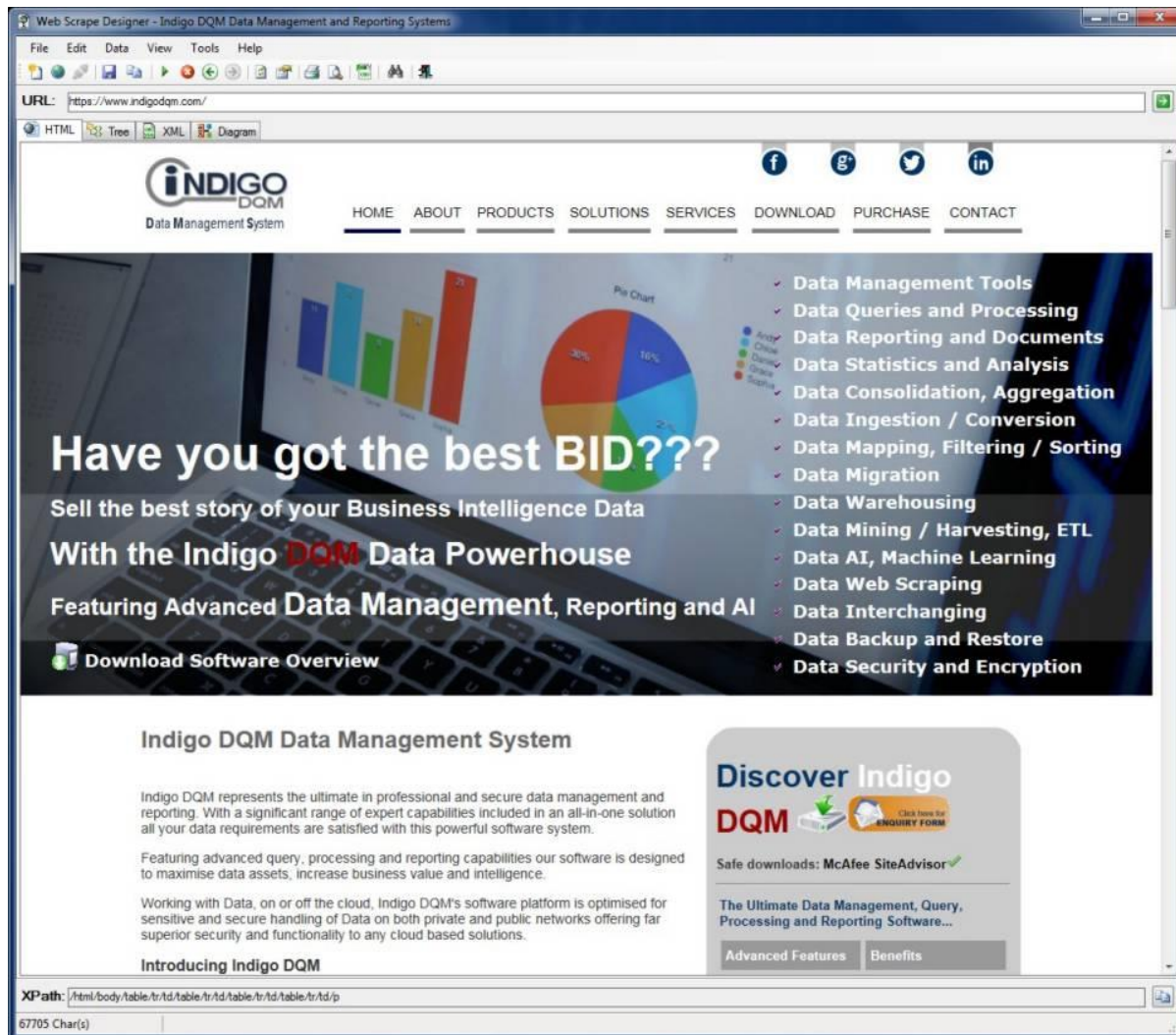


Indigo DQM Data Management System

Web Scrape Designer

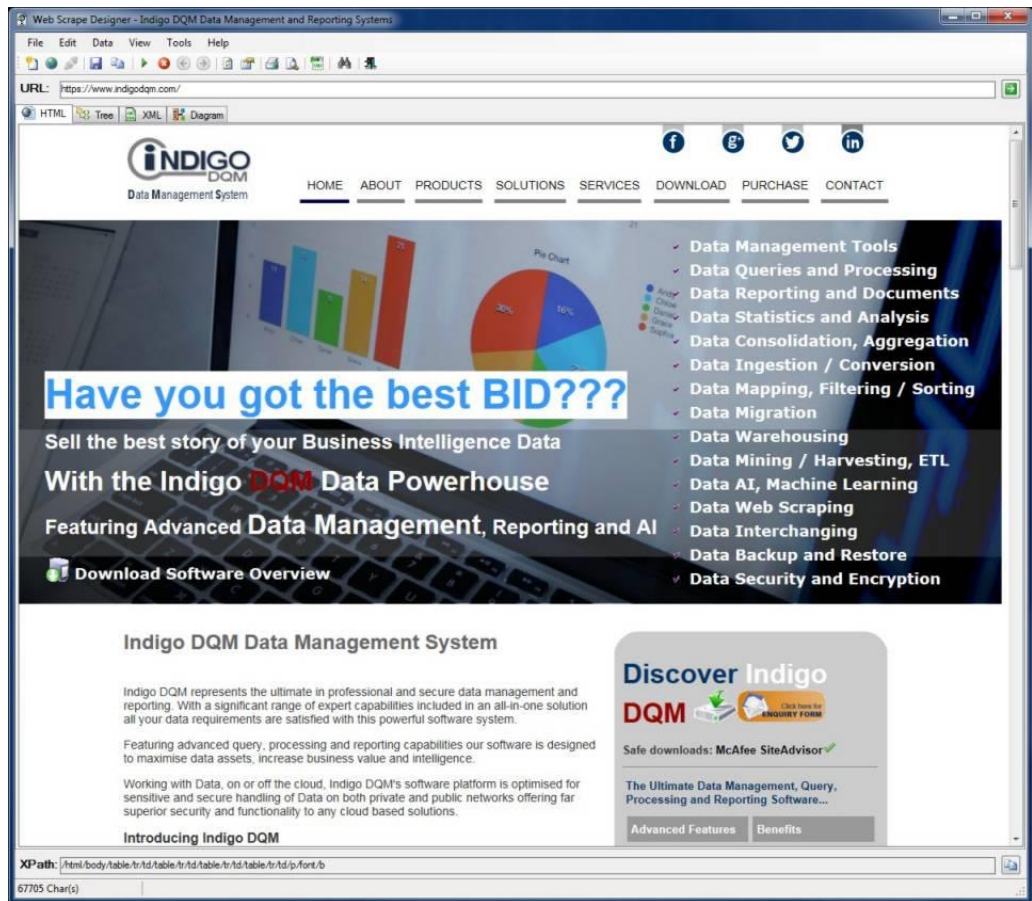
Indigo DQM Web Scrape Designer allows complex XQueries and XPath statements to be executed to extract or scrape elements from HTML Web Pages or Files.

XQuery is a query and functional programming language that is designed to query and transform collections of structured and unstructured Data, usually in the form of XML (Extensible Markup Language).

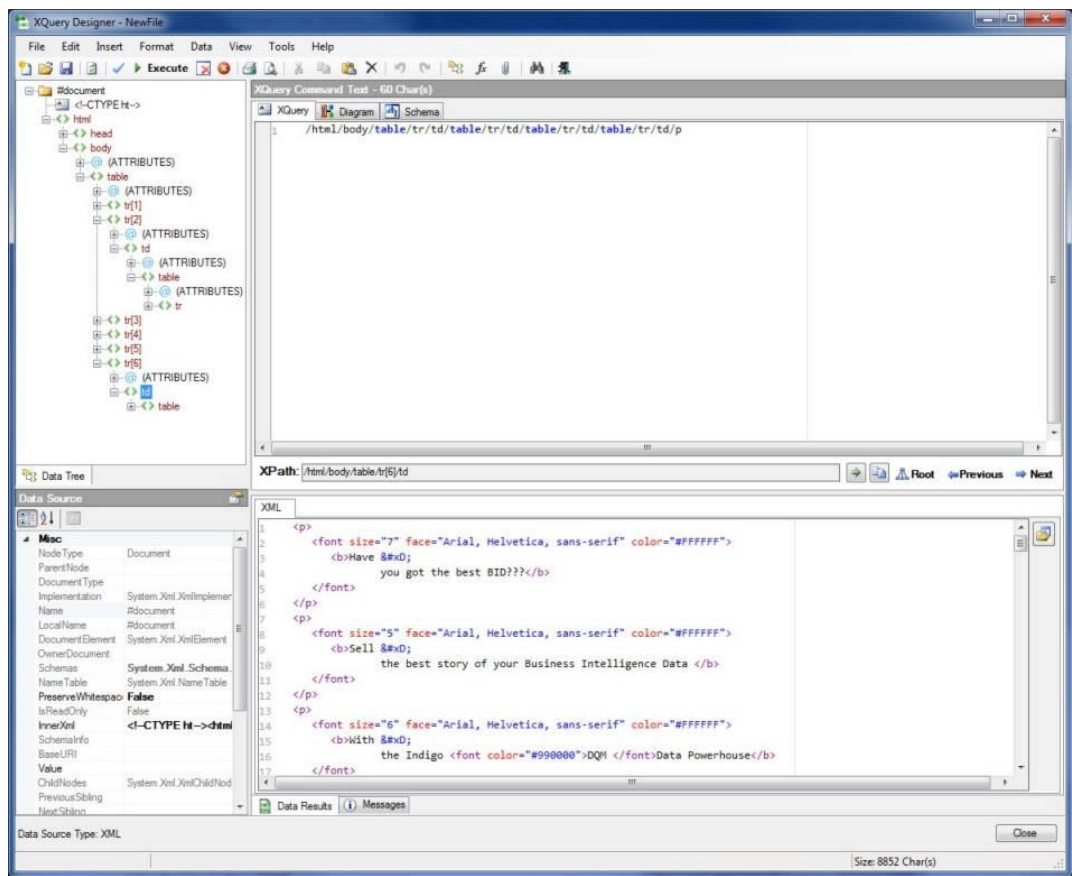


The language is based on the XQuery and XPath Data Model (XDM) which uses a tree-structured model of the information content of an XML document.

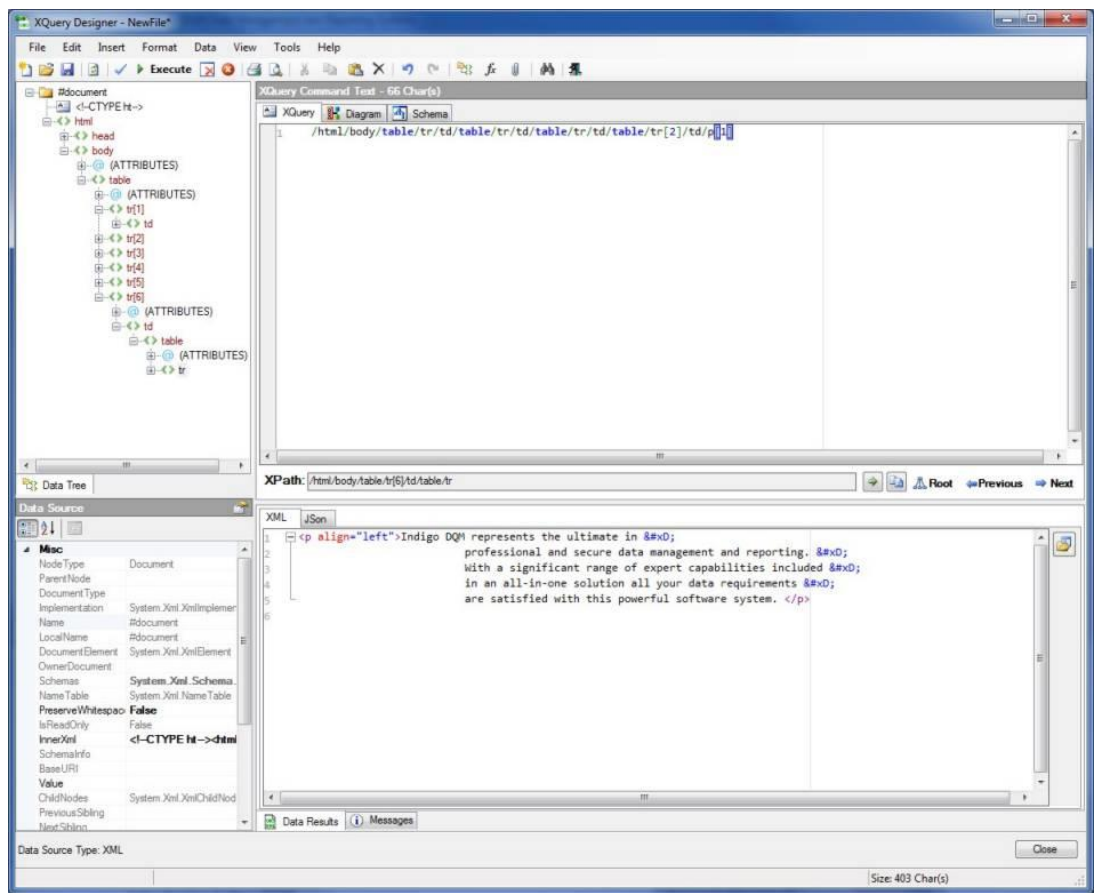
The Web Scrape Designer is a visual aid that allows web page elements to be clicked, selected or highlighted to automatically generate XPath statements.



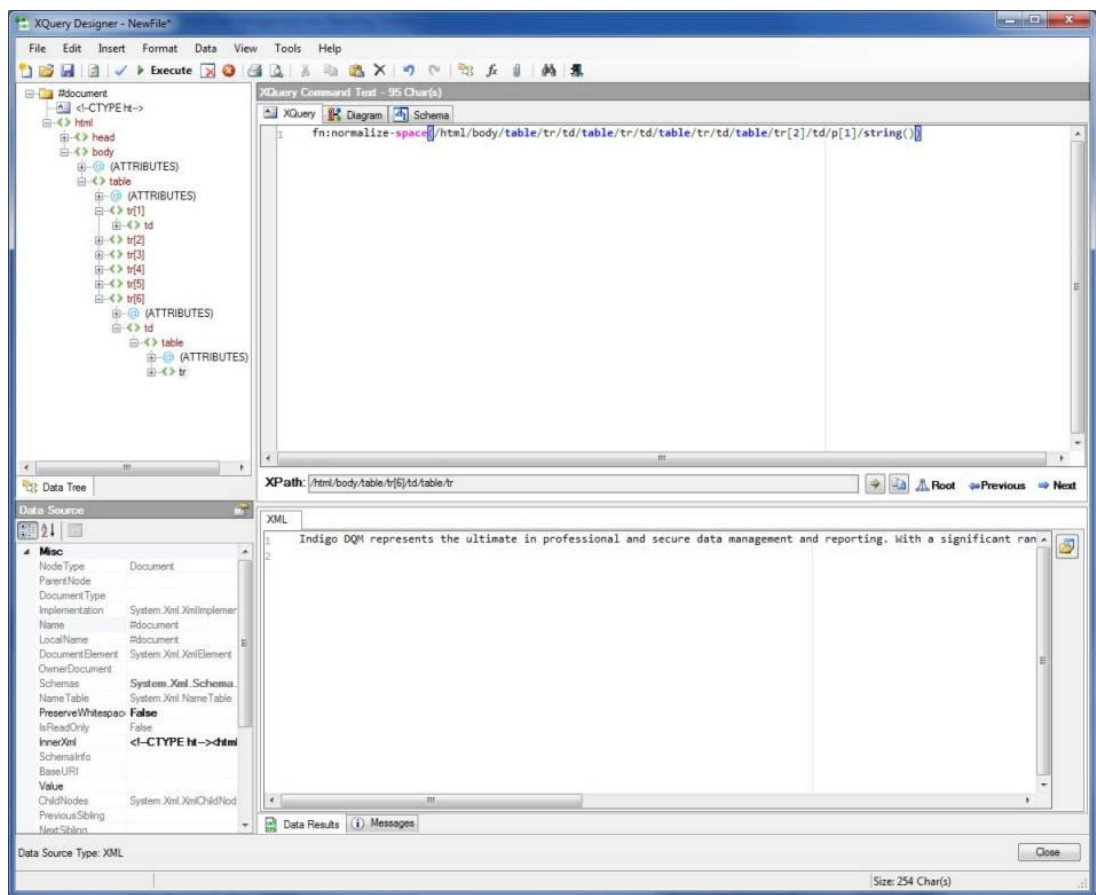
Web page data can be queried using XQuery Designer.



Extracting HTML elements from the Web Page using XQuery

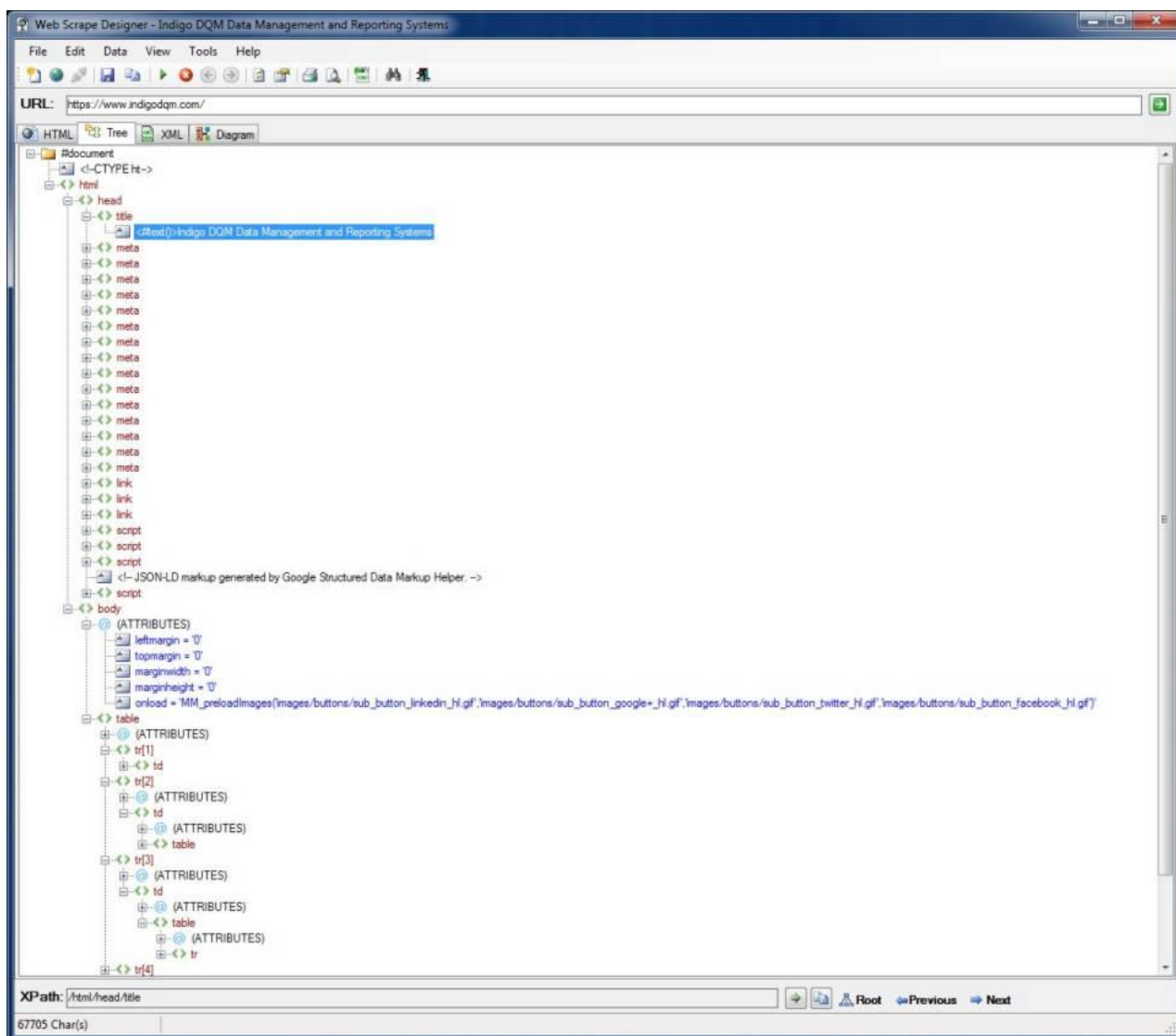


Using an XQuery function to normalize space and extract the plain text.



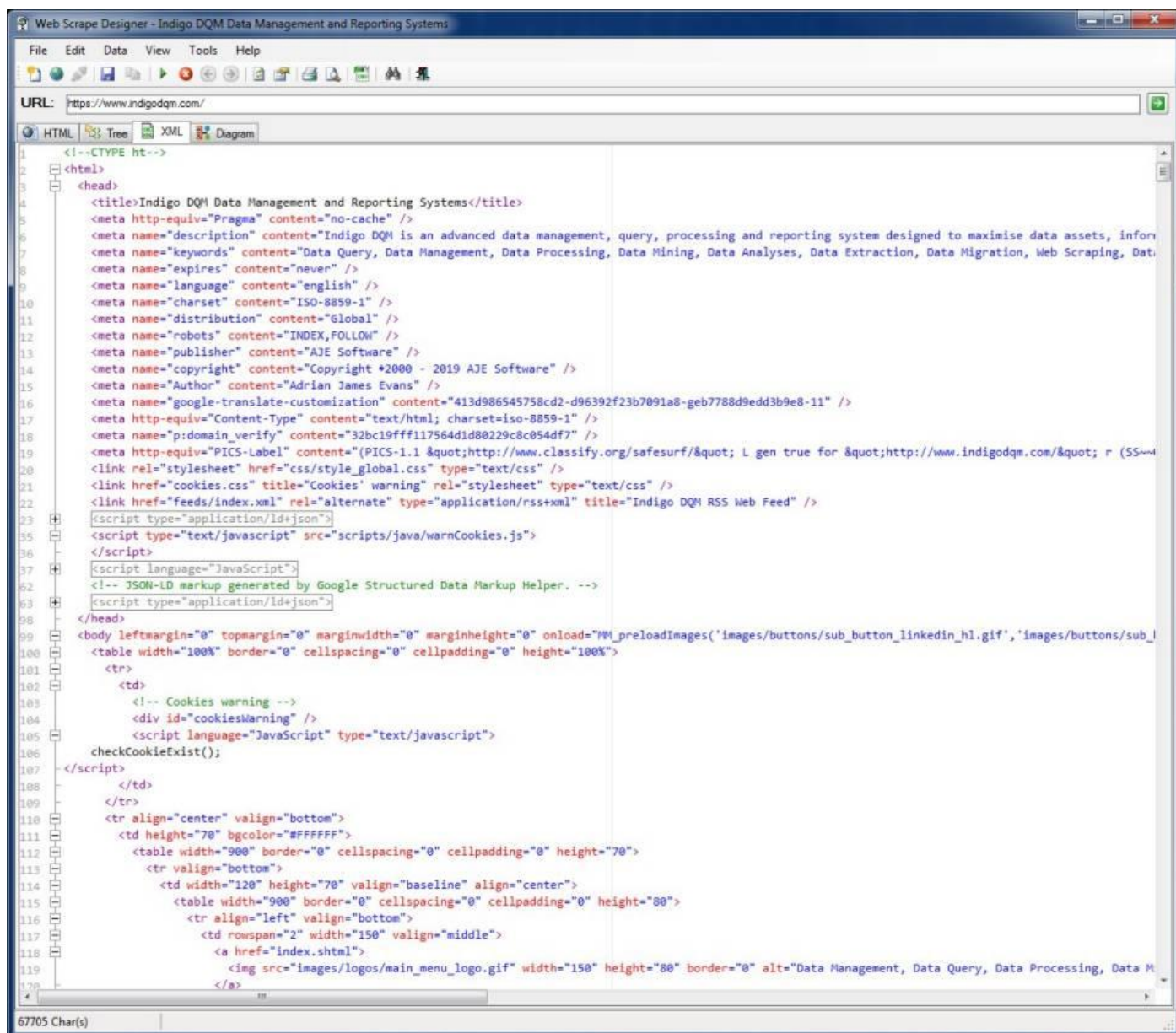
Data Tree View

Selecting a node in the Data Tree will update the current XPath for that node.



Web Page Source

Viewing the HTML to XML for the Web Page Source.

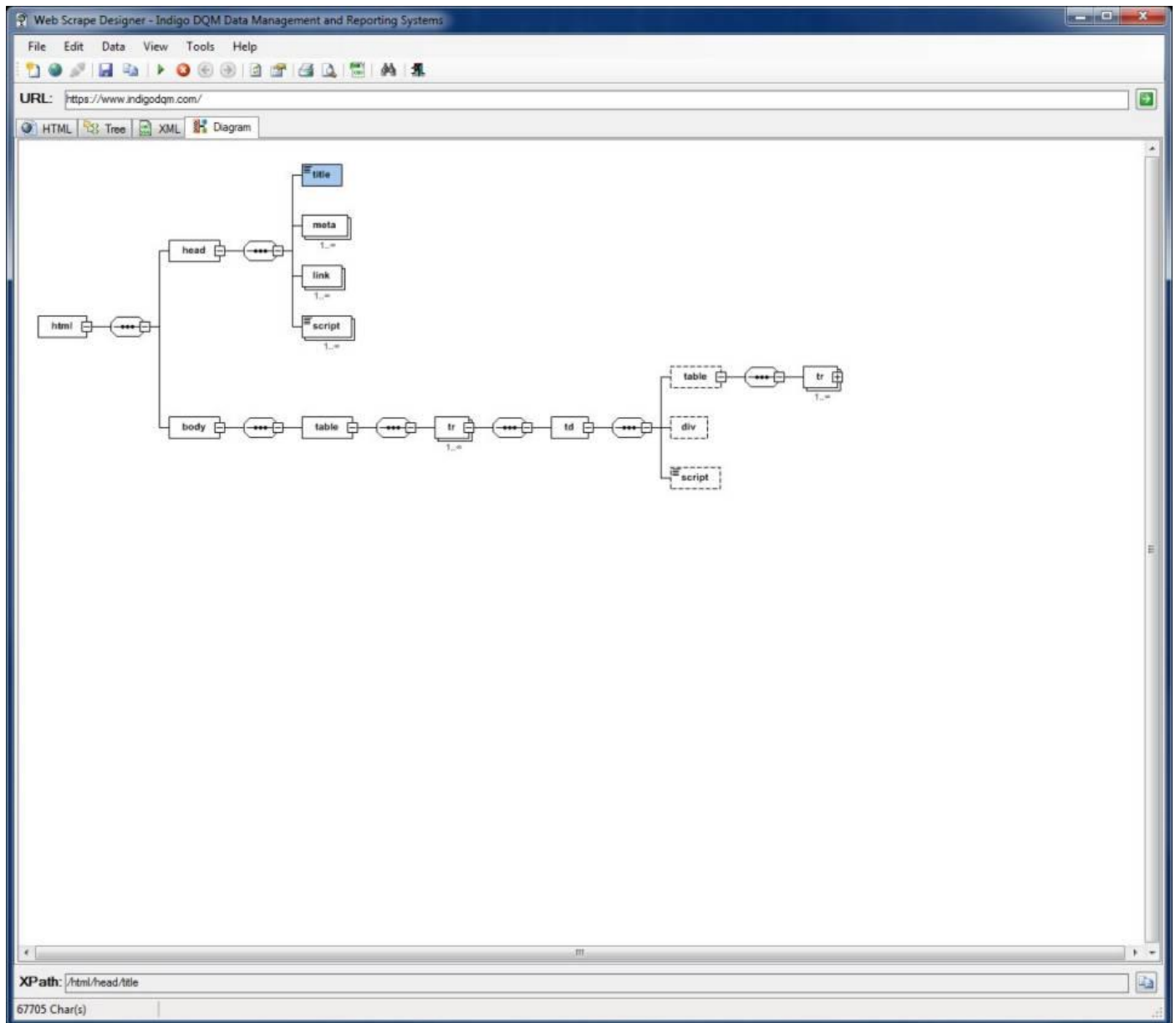


The screenshot displays the 'Web Scrape Designer' application window. The title bar reads 'Web Scrape Designer - Indigo DQM Data Management and Reporting Systems'. The menu bar includes 'File', 'Edit', 'Data', 'View', 'Tools', and 'Help'. The toolbar contains various icons for file operations and viewing. The 'URL' field shows 'https://www.indigodqm.com/'. Below the toolbar, there are tabs for 'HTML', 'Tree', 'XML', and 'Diagram', with 'HTML' currently selected. The main text area shows the raw HTML source code of the webpage, with line numbers 1 through 120 visible on the left. The code includes a DOCTYPE declaration, an HTML tag, a head section with various meta tags (title, description, keywords, expires, language, charset, distribution, robots, publisher, copyright, author, google-translate-customization, http-equiv, domain-verify, PICS-label, stylesheet, cookies, feeds/index.xml), and a body section with a table layout. The status bar at the bottom indicates '67705 Char(s)'.

```
1 <!DOCTYPE html>
2 <html>
3 <head>
4 <title>Indigo DQM Data Management and Reporting Systems</title>
5 <meta http-equiv="Pragma" content="no-cache" />
6 <meta name="description" content="Indigo DQM is an advanced data management, query, processing and reporting system designed to maximise data assets, inform
7 <meta name="keywords" content="Data Query, Data Management, Data Processing, Data Mining, Data Analyses, Data Extraction, Data Migration, Web Scraping, Data
8 <meta name="expires" content="never" />
9 <meta name="language" content="english" />
10 <meta name="charset" content="ISO-8859-1" />
11 <meta name="distribution" content="Global" />
12 <meta name="robots" content="INDEX,FOLLOW" />
13 <meta name="publisher" content="AJE Software" />
14 <meta name="copyright" content="Copyright ©2000 - 2019 AJE Software" />
15 <meta name="author" content="Adrian James Evans" />
16 <meta name="google-translate-customization" content="413d986545758cd2-d96392f23b7091a8-geb7788d9edd3b9e8-11" />
17 <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
18 <meta name="p:domain_verify" content="32bc19fff117564d1d80229c8c054df7" />
19 <meta http-equiv="PICS-Label" content="(PICS-1.1 &quot;http://www.classify.org/safesurf/&quot;; L gen true for &quot;http://www.indigodqm.com/&quot;; r (SS=4
20 <link rel="stylesheet" href="css/style_global.css" type="text/css" />
21 <link href="cookies.css" title="Cookies' warning" rel="stylesheet" type="text/css" />
22 <link href="feeds/index.xml" rel="alternate" type="application/rss+xml" title="Indigo DQM RSS Web Feed" />
23 <script type="application/ld+json">
24
25 </script>
26 <script type="text/javascript" src="scripts/java/warnCookies.js">
27
28 </script>
29 <script language="JavaScript">
30
31 </script>
32 <!-- JSON-LD markup generated by Google Structured Data Markup Helper. -->
33 <script type="application/ld+json">
34
35 </script>
36 </head>
37 <body leftmargin="0" topmargin="0" marginwidth="0" marginheight="0" onload="MM_preloadImages('images/buttons/sub_button_linkedin_hl.gif','images/buttons/sub_b
38 <table width="100%" border="0" cellpadding="0" cellspacing="0" height="100%">
39 <tr>
40 <td>
41 <!-- Cookies warning -->
42 <div id="cookiesWarning">
43 <script language="JavaScript" type="text/javascript">
44
45 checkCookieExist();
46 </script>
47 </td>
48 </tr>
49 <tr>
50 <td align="center" valign="bottom">
51 <table width="900" border="0" cellpadding="0" cellspacing="0" height="70">
52 <tr align="bottom">
53 <td align="center" rowspan="2" width="120" height="70" valign="baseline" align="center">
54 <table width="900" border="0" cellpadding="0" cellspacing="0" height="80">
55 <tr align="left" valign="bottom">
56 <td rowspan="2" width="150" height="80" border="0" alt="Data Management, Data Query, Data Processing, Data M
57 <a href="index.shtml">
58 </a>
59 </td>
60 </tr>
61 </table>
62 </td>
63 </tr>
64 </table>
65 </td>
66 </tr>
67 </table>
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
```

XSD Diagrams

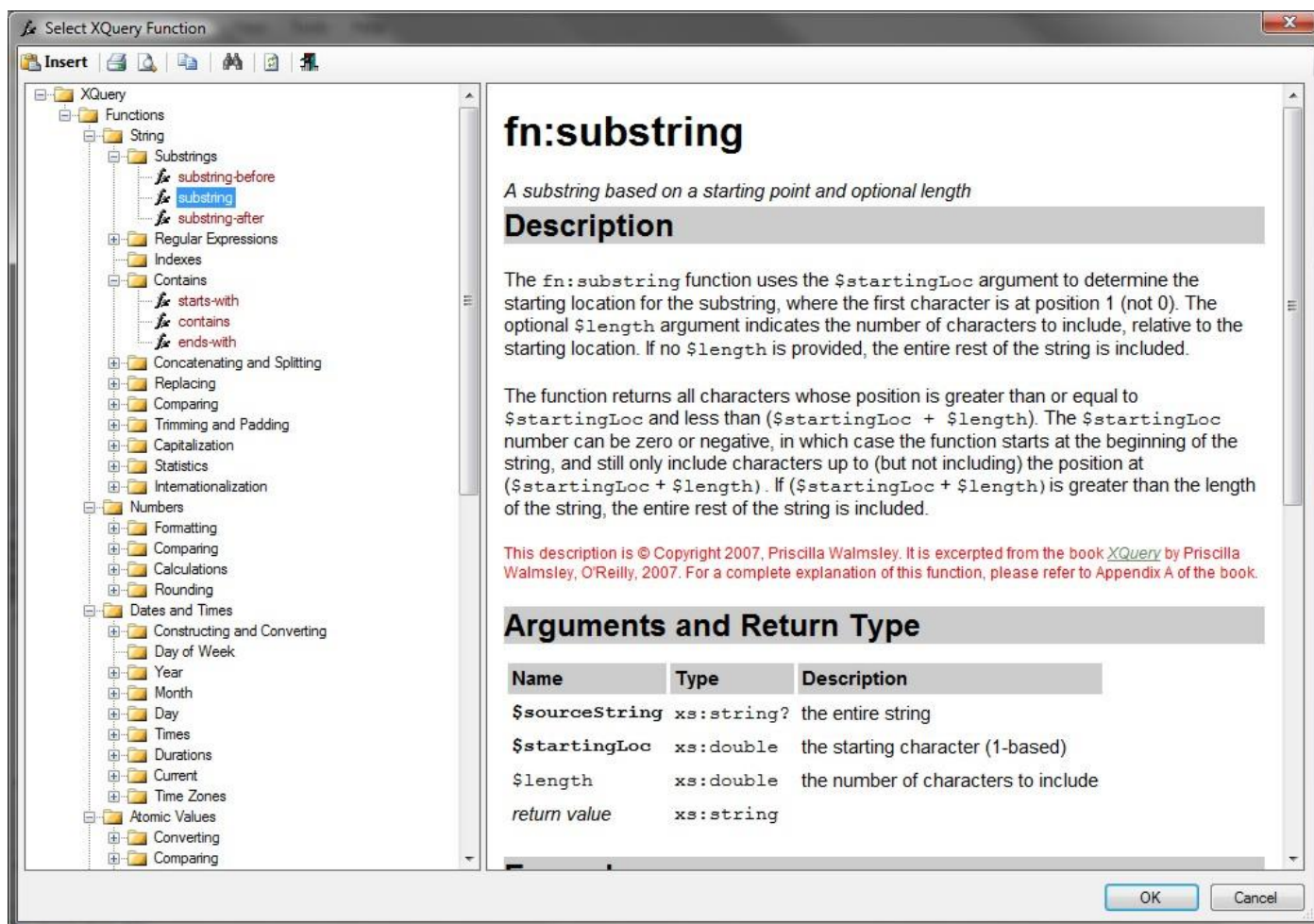
XSD Diagrams allow a visual representation of an XPath expression in the Data Schema. Click the Diagram tab and expand out the Diagram elements to show the structure of the Data Schema.



The XPath expression for the current element is shown in the XPath navigation bar. Elements can also be navigated using the navigation buttons.

Inserting an XQuery Function

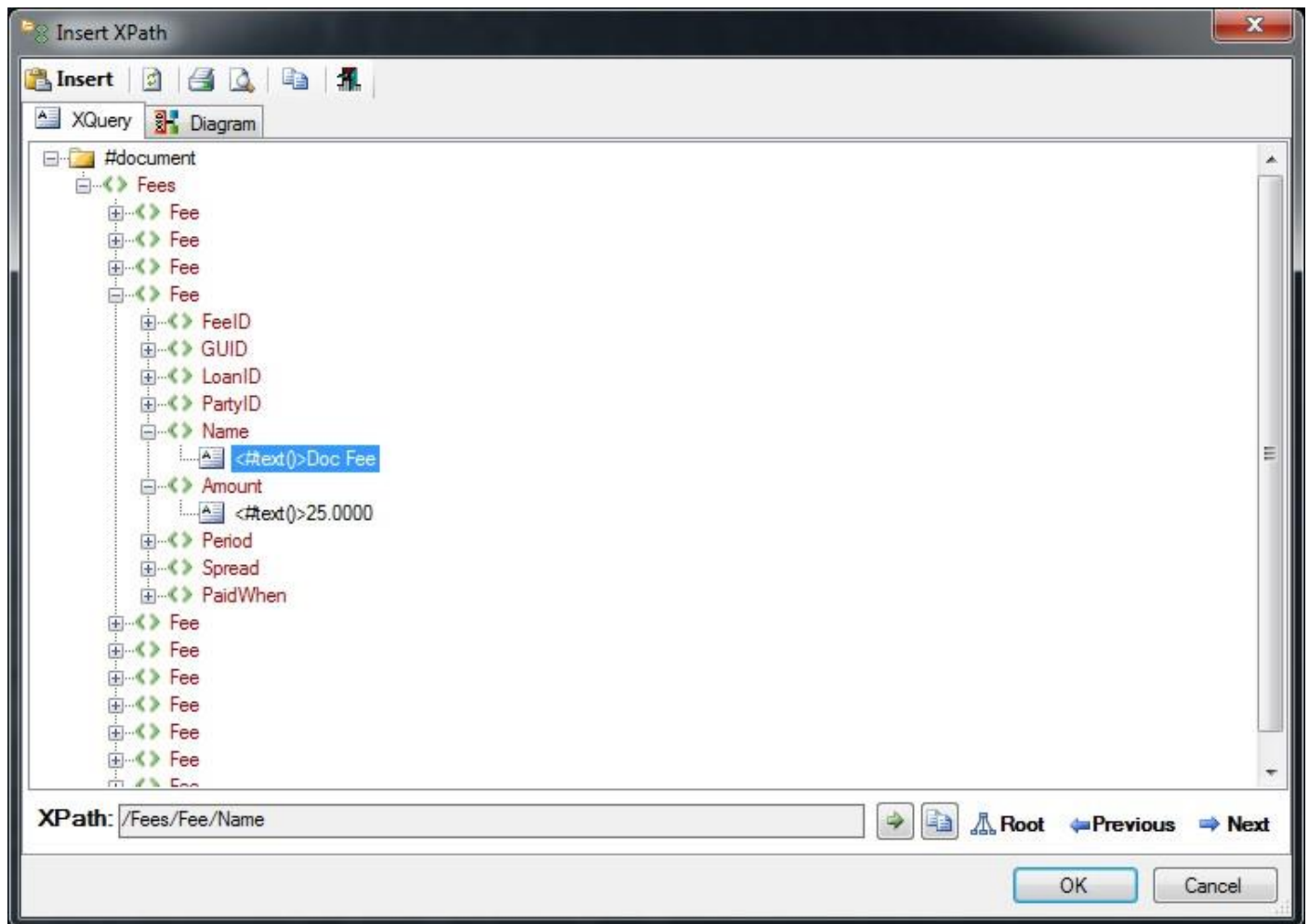
Predefined XQuery Functions can be inserted into the XQuery using the Function Tool.



XQuery contains a superset of XPath expression syntax to address specific parts of an XML document. The language is based on the XQuery and XPath Data Model (XDM) which uses a tree-structured model of the information content of an XML document.

Inserting an XPath

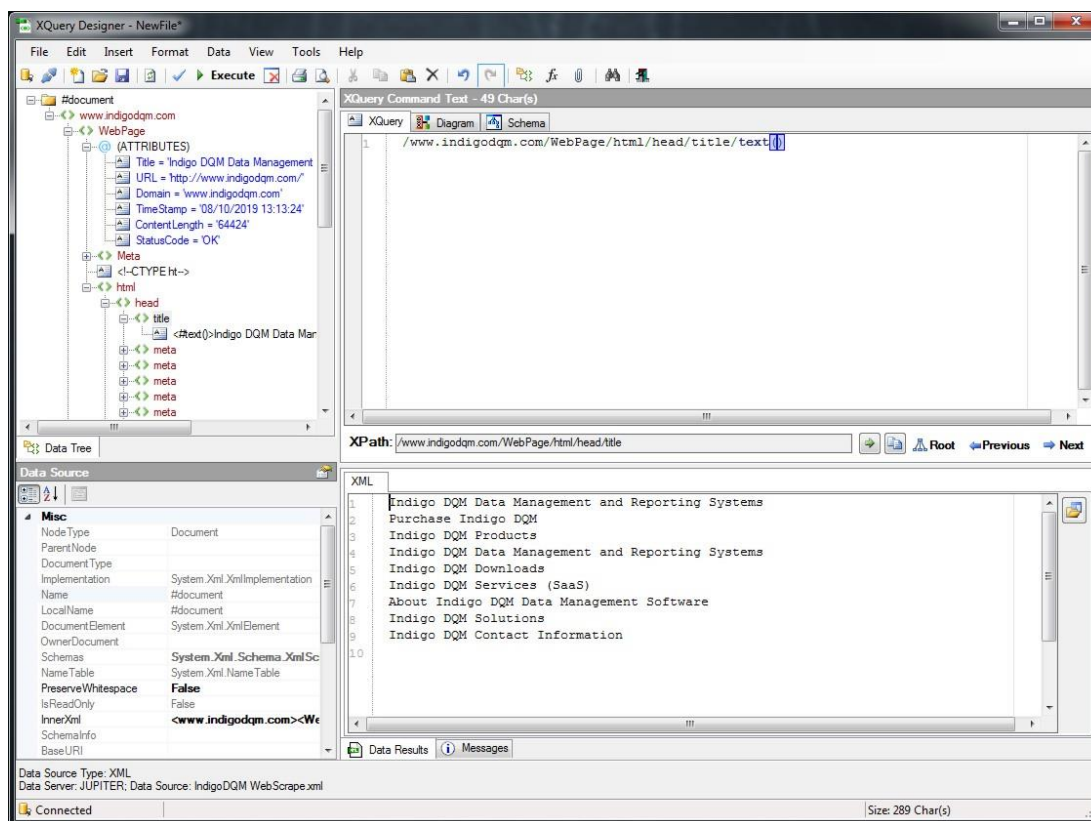
XPath can be used to navigate through elements and attributes in an XML document. XPath is a syntax for defining parts of an XML document and can be inserted by navigating the Data Tree or using the Insert Tool from the menu Insert | XPath.



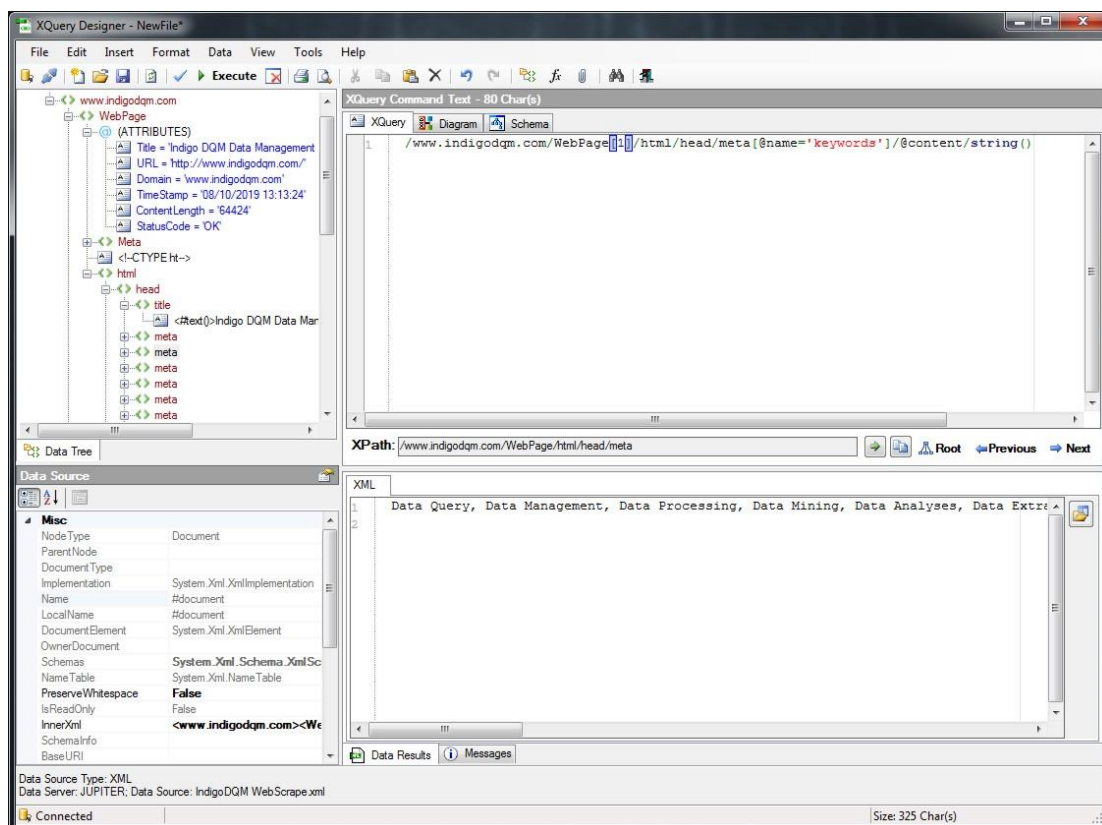
XPath uses path expressions to navigate in XML documents. Click Insert to add the current XPath expression to the XQuery Designer.

Extracting the Web Page Title using XQuery

Executing an XQuery statement for a Web Scrape to extract the Web Page Title.



Executing an XQuery statement for this Web Scrape to extract the Web Page Keywords.



XQuery contains a superset of XPath expression syntax to address specific parts of an XML document.