Table of content

[Introduction 1](#_Toc474917729)

[Template Actions 4](#_Toc474917730)

[TypeCobol Razor Templates 4](#_Toc474917731)

[Code Generator Implementation 8](#_Toc474917732)

[Source Document concept 9](#_Toc474917733)

[Linear Node Source Code Maper. 10](#_Toc474917734)

[Data Structure 12](#_Toc474917735)

[Node Linearization Phase 14](#_Toc474917736)

[Removed Nodes Phase 15](#_Toc474917737)

[Node’s Buffer Source Code association 15](#_Toc474917738)

[Function Declaration lines relocation 15](#_Toc474917739)

[Node without position attachment. 16](#_Toc474917740)

[Linear Code Generation 17](#_Toc474917741)

# Introduction

This document describes the implementation of the TypeCobol Code Generator: Codegen.

TypeCobol

TypeCobol is: Prototype of an incremental Cobol compiler front-end for IBM Enterprise Cobol 5.1 for zOS syntax.This prototype is a work in progress and is written in C#.

The Generation process follows the classical workflow:

TypeCobol Source File

Cobol85 Source file

Figure 1: Main workflow

Thus the Inputs of the Code Generation phase are: The Abstract Syntax tree formed by Nodes and the Input Text Lines of the original source file.

All Inputs can be obtained from an instance of the CompilationUnit class from which Input TokenLines and the RootNode can be retrieved.

The Code Generation phase steps like the following workflow.

1. Code template actions application
2. Node Linearization
3. Code generation in linear time.

Figure : Code generation workflow.

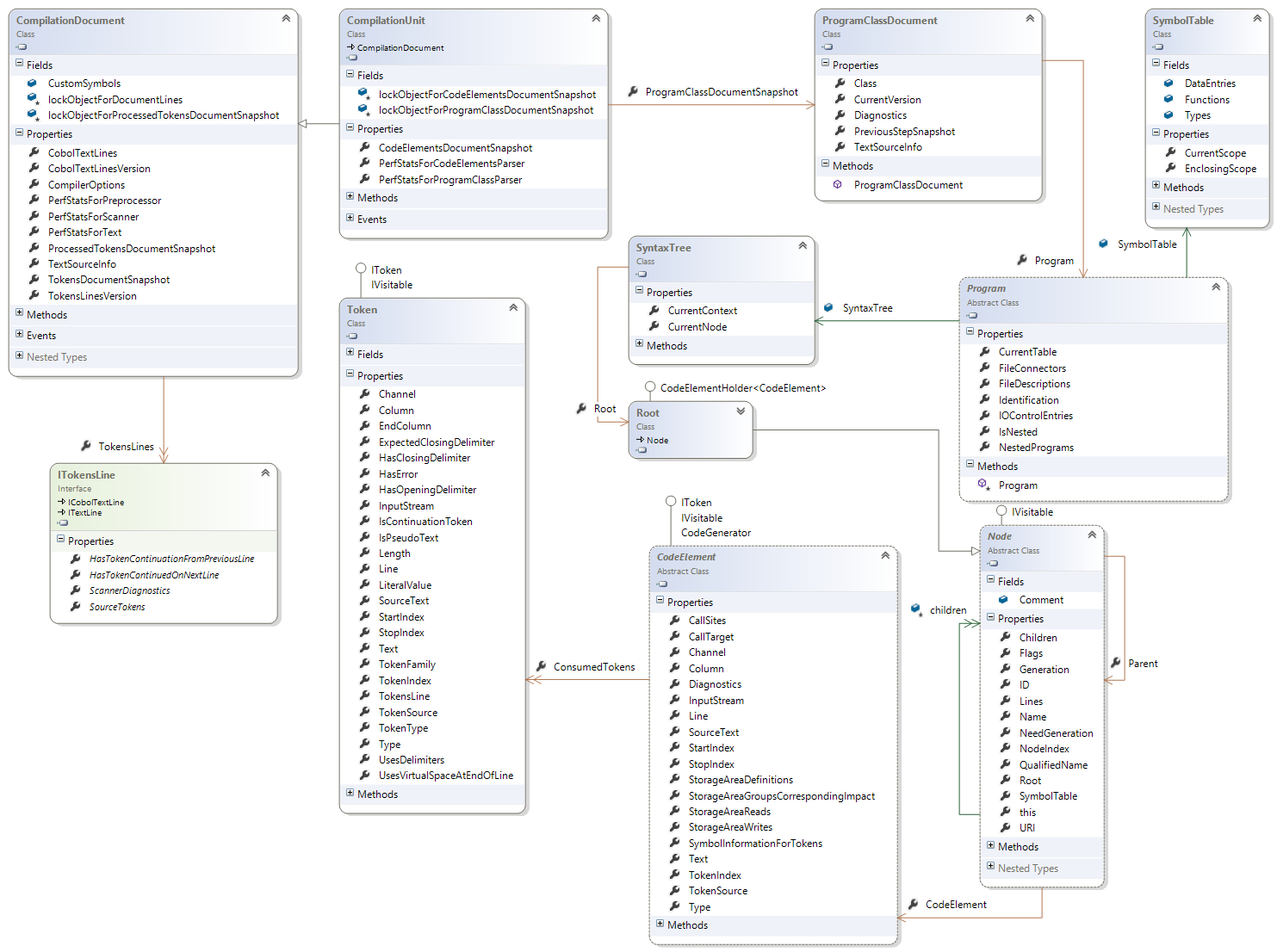


Figure 3: Class diagram: from CompilationUnit to TokenLines and Root Node.

# Template Actions

The Code generator uses code generation based on templates, by using a templating engine built upon Microsoft's Razor parsing technology:

<https://razorengine.codeplex.com/>

<https://github.com/Antaris/RazorEngine>

<https://docs.microsoft.com/en-us/aspnet/core/mvc/views/razor>

The RazorEngine allows you to use Razor syntax to build robust templates. Currently we have integrated the vanilla Html + Code support, but we hope to support other markup languages in future.  
  
Using the templating engine couldn't be easier, using a simple syntax, we can do the following:

string template = "Hello @Model.Name! Welcome to Razor!";

string result = Razor.Parse(template, new { Name = "World" });

The templating engine supports strict and anonymous types, as well as customised base templates, for instance:

Razor.SetTemplateBase(typeof(HtmlTemplateBase<>));

string template =

@"<html>

<head>

<title>Hello @Model.Name</title>

</head>

<body>

Email: @Html.TextBoxFor(m => m.Email)

</body>

</html>";

var model = new PageModel { Name = "World", Email = "someone@somewhere.com" };

string result = Razor.Parse(template, model);

## TypeCobol Razor Templates

Razor Templates used by TypeCobol Generator are in the file:

*TypeCobol\CLI\src\config\skeletons.xml*

This file is an XML file which use to specify template actions that can use Razor template. The schema of the skeletons.xml file is the following and is in the file:

*TypeCobol\Codegen\src\Skeletons\Skeleton.xsd*

|  |  |
| --- | --- |
| **Schema Elements** | **Code / Description** |
| This is the Type that enumerates the available actions that can be applied to a Node.   * Comment * Create * Erase * Expand * replace | <!-- Enumeration Type of the action attribute in a pattern element-->  <xs:simpleType name="ActionEnumType">  <xs:restriction base="xs:string">  <!-- Action to comment the original source code for this node-->  <xs:enumeration value="comment"/>  <!-- Action to create a New Generate Node-->  <xs:enumeration value="create"/>  <!-- Delete words in node input -->  <xs:enumeration value="erase"/>  <!-- The Expand action : Comment the node, remove all children add extra nodes, generate code-->  <xs:enumeration value="expand"/>  <!-- Action to replace a Node by a new Generated one -->  <xs:enumeration value="replace"/>  </xs:restriction>  </xs:simpleType> |
|  | A List of skeletons. A skeleton represents a code generation scheme that can be applied to a node that matches or verifies a set of conditions on whose nodes have attributes values:   * node : the node’s type (à la C# namespace) * name : The node’s name * level : The Cobol Level * type : A Cobol Type * sender: condition-name on SET for condition-names (to true) * receiver: For statements using items for receiving data, this is the name where the value is written. * unsafe: For statements using items for receiving data, are unsafe writes allowed ? * function: relative to function call information ??? (Not Used ???) * definitions: Get all Types and functions definitions of the current node, this definitions will be passed to the **create** and **replace** actions. * variables: not used. * Typecobol: Get all TypeCobol Qualified Symbol Reference * Visibility: Get the visibility of a function declaration if the node is a function declaration. * Copyname: The name of the first copy if one exists "?TCRFUN\_LIBRARY\_COPY?" otherwise.   Pattern is used to describe razor template to be applied to an action.   * Name: is the pattern’s identifier * Group: This * Location: COBOL paragraph where the generated code must be but * Action: the action to be performed. * Var: the variable that are used in the pattern in the format “name=value, …, name=value” * Position: Index position in the target node parent, of insertion. |

The following actions are available:

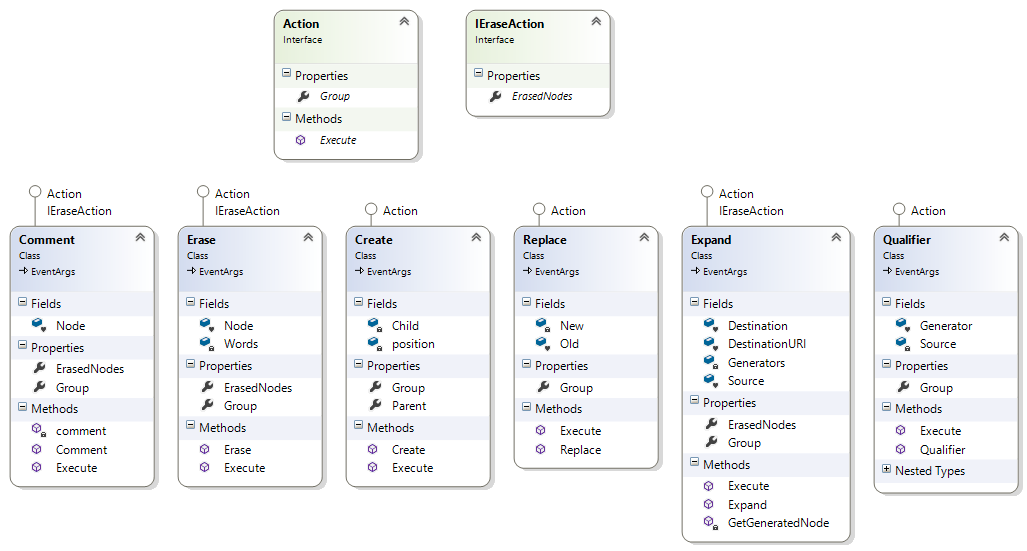
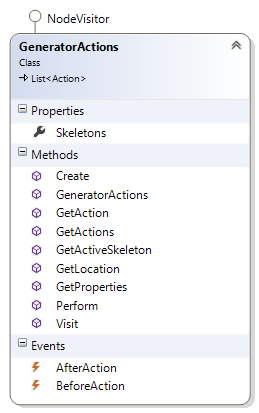


Figure : Class Diagram; Available Actions.

|  |  |
| --- | --- |
| **Action** | **Description.** |
| Comment | Mark a node to be commented with all its children. |
| Erase | Action to remove a Node from the generated code. The Removed node will be commented, all its children will be cleared.  If the erase action contains words to be erased from the input template, a new GeneratedNode is built and its output is the input without the word erased. |
| Create | Action to create a new Generate Node |
| Replace | Action to replace a Node by a New Generated one. The Old one is commented but not its children. The children are copied in the new Generated Node. |
| Expand | Action that created a new node by expanding the current node with new generated node. The source node will be commented in its children cleared.  The expanded node will be added in the Destination parent node as child at the right index. |
| Qualifier | Action used to detect if a node is subject to Type Cobol Qualifier Symbol Reference Style, and to Generate Node to perform the permutation and the replacement of :: by OF. |

All actions are performed by the class *GeneratorActions*.



# Code Generator Implementation

The main process of the Code Generator: is to create a structure that maps Nodes to their corresponding lines in the source code. Each node to be generated is associated a target source code buffer which contents the original source code corresponding to the Node. Generating the code for a Node is then replacing in the associated source code buffer is original code by the generated one, this implies the calculation of the positions of the Node within the associated source code buffer.



Figure 5: Code Generator sequence diagram

The sequence diagram above, illustrates the sequence of actions.

## Source Document concept

The concept of a source document is to be able to store a source text inside a structure that can allows to track positions inside the source code document. So that when portions of text are inserted or removed, these positions are automatically updated (moved) to always reference the tracked position.

The Implementation of such structure implies 6 classes:

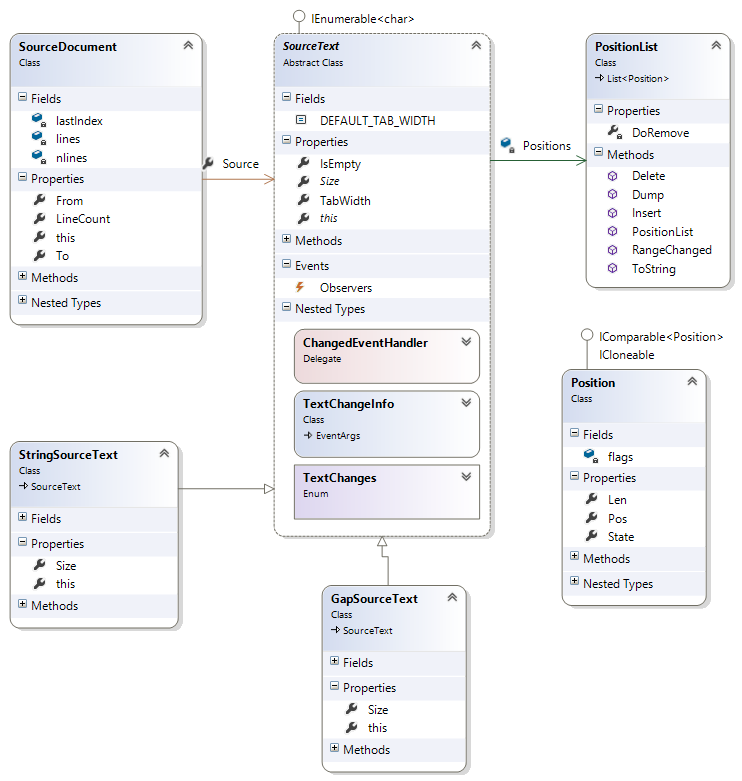


Figure 6: Class Diagram of the Source Document concept

|  |  |
| --- | --- |
| **Class** | **Description** |
| Position | Represents a moveable position inside a source code. |
| PositionList | A List to represent a collection of Positions. On Insertion or deleting the positions inside the list are recalculated. |
| SourceText | An abstract representation of a Source Text. The class Text defines methods to manage a buffer for holding source text. Together with the source text is a list of robust pointers (called Positions) into the source text.  Dependents of the Source Text are noticed by the observer mechanism and receive a change event. |
| StringSourceText | A Source Text which is represented as a String of characters. |
| GapSourceText | A Gap Source Text, with the concept of <https://en.wikipedia.org/wiki/Gap_buffer>: that allows efficient insertion and deletion operations clustered near the same location. |
| SourceDocument | Class that represents a source document, with lines tracking on insertion and deleting. |

## Linear Node Source Code Maper.

The Code generator, generates code by injecting (inserting, replacing, deleting) text in the original source code at specific Node’s position.

**Why injecting code in the original source code**? The answer to this question is that, the code generator must preserve all whitespaces, comments and preprocessor directives that were in the original source at their positions, this because such lexical items are ignored during lexical analysis.

In order to inject text at a right position for a given Node whose code is being generated, it is necessary to calculate the range of positions covered by the node within the original source code.

From the positions of a node, it is then possible to determine which lines in the source code covered the node, and from the target lines we can extract the corresponding source code in a separate buffer.

We can then create the following Mappings.

|  |  |  |
| --- | --- | --- |
| **Domain** |  | **CoDomain** |
| * Node 🡪 Positions 🡪(Line \* Source code) * Node | 🡺 |  |
| * Line | 🡺 | (List<Node> \* ) |

These mappings say that:

* Given a Node we can associate it a Source Code generation buffer in which code injection can be performed based on Node’s positions inside the buffer.
* Given a Line (Line number) we can associate to it which Nodes are generated on the line, and the target Source Code buffer.

Thus it is possible to generate linearly all node, by code injection in target line buffers, injection location are tracked using Node’s positions in the target buffer.

**BUT it still one question: how to determine Node’s positions?**

The answer to this question, is that given a Node it is possible to know its CodeElement and from its CodeElement the lexical tokens consumed by the CodeElement. Each lexical token has: Line number, a StartColum and EndColum values.

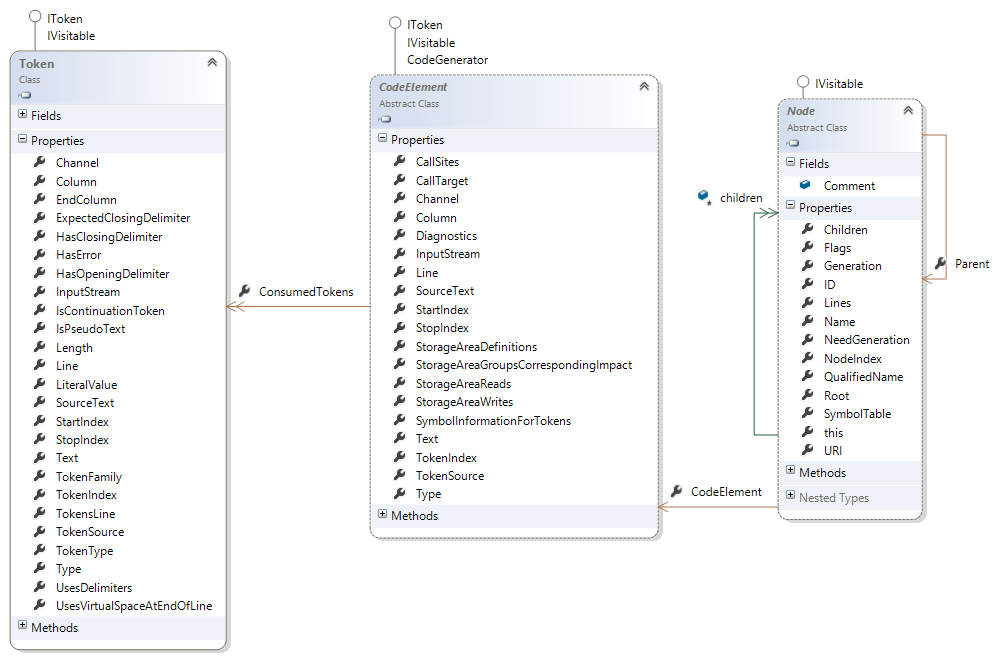


Figure 7: From Node to CodeElement, to Consumed Tokens

The following method:

public Tuple<int, int, int, List<int>, List<int>> FromToPositions(Node node)

In the class ***TypeCobol.Codegen.Generator*** calculates all position information for a given Node. Calculated positions are related to the Original Source Document. The method return a tuple whose items are:

|  |  |
| --- | --- |
| **Items** | **Description** |
| Item1 | The ***From*** column which is the column of the first valid Token |
| Item2 | The ***To*** column which is the ending column in the last line covered by Node’s tokens. |
| Item3 | The **Span** which is the indentation on the last line covered by the Node’s tokens up to the ending token’s column. |
| Item4 | The list of Line Numbers covered by the Node’s tokens. |
| Item5 | The list of offsets of each line from the first line inside the target source code buffer. |

Given these items, it is possible by translating these positions, to the portion of text covered by the Node’s lines, to calculate the (**From**, **To**) positions of the Node within the target source code buffer.

To Construct the Linear Node Source Code Mapper we can distinguish the following Steps:

1. Node Linearization Phase
2. Removed Nodes Phase
3. Node’s Buffer Source Code association
4. Function Declaration lines relocation
5. Node without position attachment.

### Data Structure

The Linear Node Source Code Mapper uses the following class as Data Structure.

|  |  |
| --- | --- |
| Classes | Description. |
|  | The class LinearNodeSourceCodeMapper is the main class whose members are described below. This class maintains the array of Line information *LineData* and the indexed list of nodes *Nodes*.     * The Inner class LineInfo is used to store data related to a line:  (*Buffer* the target source code buffer, *FunctionBodyBuffer* if the line is related to a function body buffer, *LineNodes* the list of Node index associated to the line).      * The Inner class NodeData is used to store information related to a Node.      * The Inner class NodeFunctiondata is a specialization of the class NodeData used to store specific information related to a Function Declaration Node.      * The Inner class *LinearGeneratedNode* is used to represent additional Nodes created to the Linearization, these Nodes are nodes that represents the code source of text lines that are relocated when the function declaration is relocated at the tail of the generated source document.      * The Inner Enumeration *Phase* enumerates ALL phases performed during the linearization process. |

### Node Linearization Phase

The Linearization is the process to associate to each line in the original source code its nodes, and to associate to each node its target Source Code Buffer. So that we can establish the following relations.

* Node -> SourceText
* Line -> (List<Node> \* SourceText)

The first relation says that given a Node we can associate its corresponding Source Code Buffer, and the second relation says that given a Line (Line Number) we can associate its list of nodes and its Source Code Buffer.

**Lines that have no associated Nodes are lines that must be generated without any changes from the source code.**

**More often these lines are comments or preprocessor directive or untranslated source code.**

The Linearization Phase starts from the RootNode and traverse all its children recursively.

The following method of the class TypeCobol.Codegen.Generators.LinearNodeSourceCodeMapper is executed:

|  |
| --- |
| /// <summary>  /// The Linearization Phase.  /// <param name="node">The node to linearize</param>  /// <param name="functionBody">true if the node belongs to a function body, false otherwise</param>  /// <returns>True if children of the given node must be visited, false otherwise</returns>  /// </summary>  private bool ProcessLinearization(Node node, bool functionBody = false); |

Please read the source code of this method it is documented.

#### FunctionDeclaration Phase

The *FunctionDeclaration* phase is a sub phase of the linearization phase, it is uses to associate to a Function Declaration Node all its function body nodes. It is necessary to do that because the code generated for a function declaration must be generated in a local buffer, so that the generated code can be relocated at the end of the final generated code.

This phase also determines the first line and the last line in the source code of the function declaration body, this necessary for collecting all lines that are note associated to a node inside the function body, the lines must also be relocated along the function body relocation.

The following method of the class TypeCobol.Codegen.Generators.LinearNodeSourceCodeMapper is executed:

|  |
| --- |
| /// <summary>  /// The Function Declaration Processing Phase. The Purpose of this phase is to associate  /// to a FunctionDeclaration Node all its nodes that belongs to its body.  /// It also determines the starting line and the ending line in the source code of the  /// function body.  /// <param name="node">The node that belongs to a Function Body</param>  /// <returns>True if children of the given node must be visited, false otherwise</returns>  /// </summary>  private bool ProcessFunctionDeclaration(Node node) |

### Removed Nodes Phase

The Removed Nodes Phase is the phase that marks all node that must be removed from the generation process. Their field *Removed* in the NodeData instance is set to *true*.

### Node’s Buffer Source Code association

The method is used to create the content of each source text buffer associated to a line.

From the original source code, text line are extracted and inserted in line source text buffers.

And for each Node Tracked positions (from, to) are created into the target source code buffer. The Generated code for the node will replace the portion of text between the (from, to) positions.

The following method of the class TypeCobol.Codegen.Generators.LinearNodeSourceCodeMapper is executed:

|  |
| --- |
| /// <summary>  /// Create All SourceTextBuffer Content associated to Nodes and Create  /// Node's positions inside the associated buffer.  ///  /// </summary>  private void CreateNodeSourceTextBufferContents() |

### Function Declaration lines relocation

This step creates for each line not associated to a Node inside a function body declaration, a specific node of the class *LinearGeneratedNode* inserted at the right index in the Function Declaration associated node list.

It also insert at the right position Nodes without positions inside a function declaration body.

The following methods of the class TypeCobol.Codegen.Generators.LinearNodeSourceCodeMapper are executed:

|  |
| --- |
| /// <summary>  /// For all Function declarations, this method relocates all lines within the function declaration body  /// that have not been relocated when the function declaration body has been relocated.  /// This by Dealing with all lines that are not attached to a node.  /// And relocates nodes without positions  ///  /// </summary>  private void CompleteFunctionDeclarationLinesRelocation() |
| /// <summary>  /// Collect all lines that have not been associated to a Node during Function Declaration  /// processing phase. The lines are then associated to Dummy nodes, that have a buffer containing  /// the source code of the line.  /// </summary>  /// <param name="funData">The Function Declaration Data</param>  private void CollectFunctionBodyUnNodedLines(NodeFunctionData funData) |
| /// <summary>  /// Relocate all Function declaration nodes that don't have positions.  /// Nodes without positions are relocated in the line having the last valid buffer.  /// </summary>  /// <param name="funData">The Function Declaration Data</param>  private void RelocateFunctionBodyNoPositionNodes(NodeFunctionData funData) |

### Node without position attachment.

This step deals with Nodes that have been added by the Skeleton action *Create*, during the processing of the skeletons file on the nodes.

These Nodes are created by a Factory and thus don’t have positions, they cannot be associated to a line in the source code.

The Strategy is to add these nodes in the list of associated nodes of the last line of first parent node having positions.

The following methods of the class TypeCobol.Codegen.Generators.LinearNodeSourceCodeMapper are executed:

|  |
| --- |
| /// <summary>  /// This Phase deals with node created by the Generator Factory thus nodes  /// that cannot be associated to a Position. The strategy is to associate these nodes  /// to the last line of the first parent node which has positions.  /// </summary>  /// <param name="node">The node</param>  /// <returns>true</returns>  private bool ProcessFactoryGeneratedNodeAttachment(Node node) |
| /// <summary>  /// Given a node this methods gives its first parent which has a position.  /// </summary>  /// <param name="node">The Node to get the first parent which has a position</param>  /// <returns>The first parent with a position if any, null otherwise</returns>  private Node GetFirstParentWithPosition(Node node) |
| /// <summary>  /// Get the last line of a node.  /// BECAREFUL this method must be called after the linearization phase,  /// beacuse it uses positions calculated during the linearization phase.  /// </summary>  /// <param name="node">The node to get the last line</param>  /// <param name="lastLine">output le last line number</param>  /// <param name="lastNode">The last node of the last line number</param>  /// <returns></returns>  private void GetAfterLinearizationLastLine(Node node, ref int lastLine, ref Node lastNode) |

## Linear Code Generation

From the LinearNodeSourceCodeMapper structure the code generation is run over the lines and over the node list of each line. In the worst case the generation is performed with a complexity in time of O(m\*n) where m is the number of line and n is the number of nodes. In the best case if we consider that there is one node by line the complexity in time is O(m).

The class TypeCobol.Codegen.Generators.DefaultGenerator implements the Default Code Generator

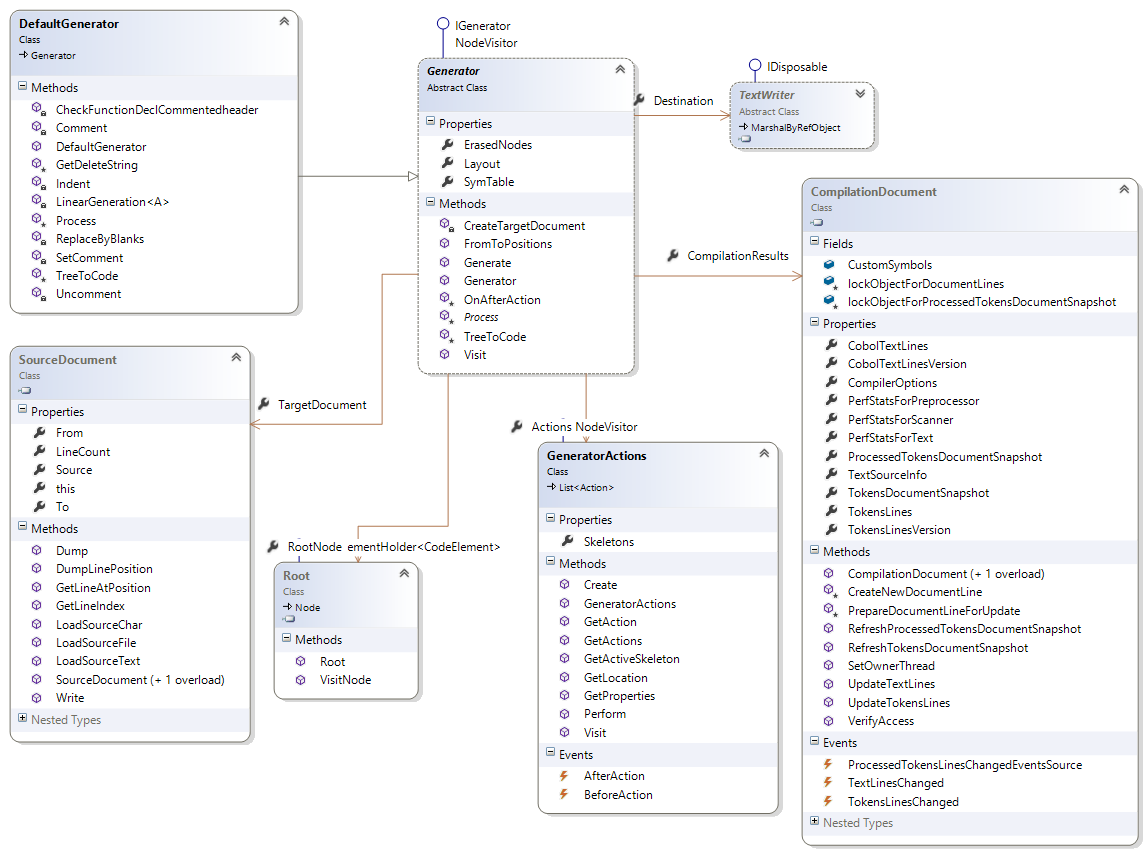


Figure 8:Global Class diagram for the class DefaultGenerator

The following method of the class TypeCobol.Codegen.Generators.DefaultGenerator is executed:

|  |
| --- |
| /// <summary>  /// Perform a linear Generation  /// //1) A Non commented line with no Associated nodes is generated without any change.  /// //2) If the line is commented then first comment all following lines that have the same intersection with the corresponding target Nodes.  /// //3) For each node related to a line, and not already generated the corresponding code.  /// //4) Flush of Function declations.  /// <param name="mapper">The linearization representation</param>  /// <param name="Input">Input source lines</param>  /// <returns>The Generated Source Text</returns>  /// </summary>  private SourceText LinearGeneration<A>(LinearNodeSourceCodeMapper mapper, IReadOnlyList<A> Input) where A : ITextLine |

The global steps of the generation process performed by the method are the following:

|  |
| --- |
| * For each line in the source document   + A non-commented line with no associated nodes is generated as it is in the source code.   + If the line is commented then the line and its following lines having nodes intersections with it are generated commented.   + For each node not already generated related to the line     - If the node is not a generated from a factory Action then       * If the node is removed then replace by blank the portion of text covered by its positions, in its target source code buffer.     - If the node is not generated by a factory Action then       * If the node is not a Function Declaration node.         + Replace in the node’s source code buffer, the portion of the original source text, by the code generated for the Node. (Code Injection).       * If the node is a Function Declaration Node         + Erase in the node’s source text buffer the header of the function declaration.         + Insert in the target Function Declaration Source Code Buffer the Commented Function header source code.         + Insert in the target Function Declaration Source Code Buffer the epilog code of the function body.         + Push the current generation context on a Stack         + Switch to the Function Generation Context (The target generation buffer, becomes the Function Declaration node source buffer).       * If the Node is an End Function Declaration Node then         + Pop the Context Stack 🡪 Leaving the Function declaration generation context. * Flush in the target source code buffer all Functions Declaration Source Code Buffers. |