Content Collector for File Systems and
Classification Module together:

1. Define a knowledge base and Classification Module field definitions.
2. Configure a file system archiving task route, including a Classification Module
task.
3. Activate the system for archiving.

This section introduces the configuration process in light of the HR legal
discovery and compliance use case for our sample company as described in

6.5.1 Defining a knowledge base and the Classification Module field
definitions

A knowledge base provides the criteria that Content Collector for File Systems
uses to determine if a document needs to be captured and, if so, how it is
processed. In addition, the Classification Module server receives documents as a
series of fields. The field definition defines the data type and the method of
language processing that the classification server performs on the field.

The following steps are the typical workflow to define a knowledge base and
configure Classification Module field definitions:

1. Create, analyze, and tune a knowledge base in Classification Workbench.
2. Add a knowledge base to the Classification Module server in Management
   Console.
3. Configure field definitions in Management Console.

Creating, analyzing, and tuning a knowledge base
Several options are available in Classification Workbench to create a knowledge
base:

- Using a categorized content set
  We recommend that you build a knowledge base from a categorized content
  set in Classification Workbench for this task.

- Using an un-categorized or partially categorized content set
  If you do not have a fully categorized content set, you can use the Taxonomy
  Proposer tool to help you with the categorization process, and then, you can
  import the categorized content set into Classification Workbench to build a
  knowledge base.
Using initialization data

When a content set is not available, you can use initialization data to build the knowledge base. *Initialization data* consists of keywords and texts associated with individual categories. *Keywords* are words or phrases that you expect will appear in documents classified by the classification server. In addition to keywords, you can associate one or more texts with categories. The Classification Module server uses this information to classify incoming documents appropriately.

There are many ways to prepare a categorized content set for building a knowledge base. In our example, we first identify a set of representative documents that are pertinent to the application environment, and we organize the documents in a file system folder structure. The category names that we choose for our knowledge base are used as the folder names. We then import this folder structure into Classification Workbench to create, analyze, and tune the knowledge base. We describe the detailed steps in Chapter 3, “Working with knowledge bases and decision plans” on page 59.

**Adding a knowledge base to the Classification Module server**

After a well-defined knowledge base is ready, perform the following steps to add it to the Classification Module server:

1. Launch the Management Console by clicking **Start → Programs → IBM Classification Module 8.6 → Management Console**.

2. In the console tree, click **Knowledge bases**, as shown in Figure 6-6.

![Figure 6-6 Selecting the Add knowledge base menu option](image-url)
3. The Add Knowledge Base window, which is shown in Figure 6-7 on page 304, shows how to define the following fields to add a knowledge base to the Classification Module server:

- Specify the knowledge base name. In our example, it is HR.
- Import a knowledge base that contains statistics by selecting the **Import statistics from file** radio button. Instruct the Classification Module server to read the knowledge base file that you want to import by selecting the **Access file from server** option and browsing to the location of the knowledge base file.
- Specify runtime options for the knowledge base:
  - Specify whether the knowledge base is to use a cache. A *cache* enables the system to handle knowledge bases that are too large to load into memory in one piece. In our example, because the HR knowledge base is within a reasonable size, we do not select the **Use a cache** option.
  - The “Back up automatically” option automatically creates backup copies of the knowledge base when you make changes to it, such as importing knowledge base statistics, changing feedback options, and adding or removing a read-only instance. Having a backup is useful, for example, if you need to reproduce results from the previous version after the knowledge base is changed. In our example, we select **Back up automatically**.
  - Specify how feedback is to be processed. Because inaccurate feedback downgrades the knowledge base accuracy, in most business environments, feedback is required to be reviewed and applied by a knowledgeable user at a later time. So in our example, we select the **Defer processing** option.
- Specify the servers and ports for running read/write and read-only instances of the knowledge base. In our example, we specify the current server to run both read/write and read-only instances.
- Specify the languages that the knowledge base is required to support. In our example, our knowledge base supports **English**.
Configuring field definitions

Classification Module processes documents as a set of *field definitions* in the Management Console. In order to enable the Classification Module server to use the proper natural language processing method for analyzing and classifying documents, perform the following steps to define the proper content type and data type of the required fields:

1. Launch the Management Console by clicking **Start → Programs → IBM Classification Module 8.6 → Management Console.**
2. In the console tree, click **Field definitions.**
Within the Content Collector and Classification Module integration, in order to have the Classification Module server process the content of a document, you have to use a field with the data type of Document. We use the predefined Document field, as shown in Figure 6-8, to build the Content Collector task route.

![Figure 6-8 Field definitions](image)

6.5.2 Building a file system archiving task route

This section shows you how to build a file system archiving task route that supports the legal discovery and compliance use case as described in 6.1.3, “Use case description” on page 294.

To build a file system archiving task route, we create a task route as shown in Figure 6-9 on page 306 for automatically classifying and archiving file share files into appropriate folders in the IBM FileNet P8 repository and for declaring critical documents as records.
Follow these steps to build the file system archiving task route:

1. Launch the Content Collector Configuration Manager by clicking **Start → Programs → IBM Content Collector → Configuration Manager.**
2. Go to the Task Routes explore pane and toolbox by clicking **Task Routes** in the Navigation pane of the Configuration Manager, as shown in Figure 6-10.

![Task Routes Explore Pane & Toolbox](image)

**Figure 6-10  The Task Routes explore pane and Toolbox of the Content Collector Configuration Manager**

3. In the Task Routes explore pane, click the **Add** icon to add a new task route, as shown in Figure 6-11 on page 308.
4. Content Collector allows you to create new task routes from the beginning or from existing templates. In our example, we show you how to create a simple file system archiving task route from the beginning.

In the New Task Route, we select **Blank task route** as a starting point for the task route creation, and we enter **HR Task Route** as the task route name, as shown in Figure 6-12 on page 309. Click **Create**.
5. A new blank task route with Start and End nodes appears in the Designer pane, as shown in Figure 6-13. Click the Save icon to save your progress.
6. For a new file system task route, a File System Collector is required to define the location from which files will be captured.

Follow these steps to add and configure a FSC Collector in your task route:

a. From the Toolbox pane, under File System Source, select FSC Collector, as shown in Figure 6-14.

b. Click anywhere in the Task Route Designer pane to drop the new FSC Collector task onto the task route, as shown in Figure 6-15 on page 311.
c. After adding the FSC Collector, you can configure it:

- To name and enable the collector on the General tab page.
- To specify when to collect files on the Schedule tab page.
- To specify where to collect files on the Collection Sources tab page.
- To specify files that must not be collected on the Filter tab page.

In our example, we describe the settings of the FSC Collector for each tab in Table 6-2 on page 312.
7. Files are stored as documents in the IBM FileNet P8 content repository. To create a document in the repository, you need to specify where to create the document and how to index it. You specify where to create the document by selecting a repository from a list of the configured repositories. You index the document by choosing a document class for the item and by specifying the values to assign to each property of that class.

Perform the following steps to add and configure a P8 4.x Create Document task in your task route:

a. From the Toolbox pane, under FileNet P8 4.x Repository, select the P8 4.x Create Document task, as shown in Figure 6-16 on page 313.

<table>
<thead>
<tr>
<th>Configuration tab name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter your task route name and its description. Also, ensure that this collector is <strong>Active</strong>.</td>
</tr>
<tr>
<td>Schedule</td>
<td>The collector can run at different frequencies, such as daily, weekly, and monthly. In our example, we set the collector to run <strong>always</strong>.</td>
</tr>
<tr>
<td>Collection Source</td>
<td>Define the file system folders that are monitored by the collector. These folders can be local folders or folders on a shared network drive. In our example, we specify the monitored folder as the C:\HRMonitored folder, and select the <strong>Monitored sub-folders</strong> option, as well.</td>
</tr>
<tr>
<td>Filter</td>
<td>Set the filter so that the collector will ignore certain documents. In our example, we set the filter options to ignore documents that are <strong>already processed</strong>, <strong>already captured</strong>, and <strong>where access is denied</strong>.</td>
</tr>
</tbody>
</table>
b. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the Start and End nodes until the task and the arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle, as shown in Figure 6-17 on page 314.
c. You can configure the P8 4.x Create Document task with the settings that are described in Table 6-3 on page 315.
Table 6-3  Settings of the P8 4.x Create Document task

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connection drop-down list, select the connection to use when creating a document. In our example, we choose a predefined P8 4.x connector connecting to a P8 content object store.</td>
</tr>
<tr>
<td>Check In Options</td>
<td>Set Check In options for your document, such as version and content capture options. These options are required by P8. For more information about each option, refer to P8 Enterprise Manager documentation. In our example, we accept the default settings.</td>
</tr>
<tr>
<td>Property Mappings</td>
<td>Select the document class that you want to use when creating the document, and enter values for the document class properties. In our example, we select the document class called document, and map the Document Title property with the File Name metadata value.</td>
</tr>
</tbody>
</table>

8. After the documents are created, they are passed to the Classification Module task for analysis. The content of each document can be used to determine whether to capture a document, rather than relying solely on the document metadata or source location metadata to make this determination. In addition, for each document, the Classification Module task populates its system metadata properties as described in 6.4.3, “System metadata” on page 299. You can use these metadata properties in the task route to determine how to process the documents.

Perform the following steps to add and configure a Classification Module task in your task route:

a. From the Toolbox pane, under Utility, select the IBM Classification Module task, as shown in Figure 6-18 on page 316.
b. Click anywhere in the Designer pane between the P8 4.x Create Document task and the End node to drop the new task onto the task route, as shown in Figure 6-19 on page 317.
c. Perform the following steps to configure the Classification Module task:

- Specify the host name of the server that hosts the Classification Module server.
- Specify the port of the Classification Module listener component. The default port is 18087.
- Select the knowledge base that you want to use for analyzing the files collected from the monitored file share.
- Select the content field that identifies the part of a document that you want Classification Module to analyze. To process the content of the document, this field must be type document.
- Specify the relevance threshold. In our example, the threshold is set to 0.3, which means that Classification Module only returns suggested categories with relevancy scores greater than 0.3. In other words, if the
relevancy score of the top category is not greater than 30%, Content Collector does not capture this document.

- Specify the maximum number of categories suggested by Classification Module to be used by Content Collector for processing in a task route.

Figure 6-20 shows the settings of the Classification Module task as configured in the HR task route example.

Figure 6-20  Configuring the Classification Module task

To support this legal discovery and compliance use case, the Content Collector task route uses the Most Relevant Score and Most Relevant Category system metadata properties to determine how to process each document. You need to create a Decision Point and two rules branching from the Classification Module task to handle the documents based on the values of the Most Relevant Score and Most Relevant Category system metadata properties.
9. Perform the following steps to add a new Decision Point to your task route:
   a. From the Toolbox pane, select **Decision Point**.
   b. Click anywhere in the Designer pane between the Classification Module task and the End node to drop the new Decision Point onto the task route, as shown in Figure 6-21.

![Figure 6-21 Adding the Decision Point to the task route](image)

10. Configure the arrow connecting the Decision Point and the End node as a rule by taking the following steps:
   a. Click the rule arrow to display the rule Configuration Pane.
   b. This branch of the task route processes the document whose **Most Relevant Score** value of its **Most Relevant Category** is greater than 90%. So, we name this rule **Score > 90%**.
   c. For the evaluation criteria, select the **Configure rule** radio button, and click **Add**.
d. In the Edit Conditional Clause window, configure the rule with the following fields, as shown in Figure 6-22:

- For the Metadata type, select **IBM Classification Module**.
- For the Property, select **Most Relevant Score**.
- For the Operator, select **Greater than**.
- Select the **Literal** radio button with the value of 0.9.

![Edit Conditional Clause window](image)

*Figure 6-22    Add conditional clause*

Figure 6-23 on page 321 shows the rule configured for the left branch of the task route for processing documents whose **Most Relevant Score** value of its **Most Relevant Category** is greater than 90%.
11. In order to browse for the document at a later point, you must add the document to a folder in the repository by using the P8 4.x File Document In Folder task.

Perform the following steps to add and configure the P8 4.x File Document In Folder task:

a. From the Toolbox pane, under FileNet P8 4.x Repository, select the **P8 4.x File Document In Folder** task.

b. Click anywhere in the Designer pane between the Decision Point and the End node to drop the P8 4.x File Document In Folder task onto the task route.

c. You can configure the P8 4.x File Document in Folder task with the settings that are described in Table 6-4 on page 322.
Table 6-4  Settings of the P8 4.x File Document in Folder task

<table>
<thead>
<tr>
<th>Configuration section name</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connections drop-down list, select the connection to use when filing a document. In our example, we choose a predefined P8 4.x connector connecting to a P8 content object store.</td>
</tr>
<tr>
<td>File in Folder Options</td>
<td>▶ Folder path: In the Folder Path text box, provide the complete path to the folder in P8. In our example, because the Content Collector task route is filing the document into the Most Relevant Category folder in P8 based on the analysis by Classification Module, the folder path is set with the metadata type of IBM Classification Module and the property of Most Relevant Category. ▶ Select to create the folder if it does not exist.</td>
</tr>
</tbody>
</table>

Figure 6-24 on page 323 shows that the P8 4.x File Document in Folder task is configured to file documents into the most relevant folder in P8 suggested by Classification Module, and it also instructs the Content Collector to create the folder in P8 if it does not exist.
The P8 4.x Declare Record task allows you to declare a record in IBM FileNet Records Manager.

**Note:** You must install and configure IBM FileNet Records Manager 4.x in order to support the usage of the P8 4.x Declare Record task in a Content Collector task route (if it is not installed).

Perform the following steps to add and configure the P8 4.x Declare a Record task to your task route:

a. From the Toolbox pane, under FileNet P8 4.x Repository, select the **P8 4.x Declare Record** task.

b. Click anywhere in the Designer pane between the P8 4.x File Document in Folder task and the End node to drop the P8 4.x Declare Record task onto the task route.
c. You can configure the P8 4.x Declare Record task with the settings that are described in Table 6-5.

Table 6-5  Settings of the P8 4.x Declare Record task

<table>
<thead>
<tr>
<th>Configuration section name</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connections drop-down list, select the file plan object store in the repository in which to declare the record.</td>
</tr>
</tbody>
</table>
| Property Mappings          | ▶ Record Class: Click the Browse button to select a record class to use when declaring the record.  
                             ▶ Classification: To add a classification to use when declaring a record, in the Classification section, click **Add**, or to remove a classification, click **Remove**. The classification can be a static value, or it can be dynamically assigned (that is, based on the name of the folder in which a record is located). |

Figure 6-25 on page 325 shows that a P8 4.x Declare Record task is configured to declare a record for the document under the management of IBM FileNet Records Manager.
13. Use the FSC Post Processing task to define what happens to a file on the file system after it has been processed. The settings that you define apply only to the document on the file system, not to the document that is added to the IBM FileNet P8 repository. As its name suggests, you need to place a Post Processing task at or near the end of a task route.

Perform the following steps to add and configure the FSC Post Processing task in your task route:

a. From the Toolbox pane, under File System Source, select the FSC Post Processing task.

b. Click anywhere in the Designer pane between the P8 4.x Declare Record task and the End node to drop the FSC Post Processing task onto the task route.

c. You can configure the FSC Post Processing task with the settings that are described in Table 6-6 on page 326.
Table 6-6  Settings of the FSC Post Processing task

<table>
<thead>
<tr>
<th>Configuration section name</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter the name and description for the task.</td>
</tr>
<tr>
<td>Post Processing Options</td>
<td>You can either delete the file from the file system after capture, or you can retain the file on the system. In our example, we choose to retain the file on the file system by selecting <strong>Do not delete file</strong> and <strong>Move file to a folder</strong>.</td>
</tr>
</tbody>
</table>

Now, you have configured the task route with the required tasks for processing the documents whose most relevant score is greater than 90%, as shown in Figure 6-26.

![Figure 6-26  Configuring the FSC Post Processing task](image_url)
14. You now need to create a second rule to handle the documents whose most relevant scores are not greater than 90%.

Perform the following steps to add and configure the second rule in your task route:

a. From the Toolbox pane, select **Link**.

b. Click anywhere in the Designer pane. A new rule arrow appears in the Designer, but neither end of the rule arrow is connected to the task route.

c. Using your mouse, select and drag the tail of the new rule to join the decision point.

d. Perform the same steps as described at Step 12 to configure the second rule of Classification Module, **Not Greater than 0.9**, as shown in Figure 6-27.

![Figure 6-27  Configuring the rule of the right branch of the task route](image)
15. Just as in Step 13, a P8 4.x File Document in Folder task is required here to add the document to a folder in the repository in order to browse it at a later point.

Perform the following steps to add and configure the P8 4.x File Document In Folder task:

a. From the Toolbox pane, under FileNet P8 4.x Repository, select the P8 4.x File Document In Folder task.

b. Click anywhere in the Designer pane between the Decision Point and the FSC Post Processing task to drop the P8 4.x File Document in Folder task onto the task route.

c. You need to configure the P8 4.x File Document in Folder task with its general and P8 connection settings as described at Step 13. The only difference is that for the documents whose most relevant score is less than 90%, those documents are configured to be placed in a designated review folder in the P8 repository for later review, as shown in Figure 6-28.

![Figure 6-28 Configuring the P8 4.x File Document in Folder task](image-url)
16. Use a *link* to join the P8 4.x File Document in Folder task and the FSC Post Processing task to complete the task route creation:

a. From the Toolbox pane, select **Link**.

b. Click anywhere in the Designer pane. A new rule arrow appears in the Designer, but neither end of the rule arrow is connected to the task route.

c. Using your mouse, select and drag the tail of the new rule to join the **P8 4.x File Document in Folder** task.

d. Using your mouse, select and drag the head of the rule to join the **FSC Post Processing** task.

Figure 6-29 shows the complete task route that was created in our legal discovery and compliance use case.
6.5.3 Activating the system for archiving

Before using the Content Collector task route to perform the file system archiving, you are required to activate the integration system by using the following tasks:

- Starting the knowledge base
- Activating the file system collector of your task route
- Starting the task route engine

**Starting the knowledge base**

Perform the following steps to start your knowledge base on the Classification Module server:

1. Launch the Classification Module Management Console by clicking **Start** → **Programs** → **IBM Classification Module 8.6** → **Management Console**.

2. Start your knowledge base if it has not yet been started, as shown in Figure 6-30.

![Figure 6-30 Starting the knowledge base](image)

**Activating the file system collector of your task route**

Follow these steps to activate the file system collector of your task route:

1. Launch the Content Collector Configuration Manager by clicking **Start** → **Programs** → **IBM Content Collector** → **Configuration Manager**.

2. In the Configuration Manager, click the **Task Routes** tab in the Navigation pane, and then, in the Task Route explore pane, select your task route. In our example, it is **HR Task Route**.

3. In the Task Route Designer pane, click the **HR FSC Collector** task to display its configuration pane on the right side of the Configuration Manager. Ensure that this collector is **Active**, as shown in Figure 6-31. Then, click **Save**.
Starting the task route engine
Start the task route engine by clicking **Start → Programs → IBM Content Collector → Start Task Routing Engine**, as shown in Figure 6-32 on page 332.
6.6 Performing the file system archiving

After the Content Collector and Classification Module integration system is properly configured and activated, file share content is effortlessly captured and added to the content repository according to the configured business rules.

In order to better the activities on the back end, perform the following steps to observe how documents are seamlessly added to the ECM repository based on your Content Collector task route definition:

1. Launch and log in to your IBM FileNet P8 Workplace.

   Browse to your P8 content object store, and review the existing content under the target folder. Observe what and how new documents are added into your P8 object store according to the Content Collector task route run.

   In our example, the system is configured to ingest and auto-classify incoming HR documents into the HR LegalCase folder in the CE_OS object store. In this demonstration, the HR LegalCase folder starts empty, as shown in Figure 6-33 on page 333.
2. To capture documents:
   a. Open the file system folder monitored by the File System Collector. In our example, open the C:\HRMonitored folder.
   b. Drag and drop a set of file share files on the monitored folder. Figure 6-34 shows the example of ten files that were added to the C:\HRMonitored folder.
3. Check the monitored file system folder to observe the changes made to the files as they get captured into the ECM repository.

In our example, eight out of ten files under the C:\HRMonitored folder are moved to the C:\HRMonitored\Processed file system folder as they get captured. Two files are not captured and left untouched in the C:\HRMonitored folder, because their relevancy score is less than 30% per the Classification Module analysis, as shown in Figure 6-35.

![Figure 6-35 Non-captured files in the monitored file system folder](image)

4. In the IBM FileNet P8 Workplace, browse to your P8 content object store and view the documents in the target P8 folders.

In our example, browse to the CE_0S P8 object store. The HR LegalCase folder contains the following newly created subfolders, as shown in Figure 6-36 on page 335:

- 401k
- EAP
- Health Care
- Leaves of Absence
- Pay
- PTO
- Stock Options

These P8 folders are created by the system and align with the Most Relevant Category suggested by the Classification Module server through the Classification Module task processing.
5. You can further review the documents auto-classified into each of the sub-folders.

   In our example, according to the rules defined in the Content Collector task route, these documents have a *Most Relevant Score* that is greater than 90%.

6. The Content Collector task route in our example is configured so that the documents with a *Most Relevant Score* no greater than 90% but greater than 30% are captured into a review folder. You can browse to the review folder and check the documents inside it.

   For example, Figure 6-37 on page 336 shows the documents that are created in the **HRReview** folder, waiting for review.

   **Note:** When working with the P8 ECM repository, you can use the Classification Center to review the documents in the review folder and take the necessary actions to tune and improve the accuracy of your knowledge base. We describe this process in Chapter 5, “Integration with IBM FileNet P8” on page 177.
7. In addition to archiving the file share files into a meaningful folder path for easy discovery, you can use the Most Relevancy Score that is assigned to the Most Relevant Category as a threshold to mark the documents as records. In our example, documents with the Most Relevancy Score greater than 90% are declared as records. In the IBM FileNet Records Manager, you can browse to the destination file plan path to review the records declared for those seven documents whose Most Relevancy Score is greater than 90%, as shown in Figure 6-38 on page 337.
Figure 6-38  Records declared in the P8 file plan object store
Integration with IBM Content Collector for E-mail

In this chapter, we introduce the key aspects of integration of IBM Content Collector for E-mail with IBM Classification Module (Classification Module) and the typical business use cases in which to use this integration asset. In addition, we describe the steps to enable the integration and provide step-by-step integration and configuration instructions for a typical use case scenario.

We cover the following topics in the chapter:

- Solution introduction
- Integration steps overview
- Enabling the integration
- Use case 1: E-mail archiving with automatic classification
- Use Case 2: E-mail classification with records management
7.1 Solution introduction

IBM Content Collector for E-mail provides an integrated solution for archiving and classifying e-mails. Classification Module extends the filtering capabilities provided by IBM Content Collector, which are based on metadata, such as sender, recipients, sent date, or e-mail domain, by analyzing e-mail content. Using Classification Module with IBM Content Collector for E-mail is an extremely practical solution because of the growing legal, compliance, and regulatory requirements around e-mails. To meet compliance and records management initiatives, or to prepare collections of data for legal discovery, Classification Module can identify content that has no business value and exclude it from the e-mail archives. For example, e-mail that discusses a marketing strategy document can be moved to an IBM FileNet P8 folder and be declared as a record in IBM FileNet Records Manager. On the contrary, e-mail that discusses a weekend football game strategy is filtered out and does not need to be archived.

In addition to compliance requirements, you can use IBM Content Collector for E-mail integrated with Classification Module to filter out automated replies, personal messages, and other irrelevant e-mails before they are added to the e-mail archiving system to reduce the overall storage costs and the legal review costs.

In this section, we describe the following topics:
- Overview of IBM Content Collector for E-mail
- Integration architecture
- Use cases

7.1.1 Overview of IBM Content Collector for E-mail

IBM Content Collector for E-mail provides a user interface to define task routes that can archive e-mails from Microsoft Exchange or Lotus Domino® E-mail servers. You can schedule the defined task routes to archive the digital contents periodically.

Typically, a task route consists of a collector source for e-mails that follows a series of task processes:
1. Parse the e-mails to create or update metadata properties.
2. Invoke algorithms to detect duplication.
3. Stub the original e-mails with links to the archived e-mail.
4. Store the e-mail instance into the supported ECM repositories.
In addition, IBM Content Collector for E-mail can use the content classification capabilities of Classification Module:

- Archive e-mails into predefined folders. For faster discovery of information from archived content, it is important to organize the contents into a known set of predefined folders.
- Identify mission critical e-mails and declare them as records. To address any third-party litigation or compliance initiatives, it is necessary to identify digital content that might address issues related to enterprise policies or business missions.
- Populate metadata of the archived instance with category names that might be configured to support parametric searches and the faceted display of search results.

7.1.2 Integration architecture

Figure 7-1 shows the architecture overview of the integration between IBM Content Collector for E-mail and Classification Module.
Within this integration, Classification Module uses a knowledge base to analyze documents to discover possible categories and their relevancy scores. You can use the top relevancy score and the top category name in IBM Content Collector (Content Collector) task routes for any of the following typical scenarios:

- Assign the folder path for the archived e-mails in an IBM FileNet P8 or IBM Content Manager (CM8) repository.
- Identify e-mails that need to be declared as records.
- Identify e-mails that must be reviewed in the Classification Center to enhance the knowledge base accuracy (applicable to IBM FileNet P8 repository type only when the Classification Module and IBM FileNet P8 integration asset is installed and configured).

### 7.1.3 Use cases

A number of scenarios exist where the Classification Module with IBM Content Collector for E-mail solution can be applied for your enterprise.

Assume that you want to filter out automated replies, personal messages, and other irrelevant e-mails before they are added to the e-mail archive to reduce overall storage costs and eventual possible legal review costs. You want to further organize certain e-mails into categories (such as 401k, Employee Assistance Program (EAP), and Health Care). You want to know why particular e-mails are filtered out so that you can be confident that relevant information is not excluded. In addition, you want to ensure that you set up the right filters to collect all potentially relevant e-mails, while filtering out any of the day-to-day irrelevant or “noisy” e-mails.

Another example is that the IT Administrator for your enterprise has been asked to create and maintain the corporate e-mail archive. In order to control the ongoing IT costs and potential legal discovery costs, your company has instructed the IT Administrator to actively manage the e-mail archive to avoid adding irrelevant data whenever possible. The IT Administrator needs to ensure that each relevant e-mail's entire message content is archived, including attachments. In addition, the IT Administrator wants to filter out irrelevant or noisy data, such as personal messages, company bulletins, and meeting and calendar invitations.

With the combined strength of Classification Module with IBM Content Collector, the IT Administrator can perform these tasks to achieve these requirements:

- Define a set of rules for automatic classification of e-mail messages into particular folders, such as Contracts, Claims, and Human Resources.
Configure rules based on corporate records management policies that define which e-mail messages to declare as records.

Define a set of rules that filters out irrelevant e-mail messages before they are archived.

Train the system using a small set of user mailboxes to serve as a set of representative e-mails.

Build the archive from existing e-mail messages over the past year, resulting in an initial archive.

Configure the system to actively add to the e-mail archive daily after performing the initial archive.

Understand how and why Classification Module is filtering various e-mail messages into various categories, and then, tune the filtering mechanism as needed.

Review e-mails that were not auto-classified or those e-mails that were sent to review for auditing purposes.

### 7.2 Integration steps overview

To set up the integration, configure a task route in IBM Content Collector for Classification Module that identifies the knowledge base that you want to use for classifying content. When IBM Content Collector processes e-mail in monitored locations, it applies the rules in the task route. Through content analysis, the knowledge base determines whether a document or e-mail needs to be captured and, if so, how it is to be classified. Classified items are moved to an IBM FileNet P8 or IBM Content Manager (CM8) repository or declared as records, as appropriate.

The following overview summarizes the steps to integrate Classification Module with IBM Content Collector for E-mail:

1. Enable the integration.
2. Use Configuration Manager to create the e-mail archiving task route.
3. Use Configuration Manager to create the e-mail archiving task route with records declaration.

**Note:** In addition to these integration steps, you can set up the Classification Center to review unclassified documents and to provide feedback to the classification engine.
7.3 Enabling the integration

In this section, we discuss how to enable the integration between IBM Content Collector for E-mail and Classification Module, for the automatic classification of e-mails and filtering and archiving e-mails.

Enabling integration consists of the following steps:

1. Configure IBM Content Collector server.
2. Configure Classification Module server.
3. Validate the integration.

7.3.1 Configuring IBM Content Collector server

Assuming that you have successfully installed IBM Content Collector Version 2.1 on a server, use the following procedures to configure the system to use Classification Module as a utility connector:

1. Install the Classification Module client module.
2. Register Classification Module utility connector.
3. Install Classification Module server and IBM FileNet P8 connectivity.

Installing the Classification Module client module

For IBM Content Collector to use the classification capabilities of Classification Module Version 8.6, you must install the Classification Module client components on the IBM Content Collector server.

Run the Classification Module installation program on the IBM Content Collector server. Select the option to install Custom components, and select the IBM Classification Module client only check box.

Figure 7-2 on page 345 shows the Classification Module software that must be installed on the IBM Content Collector server.
Registering Classification Module utility connector

You must enable Classification Module and IBM Content Collector for E-mail integration before the Classification Module task is available to use in a task route.

To integrate with a Classification Module server, IBM Content Collector for E-mail requires access to the following three Classification Module .dll files, which are located in the /bin directory of the Classification Module installation:

- bnsClient86.dll
- PackageDll23.dll
- stlport_ban46.dll

To register the Classification Module .dll files:

1. Copy the three .dll files to the ctms directory under the IBM Content Collector for E-mail installation path (for our case study, it is the C:\Program Files\IBM\ContentCollector\ctms directory), and rename the bnsClient86.dll file to the bnsClient85.dll file.

2. Open a command window, from the ctms directory, run the utility connector in registration mode by entering the following command:

   utilityconnector.exe -r
If the command returns an error message saying, "Failed to create service: ibm.ctms.utilityconnector.UtilityConnector -- The specified service already exists", most likely, you registered the .dll files in the wrong location. To correct the operation, perform the following steps:

a. Use the command below to first unregister the Classification Module .dll files in the location where you issued the initial registering command.

   utilityconnector.exe -u

b. Ensure that you change to the ctms directory and register the .dll files again using the following command:

   utilityconnector.exe -r

### Installing Classification Module server and FileNet P8 connectivity

Refer to Chapter 5, “Integration with IBM FileNet P8” on page 177 for instructions to set up Classification Module server with IBM FileNet P8 connectivity. Setting up Classification Module server with IBM FileNet P8 connectivity is required for the Classification Center to start up. With Classification Center, you can reclassify documents, review documents in folders that did not meet certain thresholds, and potentially classify valid documents into their corresponding category and provide feedback to the knowledge base.

#### 7.3.2 Configuring Classification Module server

This section provides the integration instructions for leveraging the classification capabilities of Classification Module with the detailed archiving features that are provided by IBM Content Collector. For better scalability and performance, install these products on separate servers.

Assuming that you have successfully installed Classification Module Version 8.6 on a server, use the following procedures to configure the system to classify e-mails:

1. Installing Microsoft Office Outlook 2003 or Microsoft Office Outlook 2007.
2. Configuring the Classification Module server to archive e-mail.

### Installing Microsoft Office Outlook 2003 or Microsoft Office Outlook 2007

The e-mail archiving filters in Classification Module use the Messaging Application Program Interface (MAPI) for parsing e-mails from Microsoft Exchange server.
To configure this support:

1. Install Microsoft Office Outlook® 2003 or Microsoft Office Outlook 2007 on the server that hosts Classification Module.

2. Select **Microsoft Office Outlook** as the default e-mail application in your Web browser (see Figure 7-3).

![Internet Explorer Options setting](image)

*Figure 7-3  Internet Explorer Options setting*

**Configuring the Classification Module server to archive e-mail**

In this scenario, IBM Content Collector archives e-mails from a Microsoft Exchange server, and then, Classification Module features classify the e-mails.

Follow these instructions to tune the Classification Module server for archiving e-mails:

1. Log on to the Classification Module Version 8.6 server.
2. Stop the Classification Module services:
   a. Launch Windows services.
   b. Stop the Classification Module Process Manager service.
   c. Stop the Classification Module Trace Service service.
3. If e-mails are being archived, overwrite the default document filter:
   a. Open a DOS command window.
   b. Change to the C:\IBM\ClassificationModule\Filters directory and enter these commands:
      ```
      copy docFilterManager.xml docFilterManager.xml.orig
      copy docFilterManager.E-mail.xml docFilterManager.xml
      ```
4. Start the Classification Module services:
   a. Launch Windows services.
   b. Start the Classification Module Process Manager service.
   c. Start the Classification Module Trace Service service.

### 7.3.3 Validating the integration

To validate the integration of IBM Content Collector for E-mail with Classification Module, make sure that you set up these items and configure these values properly in the system:

- Windows registry entry of `ibm.ctms.utilityconnector.ICMClassificationTask`
- Classification Module task type in IBM Content Collector Configuration Manager
- Classification Module system metadata properties in IBM Content Collector Configuration Manager

**Windows registry entry**

Confirm that the `ibm.ctms.utilityconnector.ICMClassificationTask` Windows registry entry is available by using the following steps:

1. Launch the Registry Editor by issuing the following command in the command prompt:
   ```
   regedit
   ```

2. In the Registry Editor window, navigate to **My Computer → HKEY_LOCAL_MACHINE → SOFTWARE → IBM → EMRC → 4.0 → Service Declarations2 → ibm.ctms.utilityconnector.UtilityConnector**, and validate that the registry entry of
ibm.ctms.utilityconnector.ICMClassificationTask is listed, as shown in Figure 7-4.

![Figure 7-4](image)

**Classification Module task in IBM Content Collector Configuration Manager**

Verify that the Classification Module task type is available in Configuration Manager:

1. Launch the IBM Content Collector Configuration Manager by clicking **Start → Programs → IBM Content Collector → Configuration Manager**.

2. Select **Task Routes** in the explorer pane of the IBM Content Collector Configuration Manager.

3. From the Toolbox pane, ensure that the IBM Classification Module task type is listed under the Utility tasks, as shown in Figure 7-5 on page 350.
System metadata

The integration introduces five Classification Module system metadata properties, as described in Table 7-1.

Table 7-1  Classification Module system metadata

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Relevant Categories</td>
<td>List of top categories matched</td>
</tr>
<tr>
<td>All Relevant Categories and Scores</td>
<td>Combined list of categories and scores</td>
</tr>
<tr>
<td>All Relevant Scores</td>
<td>List of top scores</td>
</tr>
<tr>
<td>Most Relevant Category</td>
<td>Winning category</td>
</tr>
<tr>
<td>Most Relevant Score</td>
<td>Winning category score</td>
</tr>
</tbody>
</table>

The Classification Module task produces the metadata values during the run time. In turn, you can use them in the task route to determine how to process documents.
Use the following steps to ensure that the Classification Module system metadata properties are available in the IBM Content Collector Configuration Manager:

1. Launch the IBM Content Collector’s Configuration Manager by clicking **Start → Programs → IBM Content Collector → Configuration Manager**.

2. Go to the Metadata and Lists box by clicking **Metadata and Lists** in the explorer pane of the Configuration Manager.

3. From the Metadata and Lists box on the left pane, select the **System Metadata** type, and then, select the **IBM Classification Manager** system metadata in the middle pane. Ensure that five Classification Module system metadata properties are displayed on the right pane, as shown in Figure 7-6.

![Figure 7-6 IBM Classification Manager system metadata](image-url)
7.4 Use case 1: E-mail archiving with automatic classification

In this section, we demonstrate the step-by-step process of using IBM Content Collector for E-mail with Classification Module to automatically classify and archive e-mails into the IBM FileNet P8 content repository. We cover the following steps:

1. Create a knowledge base.
2. Start the servers.
3. Perform pre-task route configuration.
4. Create an e-mail task route.
5. Activate the e-mail task route.
6. Verify the use case end results.

7.4.1 Creating a knowledge base

The first step in using IBM Content Collector with Classification Module for e-mail archiving is to create a knowledge base and a decision plan that will be used by the Classification Module server. The Classification Module server will be used inside the IBM Content Collector task route for the classification needs.

To create a knowledge base, the manner in which you gather sample documents to jump-start your knowledge base depends on the status of your data. You can use either one or a combination of the following ways:

- If your documents are already organized in an IBM FileNet P8 object store, follow these steps:
  a. Configure and run the Content Extractor.
  b. Import the extracted content and XML data into Classification Workbench.
  c. Build a knowledge base that is based on the IBM FileNet P8 folder or document class structure.

- If your documents are not in IBM FileNet P8 yet, but they are categorized (for example, in file system folders), follow these steps:
  a. Import documents directly into Classification Workbench.
  b. Build a knowledge base that is based on the file system directory structure.
  c. Create folders and document classes in an IBM FileNet P8 object store that correspond to categories in the knowledge base.
If your documents are not categorized, but you know how they need to be categorized, follow these steps:

a. Import documents into Classification Workbench.

b. Assign categories to documents by using Classification Workbench and build a knowledge base.

c. Create folders and document classes in an IBM FileNet P8 object store that correspond to categories in the knowledge base.

If your documents are not categorized and you do not know how they need to be categorized, follow these steps:

a. Import documents into the Taxonomy Proposer and discover categories.

b. Import categorized content items into Classification Workbench.

c. Build a knowledge base that is based on the discovered categories.

d. Create folders and document classes in an IBM FileNet P8 object store that correspond to categories in the knowledge base.

If none of the previous situations apply, use a keyword list and follow these steps:

a. Create a category structure with a starter list of keywords representative of these categories.

b. Use this list of keywords to create and analyze the knowledge base.

For our use case, we use the HR knowledge base that we used to demonstrate creating, training, and tuning a knowledge base in Chapter 3, “Working with knowledge bases and decision plans” on page 59. Figure 7-7 on page 354 shows the HR knowledge base that we use in this use case.
7.4.2 Verifying and starting the servers

You must configure, verify, and start various integration components before you can create a new task route to accomplish your business use case.

The following integration components are required to be verified and started:

- Classification Module server
- IBM Content Collector Server
- IBM FileNet P8 Server

**Classification Module server**

Perform the following steps on the Classification Module server machine:

1. Start the knowledge base that you will use for the use case if it is not already started. You can verify that the knowledge base is started by using Management Console. In Figure 7-8 on page 355, you see that Redbook HR knowledge base is started with one read-write and one read-only process.
2. Start Classification Center. Make sure that there are no connection issues from the Classification Module server to the IBM FileNet P8 systems. To verify, log on to the Classification Center using IBM FileNet P8 credentials to see the Configuration tab. See Figure 7-9 on page 356.
To start using IBM Content Collector’s Configuration Manager to configure task routes, you need to start the IBM Content Collector server. Perform the following steps on the IBM Content Collector server machine:

1. The IBM Content Collector Task Route Engine service must be stopped if it is running, as shown in Figure 7-10 on page 357.
2. Start Configuration Manager by running **Start → Program → IBM Content Collector → Configuration Manager**, as shown in Figure 7-11 on page 358.
A task starting dialog appears, as shown in Figure 7-12.

A Quickstart tutorial page appears, as shown in Figure 7-13 on page 359.
3. Click **Close**, and you see the Configuration Manager application user interface, as shown in Figure 7-14 on page 360. The highlighted sections are the major components that make up the Content Collector interface. Starting from top to bottom, going from left to right, these areas are highlighted:

- Task Routes Pane
- Toolbox Pane
- Navigation Pane
- Designer Pane
- Configuration or Task Route Pane
Figure 7-14 Configuration Manager User Interface
IBM FileNet P8 server

Perform these steps for verification:

1. Verify that you can connect to Workplace or WorkplaceXT (depending on what product you have installed) by either browsing to the Web URL or by using IBM FileNet Enterprise Manager. To go through the URL, open a browser and go to the URL, as shown in Figure 7-15:

   http://hostname:port/Workplace

   ![Figure 7-15 FileNet P8 Workplace](image)

2. If you have IBM FileNet Records Manager installed on an IBM FileNet P8 machine and you plan on declaring records using IBM Content Collector and Classification Module task route nodes, verify that you can connect to IBM FileNet Records Manager. You can verify that you can connect to IBM FileNet Records Manager by opening a browser and going to the following URL, as shown in Figure 7-16 on page 362:

   http://hostname:port/RecordsManager
7.4.3 Pre-task route configuration

You need to verify the configuration of the following Navigation pane components:

- Data Stores
- Connectors

**Data Stores**

To configure Data Stores, click **Data Stores** on the Navigation Pane and select the appropriate configured database from the explorer pane on the left bar. Table 7-2 describes the settings of the Data Store.

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter your Data Store name and its description. Also, ensure that <strong>Make this my active data store</strong> is selected.</td>
</tr>
</tbody>
</table>
In our use case example, we use the Data Store settings, as shown in Figure 7-17.

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database Information</td>
<td>Enter your database alias name, login name, and login password to access the database. Click <strong>Validate Database</strong> to check the database connection.</td>
</tr>
<tr>
<td>Database Properties</td>
<td>Click either <strong>Export</strong> or <strong>Import</strong> to export or import database properties.</td>
</tr>
</tbody>
</table>

**Connectors**

Click the **Connectors** icon under the Navigation pane to configure connectors. There are several types of connectors that you must configure for the IBM Content Collector task routes to work as required:
IBM FileNet P8 Connectors

To configure IBM FileNet P8 Connectors, click Connectors on the Navigation pane and select FileNet P8 4.x Repository from the explorer pane on the left bar. Configure IBM FileNet P8 connections to each of the object stores that you will use in your task routes. Table 7-3 describes the settings of the IBM FileNet P8 Connections.

Table 7-3 Settings under General tab of the IBM FileNet P8 Connection

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter your FileNet P8 4.x Connection name and description.</td>
</tr>
<tr>
<td>Content Engine URL</td>
<td>Enter your FileNet P8 Connection parameters, including Connection Type, Server IP address or Host name, Port Number, Path, and the Protocol to connect. The connection URL is constructed automatically with the input provided.</td>
</tr>
<tr>
<td>P8 Login Information</td>
<td>Enter the User name and Password to connect to FileNet P8. Click Retrieve object stores to get all the object stores configured on this FileNet P8 platform. The Domain is automatically filled up for you after the object store is determined. <strong>Note:</strong> Create connections to each object store using the New icon in the Designer pane.</td>
</tr>
</tbody>
</table>

In our example, we use three IBM FileNet P8 Connections for three object stores. See Figure 7-18 on page 365.
To configure the E-mail Server Connector, click the Connectors tab in the Navigation pane, and then, choose the E-Mail Server in the explorer pane on the left bar. Configure a connector to each of the e-mail servers that you will use. In our example, we use one Exchange e-mail server. Table 7-4 on page 366 describes the settings of the E-mail Server Connector General tab.
Table 7-4  Settings under General tab of the E-mail Server Connection

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter your E-mail Server name and its description. Also, select the Mail System using the drop-down list.</td>
</tr>
<tr>
<td>Log Settings</td>
<td>Select the appropriate Log Level using the drop-down list. Select the Log file Location using the Browse button. Truncate the log files and enter the number and size of the log files allowed.</td>
</tr>
<tr>
<td>Logging Type</td>
<td>Choose the appropriate Logging Type between Common base event or Plain text.</td>
</tr>
<tr>
<td>Working Directory</td>
<td>Specify the location of the Working Directory by using the Browser button.</td>
</tr>
</tbody>
</table>

Figure 7-19 on page 367 shows the setting in the General tab for the E-mail Server Connector in our example.
Table 7-5 describes the settings of the E-mail Server Connector Connection tab.

**Table 7-5 Setting under Connection tab of the E-mail Server Connection**

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connection Parameters</td>
<td>Enter the Connection parameters to the E-mail server. For example, in the case of Exchange, enter the Exchange server host name and the User ID for the Exchange server.</td>
</tr>
</tbody>
</table>

In our example, we use the Connection tab settings for the E-mail Server Connector, as shown in Figure 7-20 on page 368.
Table 7-6 describes the settings of the E-mail Server Connector Active Directory tab.

Table 7-6  Settings under Active Directory tab of the E-mail Server Connection

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credentials</td>
<td>Enter your User ID and Password to access the Active Directory.</td>
</tr>
<tr>
<td>Location</td>
<td>Choose the Location of the Active Directory as either Domain default or User defined.</td>
</tr>
</tbody>
</table>

In our example, we use the Active Directory tab settings for E-mail Server Connector, as shown in Figure 7-21 on page 369.
Click the Metadata and Lists icon under the Navigation pane to configure the system metadata and lists. In our example, we use the following system metadata, as shown in Figure 7-22 on page 370.

**Note:** Verify that IBM Classification Module is one of the installed system metadata and that it has all of the five properties as shown.
You must configure a required set of e-mail services as a pre-task route configuration step. The required set of e-mail services depends on your e-mail platform in use, such as Microsoft Outlook or Lotus Notes®. With either platform, you need to find out what e-mail components need to be captured. For each platform, you then need to install the corresponding Microsoft Outlook components as e-mail services for the IBM Content Collector server, specifically Outlook Client with a connection to the Exchange Server, a connection to a PST file, and a browser connection to the Outlook Web Mail Server. In our example, we configure the following e-mail services. To configure the e-mail services, click the E-mail Services tab in the Navigation pane to configure the following e-mail services:

- Client Configuration
- Configuration Web Service
- Information Center
- Web Application Client

Figure 7-22  Metadata and Lists configuration: System Metadata
**Client Configuration**

To configure Client Configuration, click the **E-mail Services** tab from the Navigation pane, and then, select **Client Configuration** from the explorer pane on the left bar. Table 7-7 describes the settings for the Client Configuration for E-mail Service.

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter the Name and Description of the Client Configuration.</td>
</tr>
<tr>
<td>Client Definition</td>
<td>Specify the Trigger mailbox, that is, the mailbox that is being monitored.</td>
</tr>
</tbody>
</table>

In our example, we use the Client Configuration settings for E-mail Services that are shown in Figure 7-23.

**Configuration Web Service**

To configure Configuration Web Service, click the **E-mail Services** tab from the Navigation pane, and then, select **Configuration Web Service** from the explorer pane on the left bar. Table 7-8 on page 372 describes the settings for the Configuration Web Service for E-mail Services.
Table 7-8 Settings for Configuration Web Service for E-mail Services

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter the Name and Description of the Configuration Web Service.</td>
</tr>
<tr>
<td>Configuration Web Service Definition</td>
<td>Enter the Configuration Web Service Definition, such as Host name and Port number. You can use <strong>Validate</strong> to check the Configuration Web Service. Use the check box to let IBM Content Collector use the embedded Web application server. Specify the Java Database Connectivity (JDBC™) driver directory by using <strong>Browse</strong>. Specify the JDBC Port number and the Database Server.</td>
</tr>
</tbody>
</table>

In our example, we use the Configuration Web Service settings for E-mail services that are shown in Figure 7-24.

![Figure 7-24 E-mail Services: Configuration Web Service settings](image)
**Information Center**
To configure Information Center, click the **E-mail Services** tab from the Navigation pane, and then, select **Information Center** from the explorer pane on the left bar. Table 7-9 describes the settings for the Information Center for E-mail Services.

<table>
<thead>
<tr>
<th>Property Group Name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter the Name and Description of the Information Center.</td>
</tr>
<tr>
<td>Information Definition</td>
<td>Enter the Information Definition Host name and Password.</td>
</tr>
</tbody>
</table>

In our example, we use the Configuration Web Service settings for E-mail services that are shown in Figure 7-25.

**Web Application Client**
To configure Web Application Client, click the **E-mail Services** tab from the Navigation pane, and then, select **Web Application Client** from the explorer pane on the left bar. Table 7-10 on page 374 describes the settings for the Web Application Client for E-mail Services.
Table 7-10  Settings for Web Application Client for E-mail Services

<table>
<thead>
<tr>
<th>Property group name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter the Name and Description of the Web Application Client.</td>
</tr>
<tr>
<td>Web Application Definition</td>
<td>Enter the Host name and Port number for the Web Application Client.</td>
</tr>
<tr>
<td>Repository Connection</td>
<td>Select between the two Repository Connections for connection to the ECM system.</td>
</tr>
</tbody>
</table>

In our example, we use the Web Application Client settings for E-mail services that are shown in Figure 7-26.

7.4.4 Creating an e-mail task route

In this section, we create a task route from the beginning to show you how you can use Classification Module capabilities to automatically classify e-mails into the appropriate folders and archive them into an ECM repository, such as IBM
FileNet P8 content repository. We use the HR knowledge base to classify HR related e-mails into the appropriate categories. These steps summarize building our task route for automatic classification:

1. Create an empty task route.
2. Create the **EC Collect E-mails by Rules** task node.
3. Create the **EC Extract Metadata** task node.
4. Create the **EC Prepare E-mail for Compliance** task node.
5. Create the **P8 4.x Create Document** task node.
6. Create the **Classification Module** task node.
7. Create a Decision Point and **Rules**.
8. Create the **P8 File Document to Folder** task nodes.
9. Create the **EC prepare E-Mail for Stubbing** task node.
10. Create the **EC Create E-Mail Stub** task node.

We discuss each of these steps in detail.

**Create empty task route**

In this step, we begin by creating an empty task route:

1. Start configuration Manager by clicking **Start → Program → IBM Content Collector → Configuration Manager** under IBM Content Collector server.
2. Click the **Create New** icon to start creating a new task route, as shown in Figure 7-27.

You can either start by creating a new **Blank task route** or if your task route has a similar workflow to an existing task route templates, you can create
your task route from an **Existing template**. In this example, we select **Blank task route** and click **Create**, as shown in Figure 7-28.

**Figure 7-28 Creating new task route from a blank task route**

A new blank task route with Start and End nodes appears in the Designer pane, as shown in Figure 7-29.

**Figure 7-29 Empty task route showing start and end task nodes**

3. Click the green start task node, as highlighted in Figure 7-29, to configure the general properties for the task route. Configure the name and description of the task route. We configured the task route, as shown in Figure 7-30 on page 377.
Create the Collect E-Mail By Rules task node

For a new e-mail system task route, an E-mail Collector (EC) is required to define the location from where the e-mails will be captured. Perform the following steps to add and configure an EC Collect E-Mail By Rules task node in your task route:

1. From the Toolbox pane, under E-mail Server, select **EC Collect E-Mail By Rules**, as shown in Figure 7-31 on page 378.
2. Click anywhere in the Task Route Designer pane to drop the new EC Collect E-Mail By Rules task onto the task route, as shown in Figure 7-32 on page 379.
3. After adding the EC Collect E-Mail By Rules, you can configure it:
   - Name and enable the collector under the General tab.
   - Specify when to collect files under the Schedule tab.
   - Specify from where to collect files under the Collection Sources tab.
   - Specify files that must not be collected under the Filter tab.
Table 7-11 describes the settings of the EC Collect E-Mail By Rules under each tab for our example.

Table 7-11  Settings of the EC Collect E-mail By Rules task node

<table>
<thead>
<tr>
<th>Configuration tab name</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter your task route name and its description. Also, ensure that this collector is <strong>Active</strong>.</td>
</tr>
<tr>
<td>Schedule</td>
<td>The collector can run at various frequencies, such as daily, weekly, and monthly. In this example, we set the collector to run <strong>Always</strong>.</td>
</tr>
<tr>
<td>Collection Source</td>
<td>Define a mailbox to be monitored by the collector. Under the Collection Sources, click <strong>Add</strong> to add the mailbox to be monitored. In this example, we select <strong>Mailbox</strong> as the Source Type and we provide the Mailbox Simple Mail Transfer Protocol (SMTP) address.</td>
</tr>
<tr>
<td>Filter</td>
<td>Set the filter to constraint messages, exclude or include monitored folders, and set excluded message types. In this example, we select defaults for the filter options.</td>
</tr>
</tbody>
</table>

Figure 7-33 shows the Schedule settings for EC Collect E-mail By Rules task node for our example.

![Figure 7-33  EC Collect E-mail By Rules: Schedule tab settings](image)
Figure 7-34 shows the Collection Sources settings for EC Collect E-mail By Rules task node for our example.

Click **Add** to add Collection Sources, as shown in Figure 7-35 on page 382.
Figure 7-35  EC Collect E-mail By Rules: Collection Source addition

Figure 7-36 on page 383 shows the Filter settings for EC Collect E-mail By Rules task node for our example.
At this point, we have completed the creation of the E-mail collector task node. In the next section, we create the next task node for our task route.

**Create EC Extract Metadata task node**

This task extracts e-mail metadata and calculates a de-duplication hash key for each e-mail. In this section, we create an EC Extract Metadata task node.
Perform the following steps to add and configure an EC Extract Metadata task node in your task route:

1. From the Toolbox pane, under E-mail Server, select **EC Extract Metadata**, as shown in Figure 7-37.

![Figure 7-37 Selecting EC Extract Metadata task node](image)
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the Start and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. You can click the newly created task node to configure the task node properties.

3. After adding the EC Extract Metadata, you can configure it by entering the Name, Description, and Associate Metadata under the General tab.

Figure 7-38 shows the EC Extract Metadata task node settings for our example.

![Figure 7-38 EC Extract Metadata: General tab settings](image)

At this point, we have completed the creation of Extract Metadata task node. In the next section, we create the next task node for our task route.

**Create EC Prepare E-mail for Archive task node**

This task downloads e-mail from the server and stores it as temporary files for local processing. In this section, we create an EC Prepare E-mail for archive task node. Perform the following steps to add and configure an EC Prepare E-mail for Archive task node in your task route:

1. From the Toolbox pane, under E-mail Server, select **EC Prepare E-Mail for Archive**, as shown in Figure 7-39 on page 386.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the EC Extract Metadata and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. You can click the newly created task node to configure the task node properties.
3. After adding the EC Prepare E-mail for Archive, you can configure it by entering the Name and Description under the General tab. Figure 7-40 shows the EC Prepare E-mail for Archive task node settings for our example.

![EC Prepare E-mail for Archive: General tab settings](image)

At this point, we have completed the creation of the Prepare E-mail for Archive task node. In the next section, we create the next task node for our task route.

**Create EC Finalize E-mail for Compliance task node**

This task converts the temporary files that are created by the EC Prepare E-mail For Archive task to bundled resource items (BRIs). In this section, we create an EC Finalize E-mail for Compliance task node. Perform the following steps to add and configure an EC Finalize E-mail for Compliance task node in your task route:

1. From the Toolbox pane, under E-mail Server, select **EC Finalize E-mail for Compliance**, as shown in Figure 7-41 on page 388.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the EC Prepare E-Mail for Archive and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. You can click the newly created task node to configure the task node properties.
3. After adding the EC Finalize E-mail for Compliance, you can configure by entering the Name and Description under the General tab. Figure 7-42 shows the EC Finalize E-mail for Compliance task node settings for our example.

![EC Finalize E-mail for Compliance: General tab settings](image)

At this point, we have completed the creation of the Finalize E-mail for Compliance task node. In the next section, we create the next task node for our task route.

**Create P8 4.x Create Document task node**

This task creates a document object in IBM FileNet P8 for the item being processed. You can select options to defer the check-in of the item to a later point in the task route. You can also select options indicating whether document metadata only is saved into IBM FileNet P8, or if document metadata and content are saved. (Content can be added to the document at a later point in the task route). In this section, we create a P8 4.x Create Document task node. Perform the following steps to add and configure an P8 4.x Create Document task node in your task route:

1. From the Toolbox pane, under IBM FileNet P8 4.x Repository, select **P8 4.x Create Document**, as shown in Figure 7-43 on page 390.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the EC Finalize E-mail for Compliance and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. You can click the newly created task node to configure the task node properties.
3. After adding the P8 4.x Create Document task node, you can configure it with the settings that are shown in Table 7-12.

Table 7-12  Settings for the P8 4.x Create Document task node

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connection drop-down list, select the connection to use when creating a document. In this example, we choose a predefined P8 4.x connector connecting to a P8 content object store.</td>
</tr>
<tr>
<td>Check In Options</td>
<td>Set Check In options for your document, such as version and content capture options. These options are required by FileNet P8. For more information about each option, refer to the IBM FileNet Enterprise Manager documentation. In this example, we accept the default settings.</td>
</tr>
<tr>
<td>Property Mappings</td>
<td>Select the document class that you want to use when creating the document, and enter values for the document class properties. In this example, we select the document class called document, and map the Document Title property with the E-mail Subject metadata value.</td>
</tr>
</tbody>
</table>

Figure 7-44 on page 392 shows the P8 4.x Create Document task node settings for our example.
At this point, we have completed the P8 4.x Create Document task node. In the next section, we create the next task node for our task route.
Create Classification Module task node

This task uses the document object in IBM FileNet P8 and sends it to the Classification Module server in order to find the confidence scores and the right classification folder. In this section, we create a Classification Module task node. Follow these steps to add and configure a Classification Module task node in your task route:

1. From the Toolbox pane, under Utility, select IBM Classification Module, as shown in Figure 7-45 on page 394.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the P8 4.x Create Document and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. You can click the newly created task node to configure the task node properties.
3. Configure the Classification Module task, as shown in Table 7-13 on page 395.

Table 7-13  Settings for Classification Module task node

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>Server</td>
<td>Specify the Classification Module server address and the listener port. The default port is 18087.</td>
</tr>
<tr>
<td>Classification</td>
<td>Select the knowledge base that you want to use for analyzing the files collected from the monitored file share. Also, select the content field that identifies the part of a document that you want Classification Module to analyze. To process the content of the document, this field must be type Document.</td>
</tr>
<tr>
<td>Result Set</td>
<td>Specify the relevance threshold. In our example, we need to set the threshold to 0.3, which means that Classification Module only returns suggested categories with relevancy scores greater than 0.3. Therefore, if the relevancy score of the top category is not greater than 30%, this document is not captured by IBM Content Collector. Also, specify the maximum number of categories suggested by Classification Module to be used by IBM Content Collector for processing in a task route.</td>
</tr>
</tbody>
</table>

Figure 7-46 on page 396 shows the settings for the Classification Module task node.
At this point, we have completed the creation of the Classification Module task node. In the next section, we create the next task node for our task route.

**Introduce a decision point and rules**

To support the legal discovery and compliance use case, the IBM Content Collector task route uses the *Most Relevant Score* and *Most Relevant Category* system metadata properties to determine how each e-mail is processed. To do
so, you need to create a decision point and two rules branching from the Classification Module task to deal with documents based on the values of the *Most Relevant Score* and *Most Relevant Category* system metadata properties.

A *decision point* is a node where you can define multiple rules based on certain conditions. In our example, we introduce two rules based on the relevancy score of each e-mail after being examined and returned by the Classification Module task node. Follow these steps to add a new decision point to your task route:

1. From the Toolbox pane, select **Decision Point**.
2. Click anywhere in the Designer pane between the Classification Module task and the End node to drop the new decision point onto the task route. The first rule link is highlighted by default, as shown in Figure 7-47 on page 398.
3. Configure the arrow connecting the Decision Point and the End node as a rule by taking the following steps:
   a. Click the rule arrow to display the rule Configuration Pane.
   b. Enter the name for the rule. Because this branch of the task route only processes documents with a *Most Relevant Score* value (of its *Most
Relevant Category) that is not greater than 70%, we name this rule Score < 70%.

c. For the evaluation criteria, select Configure rule, and click Add.

d. In the Edit Conditional Clause window, configure the rule with the following fields, as shown in Figure 7-48:
   - For the Metadata type, select IBM Classification Module.
   - For the Property, select Most Relevant Score.
   - For the Operator, select Not Greater than.

Select the Literal radio button with the value of 0.7.

Figure 7-48 Conditional clause for Score < 70 rule

Figure 7-49 on page 400 shows the settings for the Score < 70 decision rule.
4. We now configure the other Decision Rule by taking the following steps:
   a. Highlight the Decision Point, right-click the Decision Point, and choose 
      **Add Rule**, as shown in Figure 7-50 on page 401.
The new decision rule is now displayed in the task route, as shown in Figure 7-51 on page 402.
b. Enter a name for the rule. Because this branch of the task route processes documents with a Most Relevant Score value (of its Most Relevant Category) that is greater than 70%, we name this rule Score > 70%.

c. For the evaluation criteria, select the Configure rule radio button, and click Add.
d. In the Edit Conditional Clause window, configure the rule with the following fields, as shown in Figure 7-52:
   - For the Metadata type, select IBM Classification Module.
   - For the Property, select Most Relevant Score.
   - For the Operator, select Greater than.

Select the Literal radio button with the value of 0.7.

![Figure 7-52   Conditional metadata rule]

At this point, we have completed the creation of a decision point and two rules. In the next section, we create the next task nodes for our task route.

**Create P8 4.x File Document in Folder task nodes**

This task files document objects that were created in the repository into specific folders, enabling users to browse for the documents at a later date. In this section, we create two P8 4.x File Document in Folder task nodes, one for each of the two decision point rules that we created in “Introduce a decision point and rules” on page 396. Follow these steps to add and configure a Create P8 4.x File Document in Folder task node in your task route:

1. From the Toolbox pane, under FileNet P8 4.x Repository, select Create P8 4.x File Document in Folder, as shown in Figure 7-53 on page 404.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the Decision Point and the end of Rule link (Score >= 70%) until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. Click the newly created task node to configure the task node properties.
3. Configure the P8 4.x File Document in Folder task with the settings that are described in Table 7-14.

Table 7-14   Settings for the P8 4.x File Document in Folder task node

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connections drop-down list, select the connection to be used when filing a document. In this example, we choose a predefined P8 4.x connector connecting to a P8 content object store.</td>
</tr>
</tbody>
</table>
| File in Folder Options    | ▶ Folder path: In the Folder Path text box, provide the complete path to the folder in FileNet P8. In this example, because the IBM Content Collector task route is filing the document into the Most Relevant Category folder in FileNet P8 based on the analysis by Classification Module, the folder path is set with the metadata type of Classification Module and the property of Most Relevant Category.  
   ▶ Select to create the folder if it does not exist. |

Figure 7-54 on page 406 shows the settings for the P8 4.x File Document in Folder task node for Score >= 70% rule.
4. To configure the other P8 4.x File Document in Folder task node, from the Toolbox pane, under FileNet P8 4.x Repository, select **Create P8 4.x File Document to Folder**, as shown in Figure 7-55 on page 407.
5. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the Decision Point and End nodes for the Rule link (Score < 70%) until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. Click the newly created task node to configure the task node properties.
6. Configure the P8 4.x File Document in Folder task with the settings that are described in Table 7-15.

Table 7-15   Settings for the P8 4.x File Document in Folder task node

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connections drop-down list, select the connection to be used when filing a document. In this example, we choose a predefined P8 4.x connector connecting to a P8 content object store.</td>
</tr>
</tbody>
</table>
| File in Folder Options    | ▶ Folder path: In the Folder Path text box, provide the complete path to the folder in FileNet P8. In this example, because the IBM Content Collector task route is filing the document into the Most Relevant Category folder in FileNet P8 based on the analysis by Classification Module, the folder path is set to a fixed location for Manual Review.  
▶ Select to **create the folder if it does not exist**. |

Figure 7-56 on page 409 shows the settings for P8 4.x File Document in Folder task node for Score < 70% rule.
At this point, we have completed the creation of two P8 4.x File Document in Folder task nodes. In the next section, we create the next task node for our task route.
Create EC prepare E-Mail for Stubbing task node

This task prepares the e-mail for stubbing. In this section, we create an EC prepare E-Mail for Stubbing task node. Follow these steps to add and configure an EC Prepare E-Mail for Stubbing task node in your task route:

1. From the Toolbox pane, under E-Mail Server, select **EC Prepare E-Mail for Stubbing**, as shown in Figure 7-57.

![Figure 7-57 Adding EC Prepare E-mail for Stubbing task node](image-url)
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the P8 4.x File Document in Folder and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. Click the newly created task node to configure the task node properties.

3. Add the EC Prepare E-Mail for Stubbing task node. Enter a name and a description under the General tab.

   We set the EC Prepare E-Mail for Stubbing task node settings, as shown in Figure 7-58 on page 412, for our example.
4. We now need to add a link from the P8 4.x File Document in Folder from the Score >= 70% Rule link to the newly created EC Prepare E-mail for Stubbing task node. To create a new link, click the link icon in the Toolbox pane. You will see a new link (blue color), as shown in Figure 7-59 on page 413.
Drag and stretch the blue link to connect the P8 4.x File Document in Folder task node to EC Prepare E-Mail for Stubbing task node, as shown in Figure 7-60 on page 414.
Figure 7-60   New link connecting P8 4.x File Document in Folder task node to EC Prepare E-Mail for Stubbing task node

At this point, we have completed the creation of the Prepare E-mail for Stubbing task node. In the next section, we create the next task node for our task route.
Create EC Create E-Mail Stub task node

This task turns the original E-Mail into a stub document. That is, it removes content according to the settings in the task itself or according to the settings in a life cycle collector. In this section, we create an EC Create E-Mail Stub task node. Follow these steps to add and configure an EC Create E-Mail Stub task node in your task route:

1. From the Toolbox pane, under E-mail Server, select **EC Create E-Mail Stub**, as shown in Figure 7-61.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the EC Prepare E-Mail for Stubbing and End nodes until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. Click the newly created task node to configure the task node properties.

3. Add the EC Create E-Mail Stub task node. Enter a name and a description under the General tab.

We set the EC Create E-Mail Stub task node settings, as shown in Figure 7-62 on page 417, in our example.
At this point, we have successfully completed the entire task route for our use case scenario. In the next section, we describe the procedure to activate this task route to automatically classify e-mails in the required folder.
7.4.5 Activate the integration system

In this section, we describe the process of activating the system to start automatic e-mail archiving with automatic classification. We describe the following steps:

- Activate the task route.
- Start the task routing engine service.
- Activate the audit log.

**Activate task route**

To activate the task route, start by highlighting the **EC Collect E-Mail By Rules** task note and make sure that **Active** is selected, as highlighted in Figure 7-63 on page 419.

**Note:** To deactivate the task route, clear the selection from the Active flag.
Figure 7-63  Active flag for task route

**Start the Task Routing Engine service**

Go to **Start → Control Panel → Administrative Tools → Services**. Find the IBM Content Collector Task Routing Engine Service and right-click it to start the service, as shown in Figure 7-64 on page 420.
Activate the audit log

To activate the audit log, click the **task node** icon in the task route, as shown in Figure 7-65 on page 421, and select **Include audit log task**.
Click the highlighted audit node in Figure 7-65 to set the correct audit level, as shown in Figure 7-66 on page 422.
Figure 7-66   Audit Log Node General Properties
7.4.6 Verify the automatic classification and archiving results

Now that we have activated the integration system, we will verify whether our use case results have been achieved. To verify the automatic classification and archiving results, we create three e-mails in the Exchange mailbox:

1. E-mails consisting of Stock Options and Pay-related concepts or subjects
2. E-mail consisting of unknown category-related subject

E-mails consisting of Stock Options and Pay-related concepts

Follow the steps that are described in this section to verify the automatic classification and archiving of e-mails consisting of valid concepts from the knowledge base, such as *Stock Options* and *Pay*:

1. Use the Outlook E-Mail client to create a new e-mail using the Stock Options subject. In this example, we create a new e-mail with the following settings:
   - To: L2user3
   - Subject: Stock Options
   - Body of e-mail: When do my stock options expire?

2. Use Microsoft Outlook client to send the e-mail.

3. In a browser window, launch the Workplace application and browse to the appropriate object store. In this example, because the e-mail content is categorized under the Redbook/HR/Stock Options category by the Classification Module task node as the top category, a new folder under Object Stores/CE_OS is created and highlighted, which is called Redbook/HR/Stock Options. The e-mail is stored as a document called Stock Options, as highlighted in Figure 7-67 on page 424.
4. The IBM Content Collector task route is configured so that, after the e-mail is classified by the Classification Module and archived into IBM FileNet P8 repository under the appropriate folder, a stub in the Microsoft Outlook e-mail client is left behind. In this example, verify that the Outlook client e-mail body is now replaced by a Link to archived E-mail body stub.

5. Similarly, you can verify the classification by sending an e-mail with Pay as the subject and see that the e-mail gets classified and archived into the Redbook/HR/Pay category. In this example, we already have sent an e-mail with Pay Scale as the subject, and the e-mail is already classified and archived into the correct IBM FileNet P8 folder, that is, Object Stores/CE_OS/Redbook/HR/Pay, as shown in Figure 7-68 on page 425.
E-mail consisting of unknown category-related subject

To work with e-mails consisting of non-matched categories, follow these steps:

1. Use the Outlook E-Mail client to create a new e-mail using no relevant concept or subject from the knowledge base. In this example, we create a new e-mail with the following information:
   - To: L2user3
   - Subject: Not Categorized

2. Use Outlook client to send and receive the e-mail.

3. In a browser window, launch the Workplace application and browse to the appropriate object store. In this example, because the e-mail content does not match any existing category in the knowledge base according to the Score < 70% rule, these types of documents will be placed under the Redbook/manualReview category by the Classification Module task node. A new folder under Object Stores/CE_OS is created and highlighted, which is called Redbook/manualReview. The e-mail is stored as a document called Not Categorized, as highlighted in Figure 7-69 on page 426.
4. The IBM Content Collector task route is configured so that, after the e-mail is classified by the Classification Module and archived into IBM FileNet P8 repository under the appropriate folder, a stub in the Outlook e-mail client is left behind. Verify that the Outlook client e-mail body is now replaced by a Link to archived E-mail body stub.

7.5 Use Case 2: E-mail classification with records management

In this use case, we create a simple scenario in which e-mails from an existing IBM FileNet P8 content object store are used to create a knowledge base. IBM Content Collector is then used with the Classification Module task route pointing to the HR knowledge base to automatically classify the newly ingested e-mails and to declare them as records.

Using the HR knowledge base, create a new Classification Module task route so that the ingested e-mail document gets automatically classified into the appropriate content object store and filed as a record with the correct retention and disposition schedule into the appropriate file plan object store. We will use the same HR knowledge base that we used for 7.4, “Use case 1: E-mail archiving
These steps are an overview of the steps that involved in implementing this use case:

1. Create the knowledge base.
2. Verify the integration settings.
3. Create an e-mail task route.
4. Activate the task route.
5. Verify the use case end results.

The steps listed here are exactly the same steps that we described in 7.4, “Use case 1: E-mail archiving with automatic classification” on page 352; however, we need to modify the task route with the addition of the P8 4.x Declare Record task node. In this section, we describe how to modify the task route that you created earlier in order to add the records declaration capability. We also verify the use case end results on the Exchange Outlook client and the IBM FileNet P8 records management folder. We show you how to modify the existing task route and to verify the record declaration results.

7.5.1 Modify the existing task route to add the P8 4.x Declare Record task node

This task allows you to configure the required properties to declare a record in IBM FileNet Records Manager. Follow these steps to add and configure a P8 4.x Declare Record task node in your task route:

1. From the Toolbox pane, under FileNet P8 4.x Repository, select P8 4.x Declare Record, as shown in Figure 7-70 on page 428.
2. Click anywhere in the Designer pane (white space) to drop the new task onto the task route. After you have dropped the task onto the task route, you can drag it over the arrow connecting the P8 4.x File Document in Folder and EC Prepare E-Mail for Stubbing on the Score >= 70% Rule Link, until the task and arrow are highlighted and the task icon has an exclamation point (!) in a yellow triangle. Click the newly created task node to configure the task node properties.
3. Add the P8 4.x Declare Records task node, and configure it with the settings that are shown in Table 7-16.

Table 7-16  Settings for the P8 4.x Declare Record task node

<table>
<thead>
<tr>
<th>Configuration section</th>
<th>Detailed settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Enter a name and a description for the task.</td>
</tr>
<tr>
<td>P8 Connection</td>
<td>From the Connection drop-down list, select the connection to use when creating a document. In this example, we choose a predefined P8 4.x connector connecting to a FileNet P8 content object store.</td>
</tr>
<tr>
<td>Property Mappings</td>
<td>Select the record class that you want to use when creating the record, and enter the classification path for the record declaration. In this example, we select the record class called ElectronicRecord, for Classification/RecordsManagement/File Plan/Email and map the Document Title, From and Sent On property with the Email Subject, From, and Sent On metadata values.</td>
</tr>
</tbody>
</table>

4. Select the Classification Path by clicking Add under the Property Mappings group. A P8 Classification dialog appears. In this example, we select the P8 Classification path, as shown in Figure 7-71.

Figure 7-71  P8 4.x Declare Record Classification folder selection

Figure 7-72 on page 430 shows the P8 4.x Declare Record task node settings for our example.
At this point, we have successfully completed the modification of the existing task route to add the P8 4.x Declare Record task node for our use case scenario. In the next section, we activate this task route and verify the declare records results.
7.5.2 Verify record declaration results

After modifying the task route as explained in 7.5.1, “Modify the existing task route to add the P8 4.x Declare Record task node” on page 427, follow 7.4.5, “Activate the integration system” on page 418 to activate the integration. Follow 7.4.6, “Verify the automatic classification and archiving results” on page 423 to verify the automatic classification and archiving results.

To verify the records declaration, use the Records Manager browser interface and browse to the File Plan Object Store location used in the Classification path. In this example, we have used the same two existing e-mails with the subjects Stock Options and Pay Scale to declare records. The two documents are now declared as records under IBM Records Manager Email File Plan, as highlighted and shown in Figure 7-73.
**Note:** When working with the IBM FileNet P8 repository, you can use the Classification Center to review the documents in the review folder and take necessary actions to tune and improve the accuracy of your knowledge base.
Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks publications

For information about ordering these publications, see “How to get IBM Redbooks publications” on page 434. Note that some of the documents referenced here might be available in softcopy only.

- Working with IBM Records Manager, SG24-7389
- Content Manager Implementation and Migration Cookbook, SG24-7051
- Understanding IBM FileNet Records Manager, SG24-7623
- IBM FileNet Content Manager Implementation Best Practices and Recommendations, SG24-7547
- Introducing IBM FileNet Business Process Manager, SG24-7509
- Introducing OmniFind Analytics Edition: Customizing Text Analytics, SG24-7568 (The product is currently known as IBM Content Analyzer)

Online resources

These Web sites are also relevant as further information sources:

- IBM InfoSphere Classification Module Version 8.7 Information Center
  http://publib.boulder.ibm.com/infocenter/classify/v8r7/
- IBM Classification Module Version 8.6 publication library
  http://www.ibm.com/support/docview.wss?rs=3376&uid=swg27012760
- IBM Classification Module Web site
  http://www.ibm.com/software/data/enterprise-search/classification
- IBM Enterprise Content Management Web site
  http://www.ibm.com/software/data/content-management
How to get IBM Redbooks publications

You can search for, view, or download IBM Redbooks publications, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks, at this Web site:

ibm.com/redbooks

To get a list of IBM Enterprise Content Management-related IBM Redbooks publications, enter ecmredbooks as the search string at the Web site.

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services
Index

A
action
decision plan, rule 26
active view 71, 80, 108
AddOn 188, 190–191
Open XML file 190
Administration process
Classification Module server component 31
analysis
semantic 22
analysis data 76, 105–106
application programming interface (API) 49
automatic classification 17, 39, 178, 180, 187, 276, 342, 352
E-mail archiving 418

B
business use cases 9

C
Categories 22
categorization 23
Classification Center 17, 29, 34, 50, 178, 180, 183, 273, 293, 335, 342–343, 346, 355, 432
classification results 283
document review filter 277
IBM FileNet P8 integration 216
review setup 180
business use cases 9
classification capabilities 296
combined strength 342
configuration process 7
correct classification 91, 230
IBM Content Collector 352
input 6
integration options 16
knowledge base 37
Many applications 16
metadata type 16
output 6
second rule 327
setup connectivity 198
supported content repository 16
unique capability 102
Classification Module Server 29
Classification Module server 22, 29, 344–346
system architecture 30
classification process 1, 10–11, 34, 50, 249, 251–252, 261, 263–264
classification scheme 7
Classification server
administration 40, 44
classification technology 25
CSV file 101
Decision Plan project 234
export wizard 236
extracted content XML output 57
field definition settings 28
file system folder 64
file system folder structure 61
import content set 30
Import documents 353
saved XML data 283
Client APIs 30
compliance 10
use case 9
concepts 22
configuration file 106, 198
Configuration Manager 297–298, 300, 307, 343, 348–349
Classification Module system metadata 348
extplorer pane 351
explorer pane right side 330
content
route 14
unstructured 7
unstructured, taxonomy 7
Content Extractor 29, 34, 50, 179–180, 205, 352
Content field
definition administration 40, 43
values section 98
content field 69, 90, 95, 221–222, 317, 395
IBM Content Collector 18
folder structure 61, 185, 272, 302
representative sample files 61
FSC Collector 310–312
configuration tabs 312

group
rule 26

H
host name 44, 201, 269, 317, 367, 372, 374
HR knowledge base 61, 77, 79, 90, 101, 303, 353–354, 375, 426
keyword CSV file 101

I
IBM Content Collector 18
  Classification Module 340, 342
  Integration 18
IBM Content Collector (ICC) 34, 175, 291–292, 294, 339–341
  administration application 187
  administrator Id 255, 273
  Classification Module 34, 50
  Classification Module integration components 254
  Classification Module v8.6 integration component 183
  connectivity 344, 346
  Connector component 50
  Connectors 364
customer 17
document classes 53
document object 389
document property fields 261
e-mail systems 18
folder structure 185
Integration 17
integration 17, 89, 177, 183–184
integration component 183
integration components 183
offering 17
parameter 187
right folder structure 40
sample content 181
use Classification Module 182
IBM FileNet P8 document
class 210, 221, 272
property 251, 261
IBM FileNet P8 integration asset 29
ICC Configuration Manager 297–298, 300
  Classification Module system metadata properties 297
  Classification Module task 298
ICC task route 292–295, 299
  Record task 323
imported content set
decision plan 96
input 6
integration options 16

K
Knowledge base
  basic elements 22
certain set 41
  instance pooling 31
load-balance read-write processes 47
read-only instance 46
read-write instance 32
knowledge base 23
backup copies 303
Category names 215, 270
category names 185
Classification Module works 204
direct impact 103
existing category 425
immediate, visual sense 79
optimize accuracy 30
optimized 29
previous version 303
project type 63
read/write and read-only instances 303
relevant concept 425
runtime options 303
statistical models 276
textual analysis capabilities 8
valid concepts 423
knowledge base (KB) 1, 5–6, 8, 22–24, 34, 39, 45, 59–60, 82, 179–180, 185, 246–247, 293–295, 342–343, 346, 354, 395, 432
L
learning 23
learning process 22
left pane 103, 197, 219, 243, 300, 351
decision plan name 248
Listener
  Classification Module server component 31
Lotus Quickr 15

M
mail
  routing, use case 15
Management Console 29–30
matching 23
Messaging Application Program Interface (MAPI) 346
metadata 6, 20, 180–181, 205, 292, 297, 299, 340–341, 348
  ECM system 20
metadata field 6
Metadata type 299–300, 320, 350–351, 399
Model-View-Controller (MVC) 50
Module server (MS) 346

N
natural language processing
  Configure fields 36
textual content 36
Natural Language Processing (NLP) 25
natural language processing (NLP) 5, 22, 25, 36, 261
Navigation Pane
  click Connectors 364
  click Data Stores 362
  Connectors tab 365
  E-mail Services tab 370
  Lists icon 369
next section 179, 387, 389, 392
NLP Engine 25
non-matched category 425

O
object store 180, 185, 205, 332, 334, 352–353, 364
  Document base class 195
document classes 205
  IBM FileNet P8 Connections 364
  imported properties 194
  online feedback 23
  online learning 23
  optical character recognition (OCR) 6
  original e-mail 415
  output 6

R
Read/only process
  Classification Module server component 31
Read/write process
  Classification Module server component 31
Read-only (R/O) 46
read-only instance 32, 48, 303
read-only request 32, 47
Read-write (R/W) 46
read-write instance 32, 46
real life 23, 61
Recall graph 76–77, 79, 151
  Total Precision 79
records declaration 11
Records management
  system 10, 26, 292
records management 15
Redbooks Web site 434
  Contact us xxv
related knowledge base
  same check 252
relevancy 23
relevancy score 23, 73, 293, 317, 342, 395, 397
relevant category 299, 322, 350, 396, 399
Relevant Score 299, 318–319, 327, 350
report management
  use case 16
representative e-mails 343
return on investment (ROI) 9
review 8
review folder 216, 218, 225, 284, 287, 335, 432
right pane 44, 103, 188, 197, 217, 300, 351
  Select one 244
    server name 44
root folder 66
    fully qualified path 66
route content 14
rule
  group 26
Rule 2 87, 90–91
rules
decision plan 26
Runtime setting 54, 251, 267–268

S
score 23
content 14
select_branch KB
 score result 228
selected knowledge base
category scores 222
semantic analysis 22
Semantic Modeling Engine 25
Semantic Modeling Language (SML) 25
Simple Object Access Protocol (SOAP) 32
SOAP Layer 30
supported ECM repository
e-mail instance 340
system architecture
 Classification Module server 30
System metadata 299–300, 315, 318, 350–351, 369

T
Active flag 418
FSC Post Processing task 325
genral properties 376
later point 389
left branch 320–321
metadata properties 315
new FSC Collector task 310
new task 313, 316
Record task 323
right branch 327
task node icon 420
task route decision 19
taxonomy 7, 29
generation 13
Taxonomy Proposer 29–30, 37–38, 100, 353
 Import documents 353
 pictorial representation 37
toolbar 71, 76, 80
Toolbox pane 298, 310, 312, 349, 359, 377
 link icon 412
training process 22
triggers
decision plan, rule 26

U
unfile document 89
unstructured content 7
upper-right pane 219
use case 9, 15, 17, 177, 180–181, 185, 222, 291, 294, 301, 339, 352–353
use cases
 business 9
 compliance 9
 mail routing 15
 report management 16

W
white space 313, 385–386, 388
Workflow Assistant
 output 38
 window 217
WSDL 33, 49

X
XML format 179, 283, 286
click Save Document 286
Save document 286
XML output
directory 205, 207
file 107, 205
XmlDirectory property 207

Y
yellow triangle 313, 385–386, 388, 428
exclamation point 313
IBM Classification Module Version 8.6 is an advanced enterprise software platform tool designed to allow organizations to automate the classification of unstructured content. By deploying the module in various areas of a business, organizations can reduce or avoid manual processes associated with subjective decision making around unstructured content and streamline the ingestion of that content into their business systems in order to use the information within their organization. At the same time, they can safely remove irrelevant or obsolete information and therefore utilize their storage infrastructure more efficiently. By reducing the human element in this process, IBM Classification Module ensures accuracy, consistency, and enables auditing while simultaneously driving down labor costs.

This IBM Redbooks publication explains what IBM Classification Module Version does, the key concepts to understand, and its integration with other products and systems. With this book, we show you how IBM Classification Module can help your organizations to automate the classification of large volumes of unstructured content in a consistent and accurate manner. We also cover several of the major use cases for IBM Classification Module and show you how to implement each use case.

This book is intended to educate both technical specialists and nontechnical personnel in how to make IBM Classification Module work for your organizations.