

# PEPSI-LiTE Curation Tool

## User Guide

PEPSI-LiTE curation tool is a tool with the following features:

- generating chemical structure from a given compound name
- generating IUPAC Name from a given structure
- converting image to a structure
- converting text from text image in pdf document
- review/update structure and structure properties
- generating xml file from the .sdf file
- validate xml file against W3C rules of wellformedness
- validate xml file against a given .xsd schema
- review/update xml file.

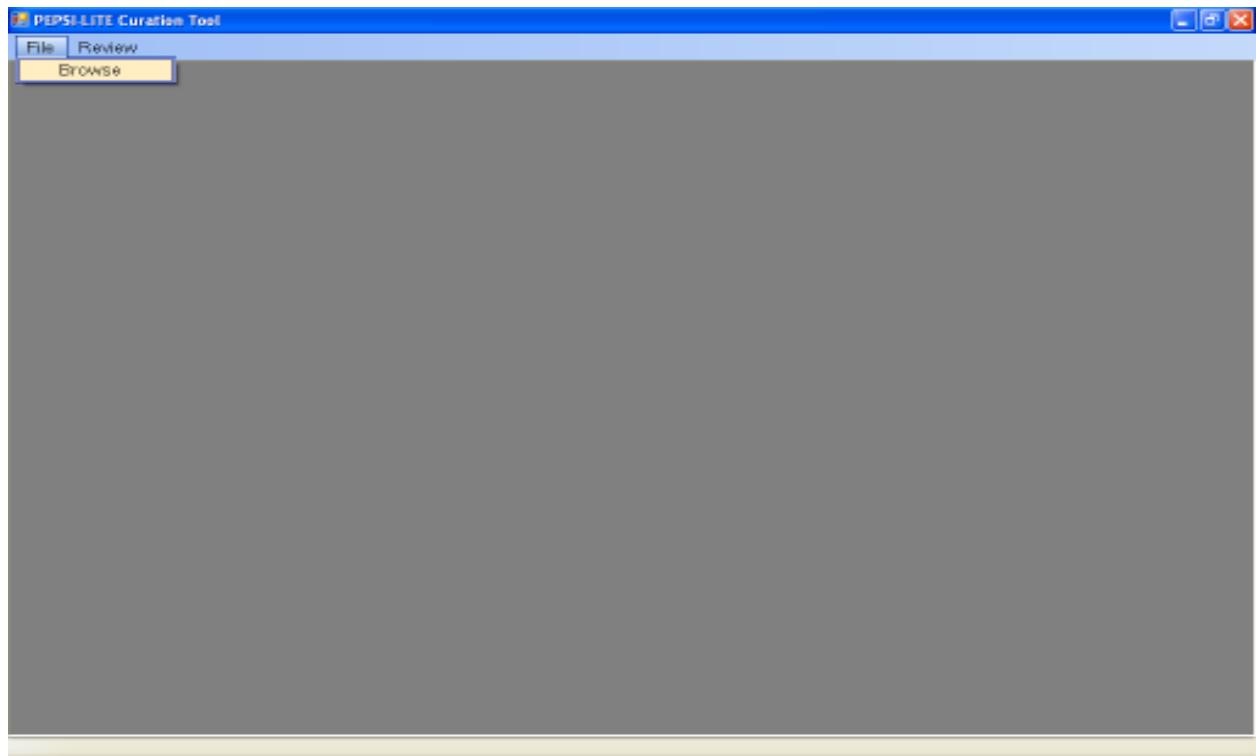
**PEPSI-LiTE curation tool is developed with an objective of minimizing man hours in the curation process of the PEPSI-Lite curation project.**

With this tool

- We can minimize the time to convert chemical compound names to structures without manually drawing structures.
- We can generate IUPAC Names automatically.
- We can convert text image to text in the process of generating structure from a compound name.
- We can review and update .sdf file very efficiently.
- We can generate xml file from the .sdf file.
- We can validate xml file against the given .xsd schema by the client without using any third party software.

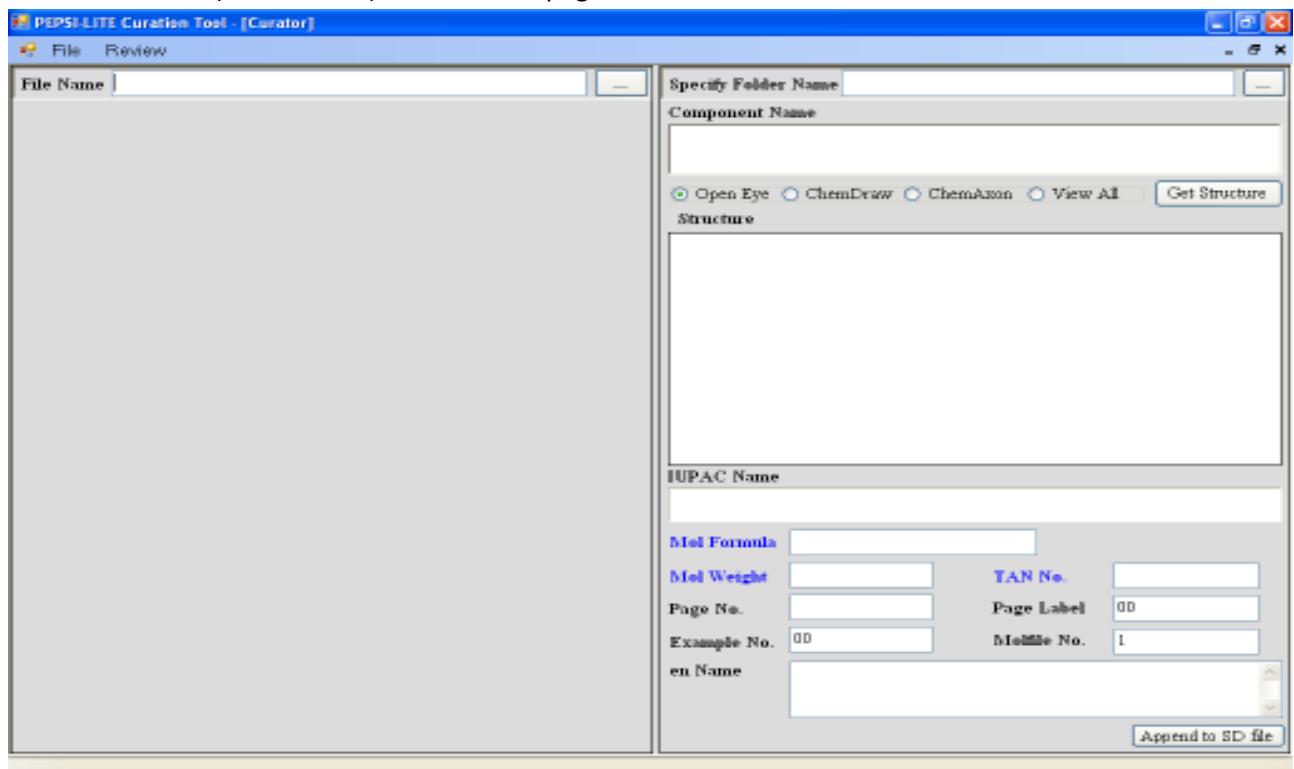
**Note:** Before using the software, please install the list of softwares mentioned in the PEPSI-LiTE Curation Prerequisites.doc file.

-for any queries/issues while installing/using the application, please feel free to contact  
[sairam.punyamantula@gvkbio.com](mailto:sairam.punyamantula@gvkbio.com)



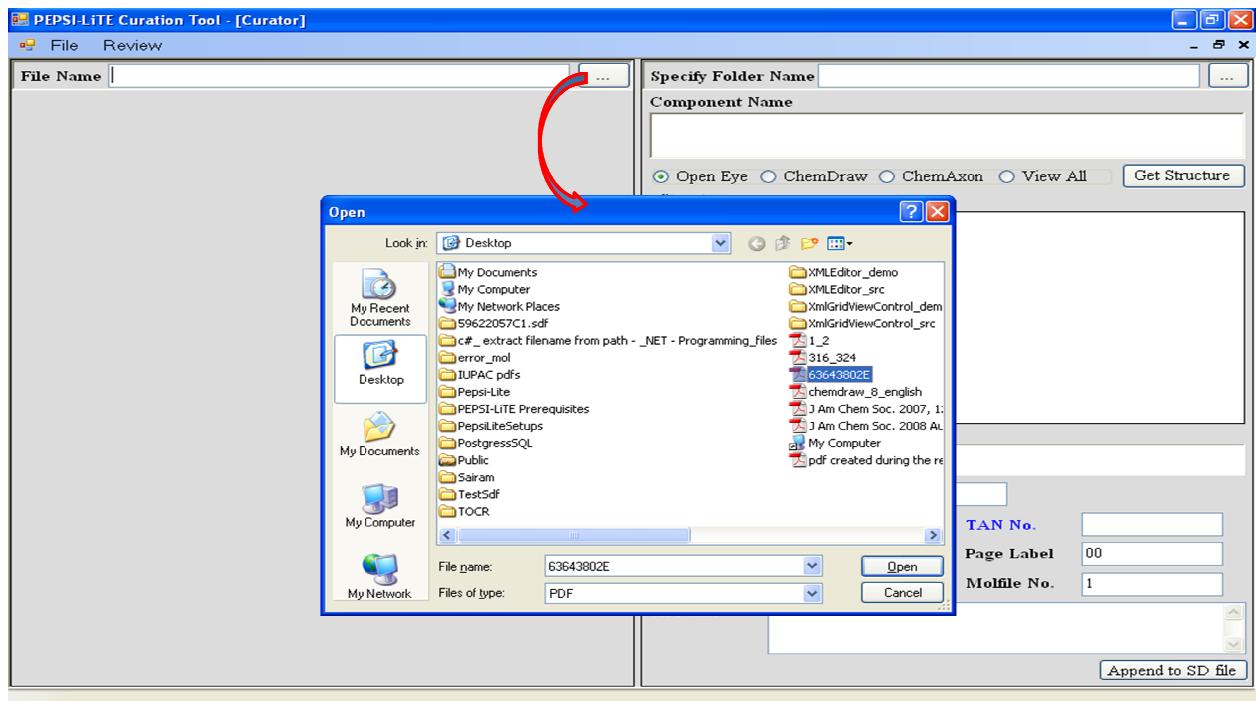
### Screen No: 1

The above screen(**Screen No:1**) is the master page of the PEPSI-LiTE Curation tool.



### Screen No: 2

Once you click on Browse button, Curator page will be displayed, as shown in **Screen No:2**.



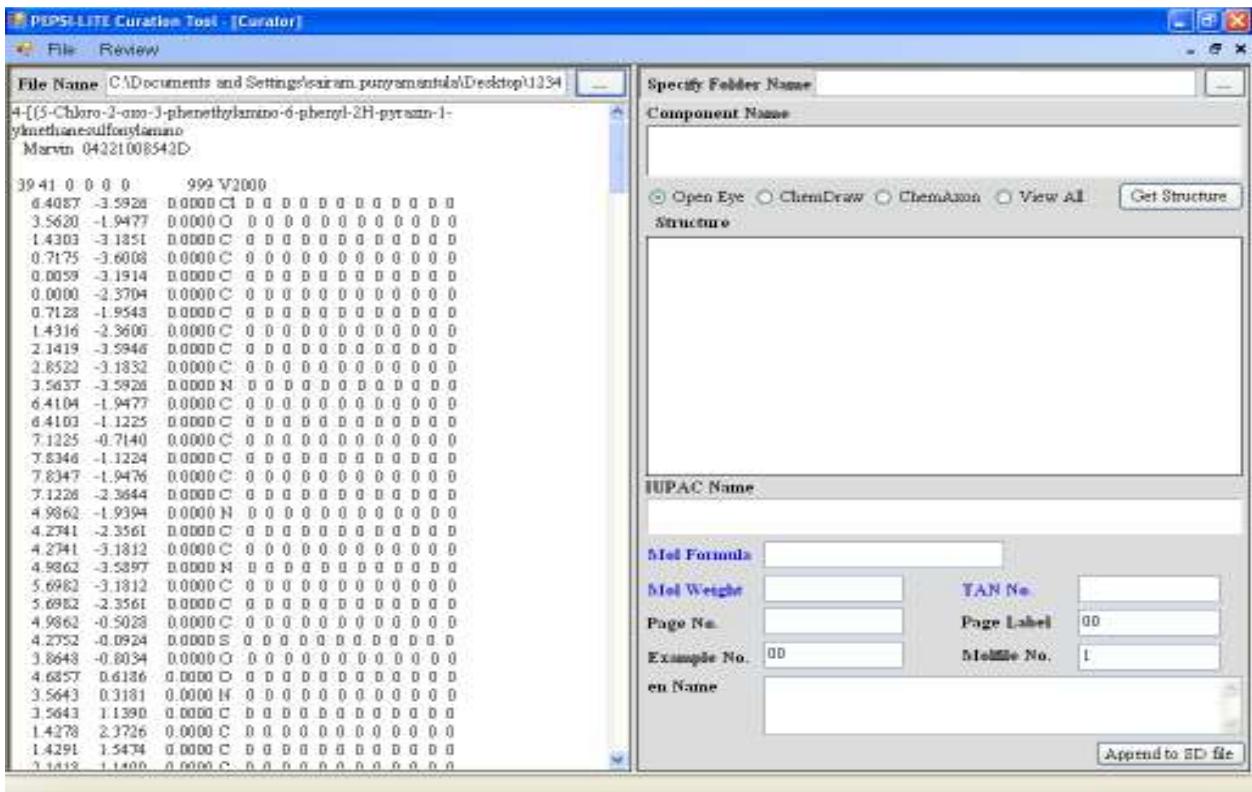
### Screen No: 3

When you click on browse file button open file dialog box will be shown. In this dialog box, user have the option to select either .txt / .pdf file. Once you select a file the corresponding file will be opened in the left pane, as shown in **Screen No: 4**(if pdf file) and **Screen No: 5**(if txt file).

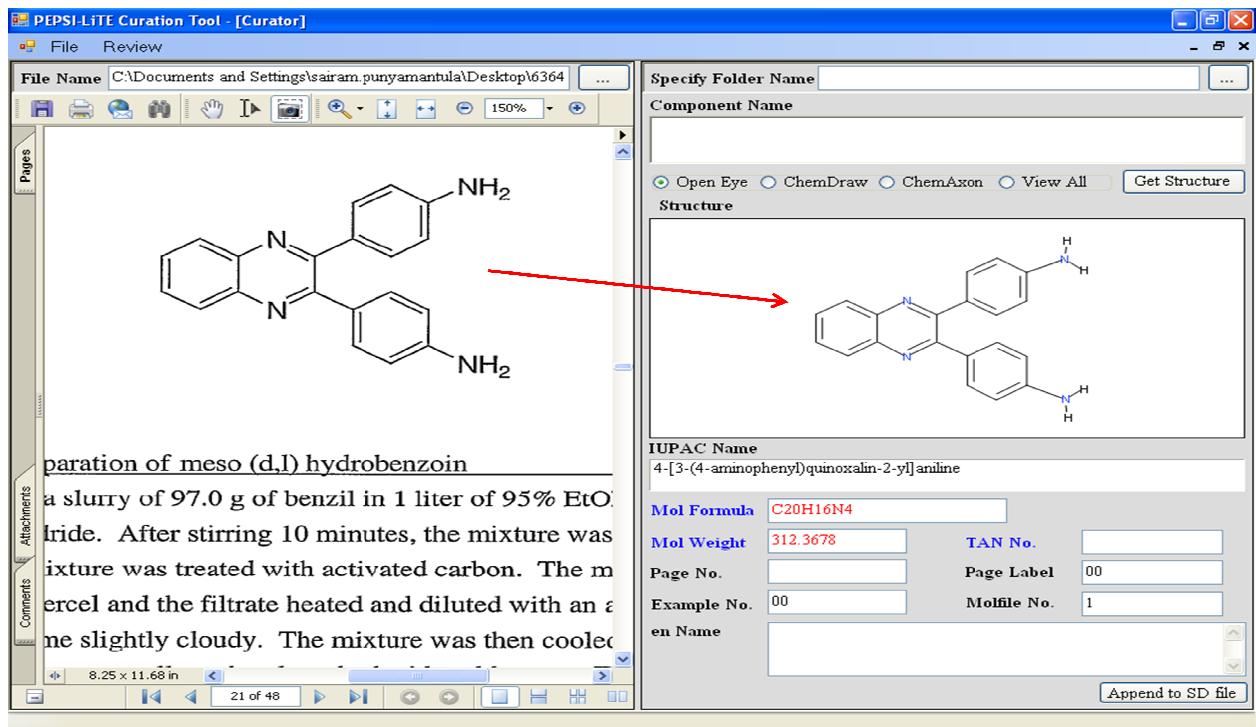
paration of meso (d,l) hydrobenzoin

a slurry of 97.0 g of benzil in 1 liter of 95% EtO. Iride. After stirring 10 minutes, the mixture was mixture was treated with activated carbon. The m ercel and the filtrate heated and diluted with an en slightly cloudy. The mixture was then cooled s were collected and washed with cold water. T

### Screen No: 4



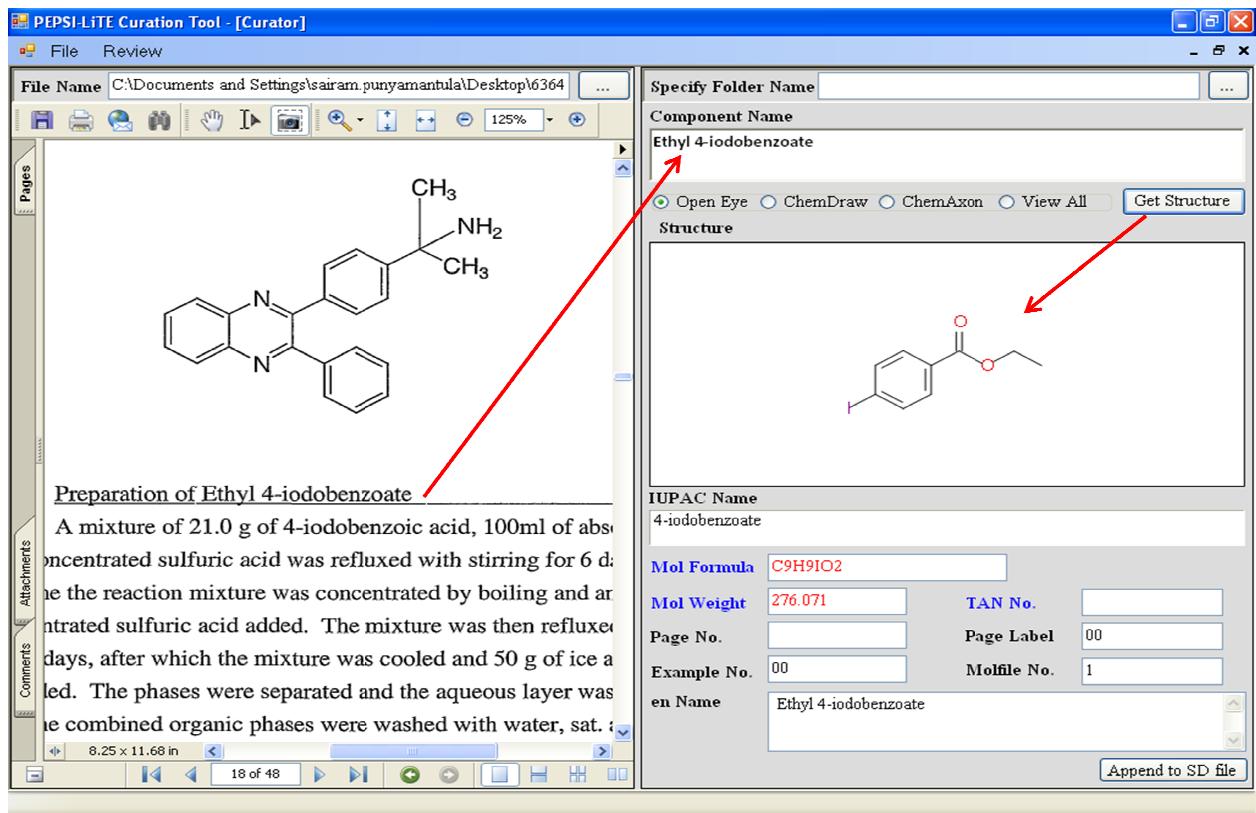
Screen No: 5



### Screen No:6

When user copy any image structure in a pdf using pdf image tool and click in structure pane, PEPSI-LITE Curation tool will automatically converts image to structure.

**Note: Zoom the pdf document to 150% to get correct structure.**

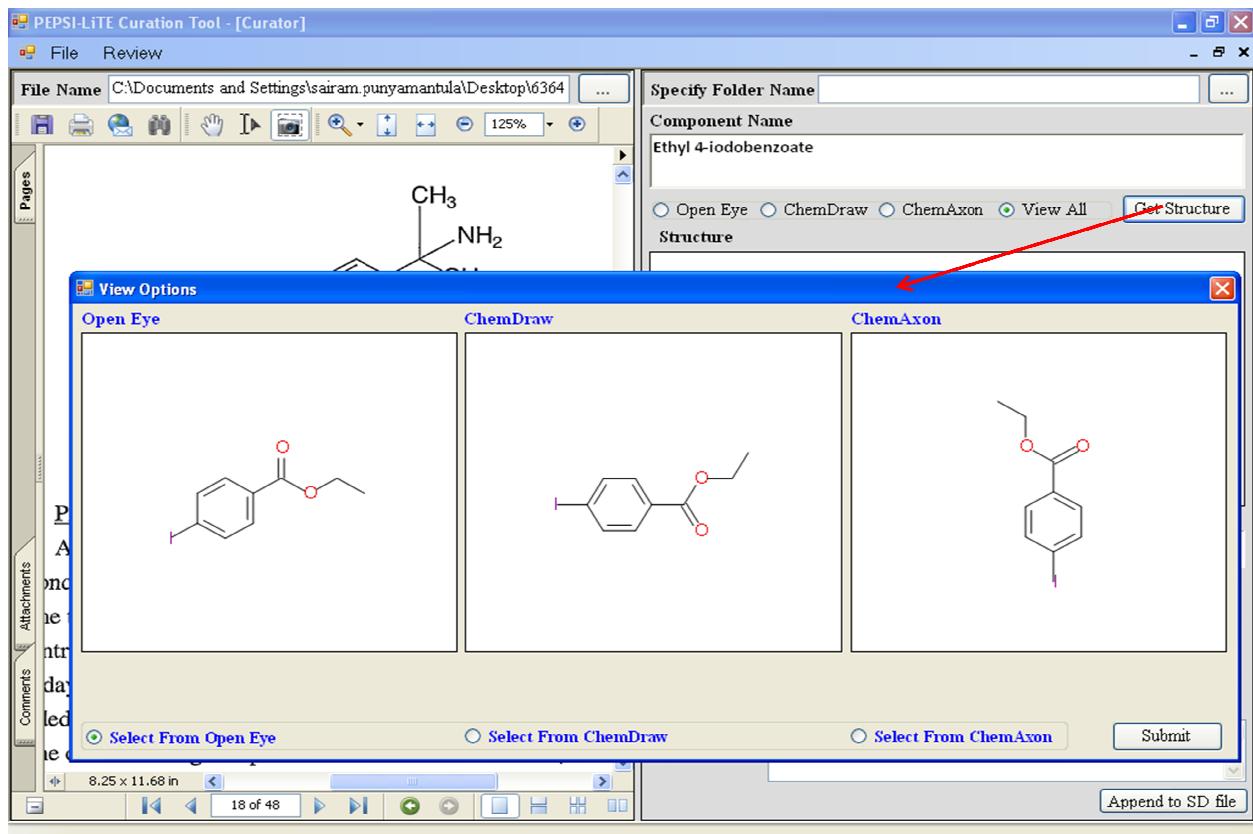


### Screen No: 6

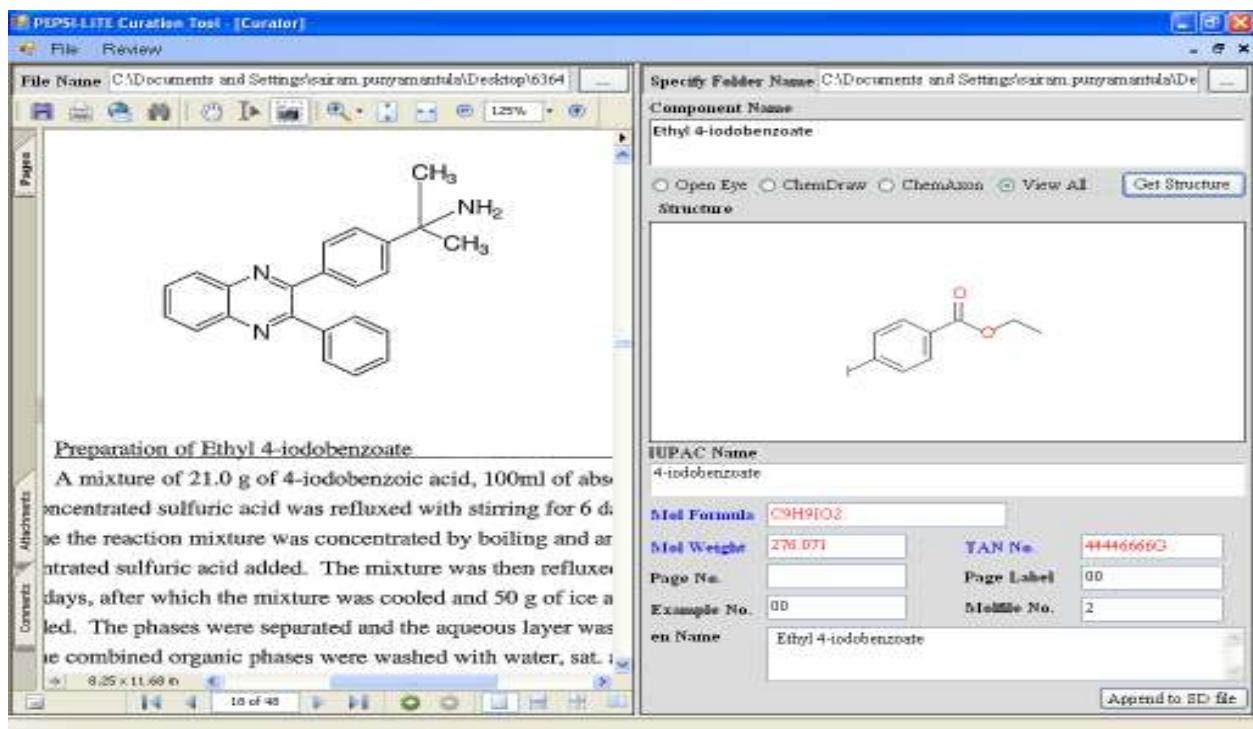
When you copy a text image from the pdf browser and when you click on compound name text box, the copied text image will be converted into text automatically. If the converted text is correct then click on **Get Structure** button or do the required changes, select **Open Eye/ChemDraw/ChemAxon** view option and click on **Get Structure** button.

When you click on Get Structure button the compound name will be automatically converted into a chemical structure and **IUPAC Name** for the structure will be displayed in the **IUPAC Name label**, and **Mol Formula** and **Mol Weight** for the structure will be calculated automatically, as shown in **Screen No: 6**

The user have the option to view all the view options by selecting **View All** option and click on **Get Structure** button. The user will be redirected to **View Options** form. In **View Options** form user can see the converted structure in **Open Eye, ChemDraw and ChemAxon**. Once user selects any option and click on **Submit** button, that selected structure will be shown in Structure pane, as shown in **Screen No: 6A and Screen No: 6B**

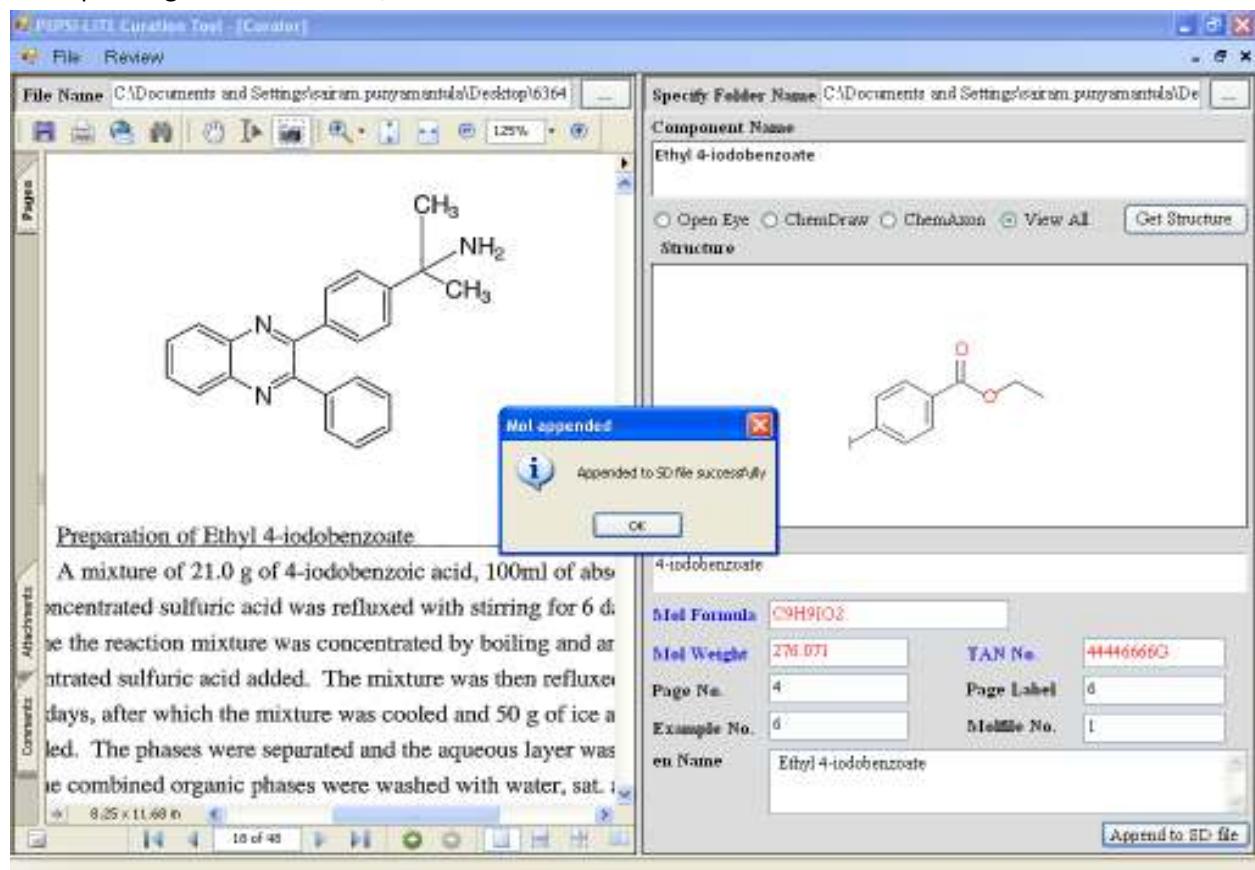


Screen No: 6A



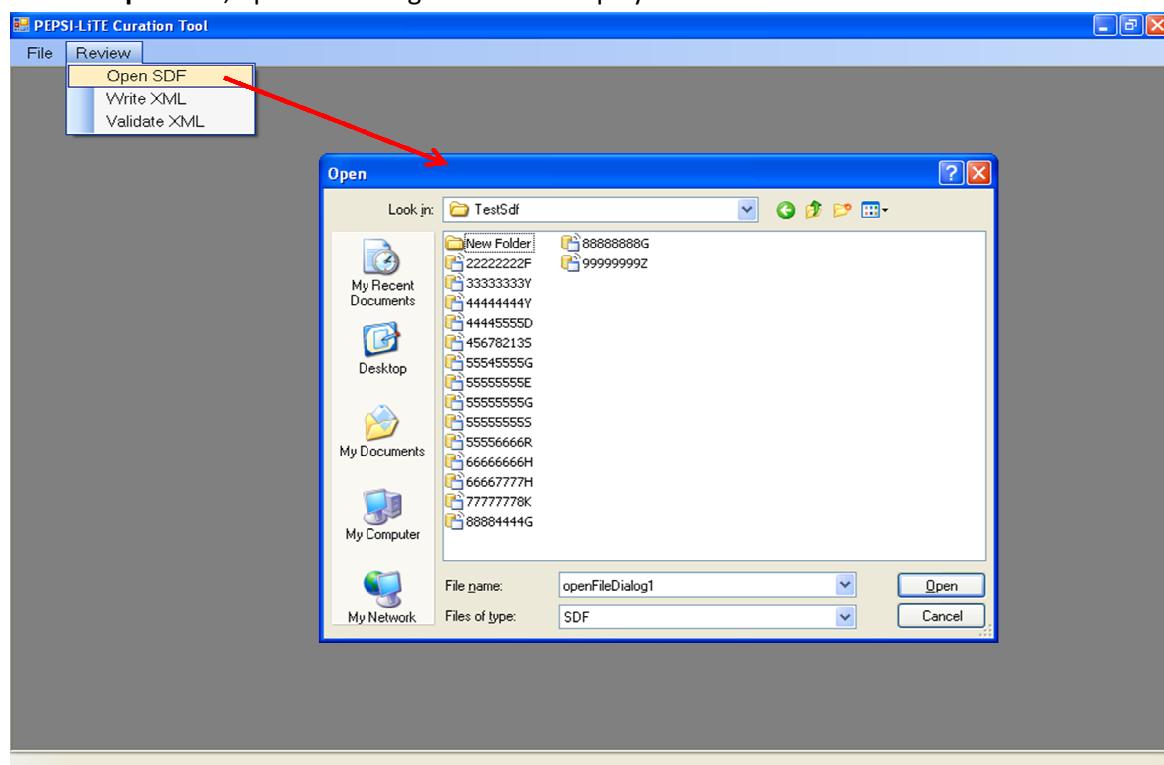
Screen No: 6B

The end user have to specify a folder/directory name in the Specify Folder Name textbox, to save the generated \*.sdf file in the corresponding folder. Once the user enter all the required fileds like TAN NO, Page No, etc.., the structure with the properties will be saved in the corresponding folder with the name **TANNumber.sdf**. All the other generated records will be appended to the corresponding **TANNumber.sdf**, as shown in **Screen No: 7**.



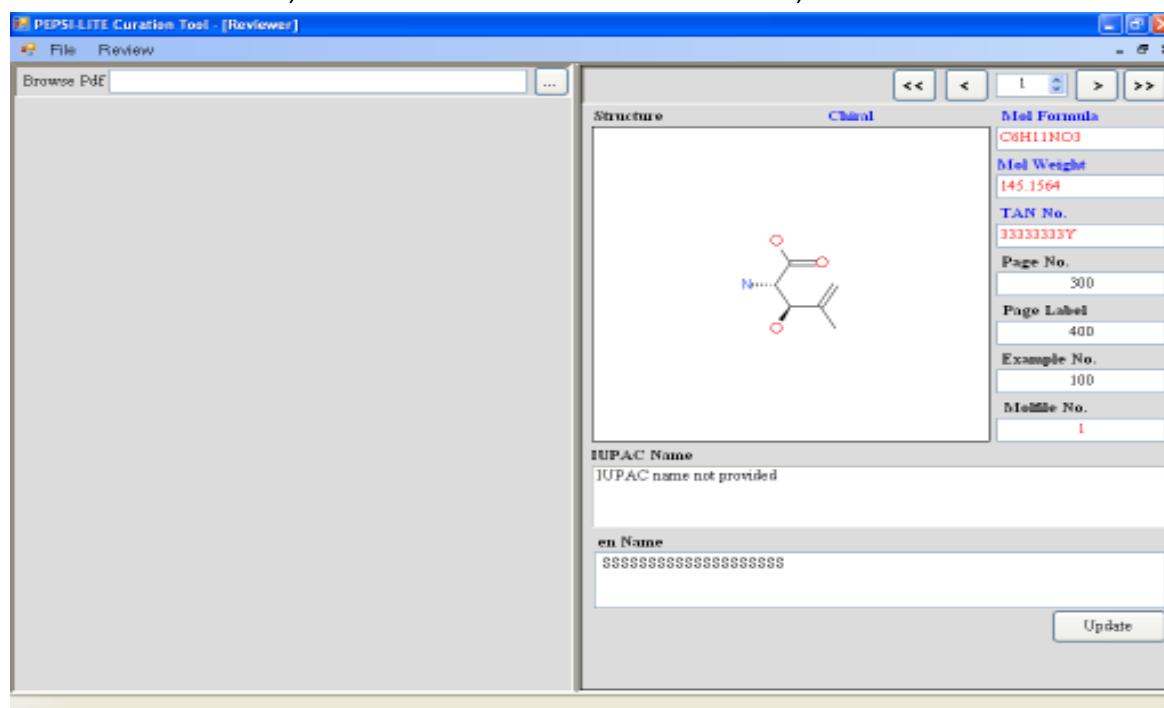
Screen No: 7

Once all the records are curated, user can review the .sdf file in the Review menu.  
Click on **Open SDF**, open file dialog box will be displayed as shown in **Screen No: 8**.



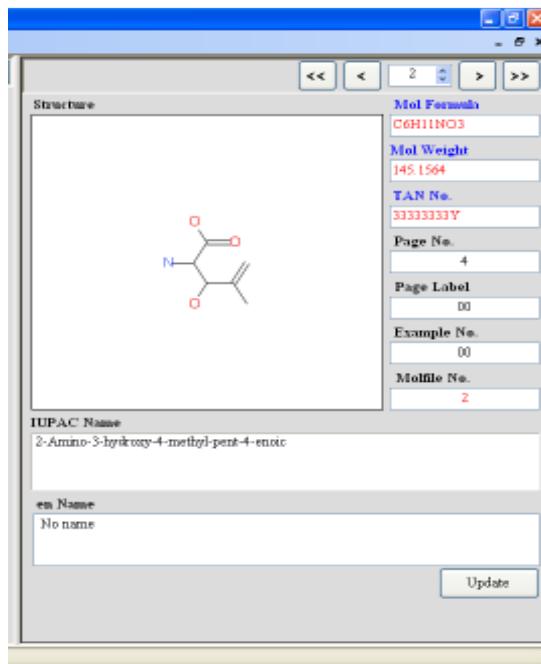
**Screen No: 8**

Once user select sdf file, the sdf file is shown in the Reviewer form, as shown in **Screen No:9**.

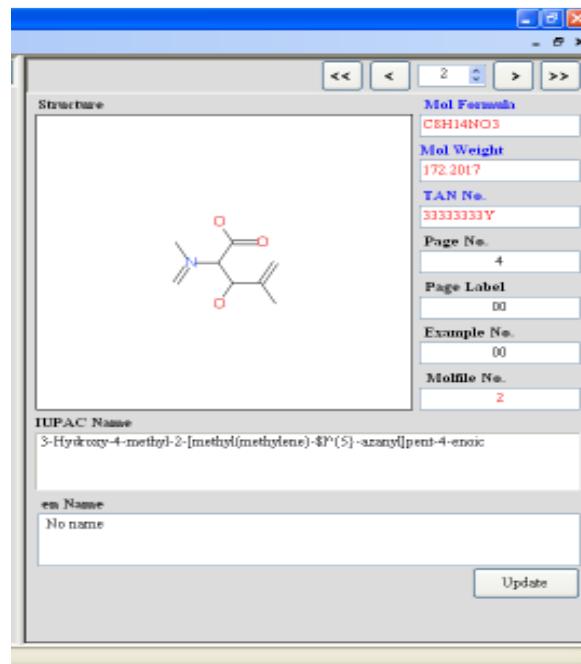


**Screen No: 9**

With the navigation buttons <<, <, >, >> user can navigate through the records in the sdf file. If you want to update the structure, just double click on the structure, the application will be navigated to the **Symyx draw control**. Once you change the structure and close the Symyx draw control , the updated structure will be displayed in the Reviewer page and the Mol Formula, Mol Weight and IUPAC Name will be calculated automatically, as shown in **Screen No: 10A & 10B**.

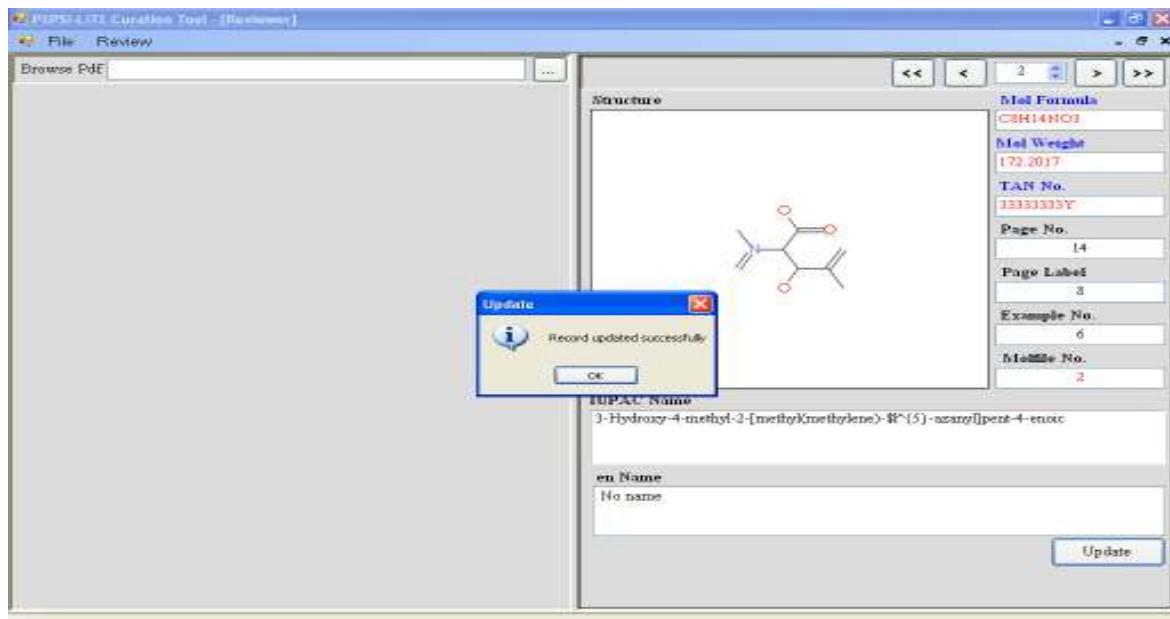


**Screen No: 10 A**



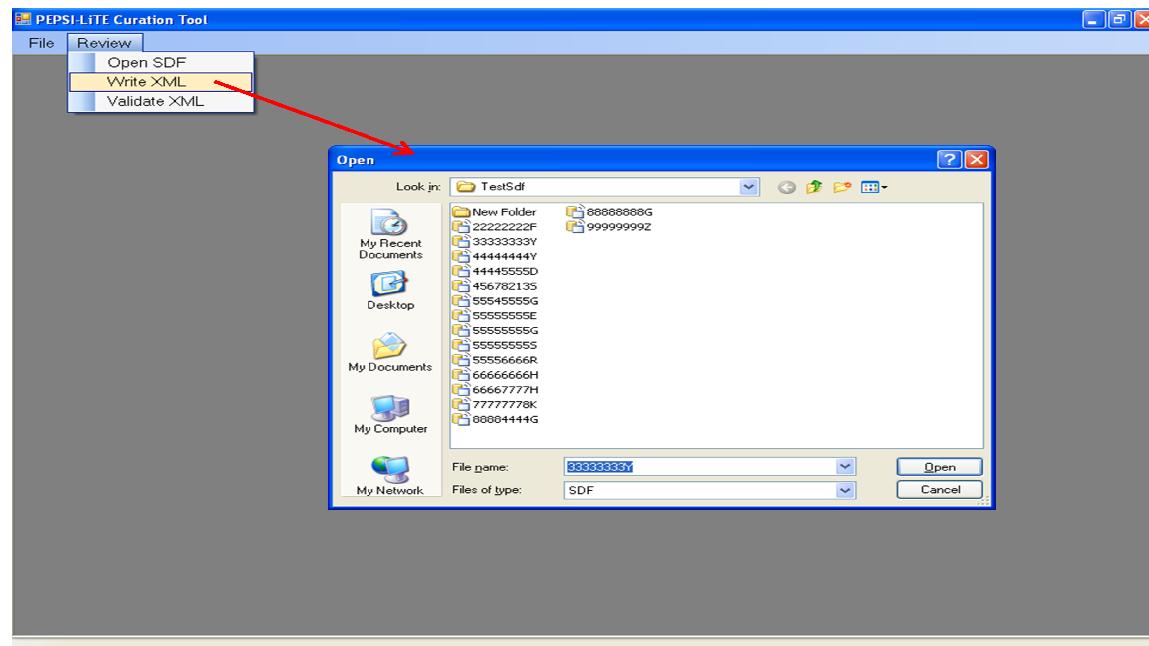
**Screen No: 10 B**

Once the required changes are done click on Update button, the record will be updated in the corresponding sdf file and status will be displayed as an message to the user, as shown in **Screen No:11**.



**Screen No:11**

Once the review is done, Click on Write XML submenu under Review menu, as shown in **Screen No:12.**



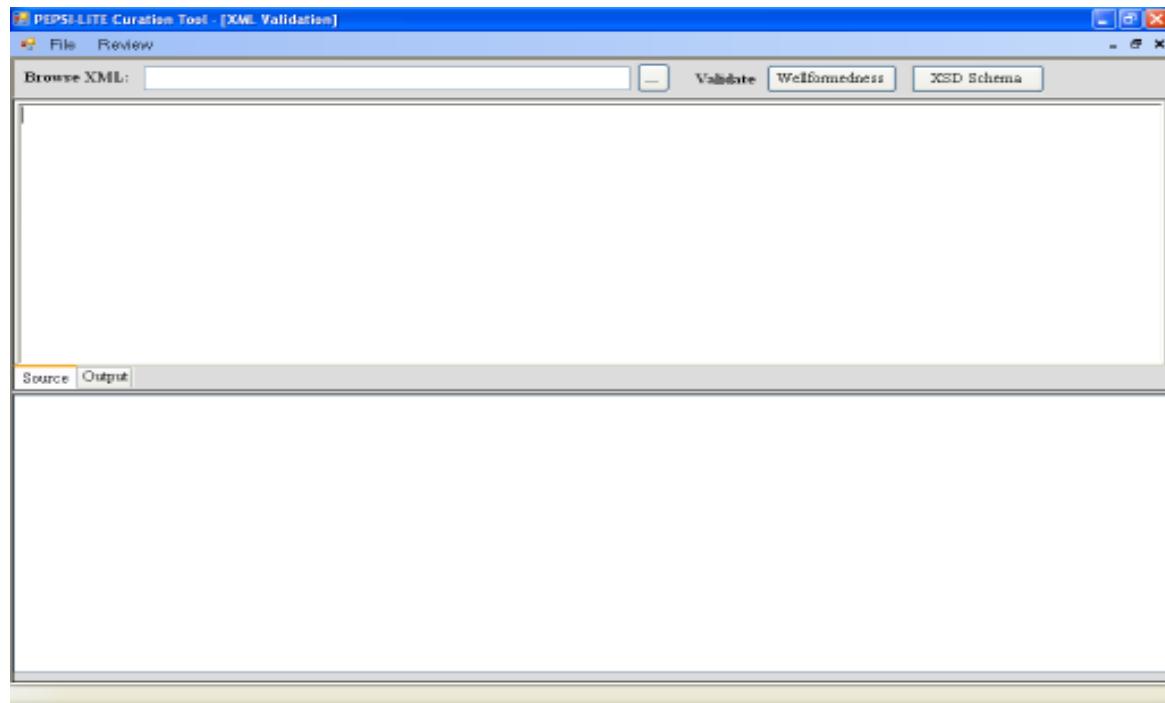
**Screen No:12**

Once user select a particular .sdf file, an xml file will generated in the same folder with the same name as sdf file after the validation against the given .XSD schema and the generated .xml file is shown to the user, as shown in **Screen No:13.**

```
<?xml version="1.0" encoding="utf-8" ?>
- <patent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xsi:noNamespaceSchemaLocation="PatentEnhancedPrioritySubstanceIndexing-2.9.xsd">
  - <patientInfo>
    <tan>33333333Y</tan>
    <language>en</language>
  </patientInfo>
  - <propheticSubstances>
    - <propheticSubstance>
      <patientLocation>
        <pageNumber>300</pageNumber>
        <pageLabel>400</pageLabel>
        <exampleLabel>100</exampleLabel>
      </patientLocation>
      <names>
        <name lang="en">SSSSSSSSSSSSSSSS</name>
        <IUPACName>IUPAC name not provided</IUPACName>
      </names>
      <structureData type="MOL Molfile V2000"
        encoding="Base64">KD3TLDN8KS0yLWfTaWSvLTMrtaHlk0m94e81idxRhbmo9pYyBhY2lkCgAgTWFydmiluICAwNDISMTAxNDUyMkQgIC
      </propheticSubstance>
    - <propheticSubstance>
      <patientLocation>
        <pageNumber>14</pageNumber>
        <pageLabel>8</pageLabel>
        <exampleLabel>6</exampleLabel>
      </patientLocation>
      <names>
        <name lang="en">No name</name>
        <IUPACName>3-Hydroxy-4-methyl-2-[methyl(methylene)-1-(5-azanyl)pent-4-enolic</IUPACName>
      </names>
      <structureData type="MOL Molfile V2000"
        encoding="Base64">M1hbWluby0zLWh5ZHveHktYnV0YW5vaWMgYWhpZAgIE1hcncZpbAgMDQyOTEwMTQ1MjDEICAgICAgICAg
      </propheticSubstance>
    - <propheticSubstance>
```

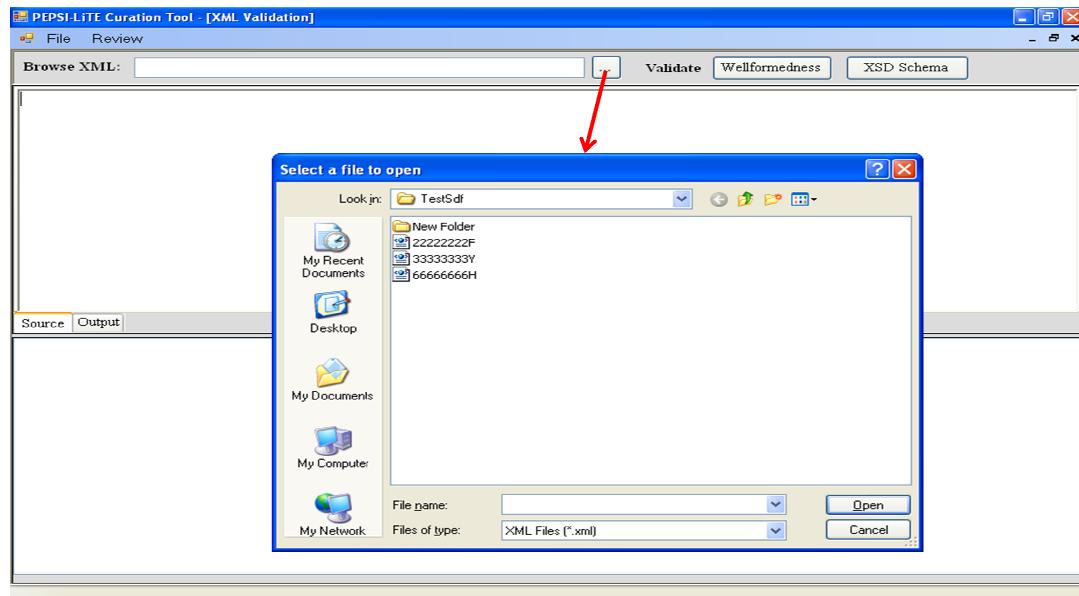
**Screen No: 13**

If the user wants to validate .xml file manually, he/she have the option to validate .xml file in the **Validate XML** submenu in the **Review** menu. Once click on Validate XML menu, Validate XML form will be displayed to the user, as shown in **Screen No:14**.



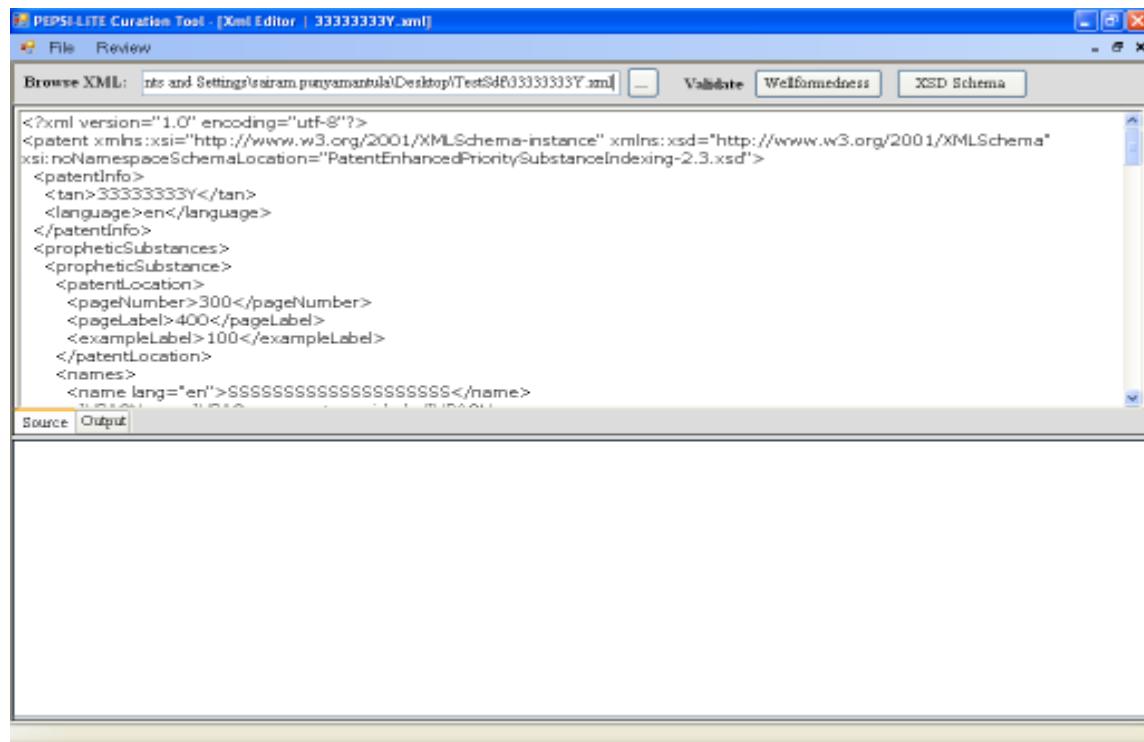
**Screen No:14**

Once the user click on Browse XML button open file dialog will be displayed to the user, as shown in **Screen No:15**.



**Screen No:15**

When the user select any xml file, that xml file content will be displayed in the Source tab of the Validate Xml form, as shown in **Screen No:16**.



The screenshot shows the PEPSI-LITE Curation Tool interface. The title bar reads "PEPSI-LITE Curation Tool - [Xml Editor | 33333333Y.xml]". The menu bar includes "File" and "Review". Below the menu is a toolbar with "Validate", "Wellformedness", and "XSD Schema" buttons. The main area is titled "Browse XML:" and shows the XML code for "33333333Y.xml". The XML content is as follows:

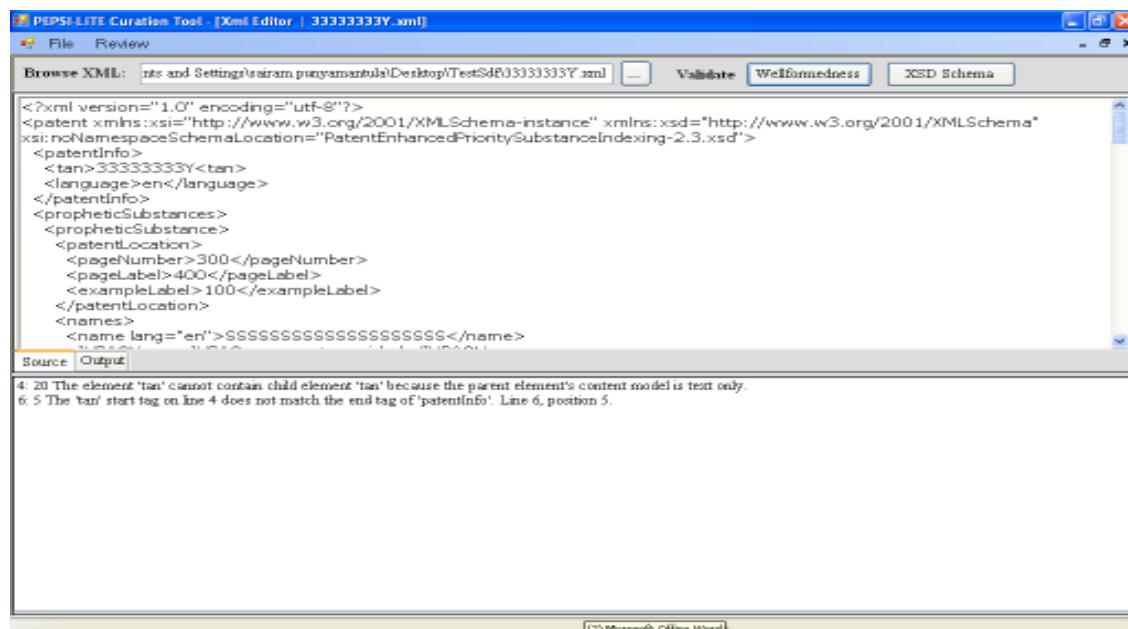
```
<?xml version="1.0" encoding="utf-8"?>
<patent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xsi:noNamespaceSchemaLocation="PatentEnhancedPrioritySubstanceIndexing-2.3.xsd">
  <patentInfo>
    <tan>33333333Y</tan>
    <language>en</language>
  </patentInfo>
  <propheticSubstances>
    <propheticSubstance>
      <patentLocation>
        <pageNumber>300</pageNumber>
        <pageLabel>400</pageLabel>
        <exampleLabel>100</exampleLabel>
      </patentLocation>
      <names>
        <name lang="en">SSSSSSSSSSSSSSSSSS</name>
      </names>
    </propheticSubstance>
  </propheticSubstances>
</patent>
```

Below the XML editor, there are tabs for "Source" and "Output".

**Screen No: 16**

User have to option to validate xml file either Wellformedness or XSD Schema.

If any errors in the validation process, those errors will be displayed in the error panel, as shown in **Screen No:17**.



The screenshot shows the PEPSI-LITE Curation Tool interface, similar to Screen No:16, but with validation errors present. The "Wellformedness" button is highlighted in the toolbar. The error panel at the bottom displays two messages:

- 4: 20 The element 'tan' cannot contain child element 'tan' because the parent element's content model is text only.
- 6: 5 The 'tan' start tag on line 4 does not match the end tag of 'patentInfo'. Line 6, position 5.

**Screen No:17**

When user click on any error in the error panel, the user will be directed to the corresponding error in the xml file, as shown in **Screen No:18**

The screenshot shows the PEPSI-LITE Curation Tool interface. The main window title is "PEPSI-LITE Curation Tool - [Xml Editor | 33333333Y.xml]". The menu bar includes "File" and "Review". Below the menu is a toolbar with buttons for "Validate", "Wellformedness", and "XSD Schema". The "Wellformedness" button is highlighted. The "Browse XML" tab is selected, showing the XML code:

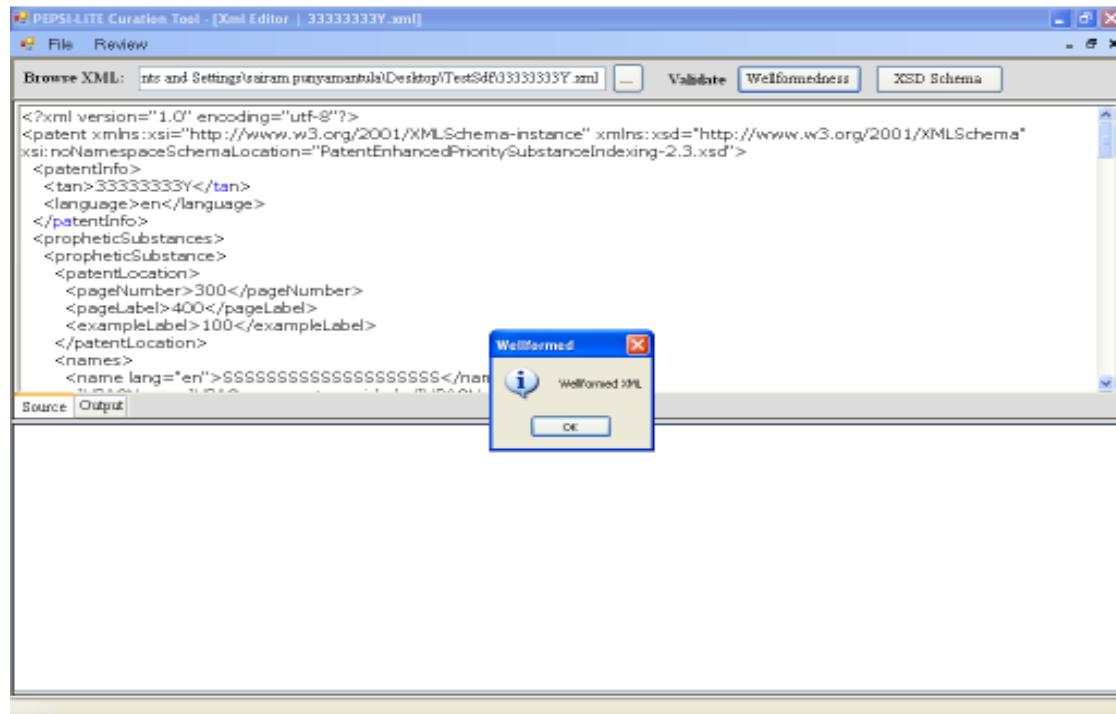
```
<?xml version="1.0" encoding="utf-8"?>
<patent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xsi:noNamespaceSchemaLocation="PatentEnhancedPrioritySubstanceIndexing-2.3.xsd">
<patentInfo>
  <tan>33333333Y</tan>
  <language>en</language>
</patentInfo>
<propheticSubstances>
  <propheticSubstance>
    <patentLocation>
      <pageNumber>300</pageNumber>
      <pageLabel>400</pageLabel>
      <exampleLabel>100</exampleLabel>
    </patentLocation>
    <names>
      <name lang="en">SSSSSSSSSSSSSSSSSS</name>
    </names>
  </propheticSubstance>
</propheticSubstances>
```

The "Output" tab is selected. A message box at the bottom displays two validation errors:

- 4: 20 The element 'tan' cannot contain child element 'tan' because the parent element's content model is text only.
- 6: 5 The 'tan' start tag on line 4 does not match the end tag of 'patentInfo'. Line 6, position 5.

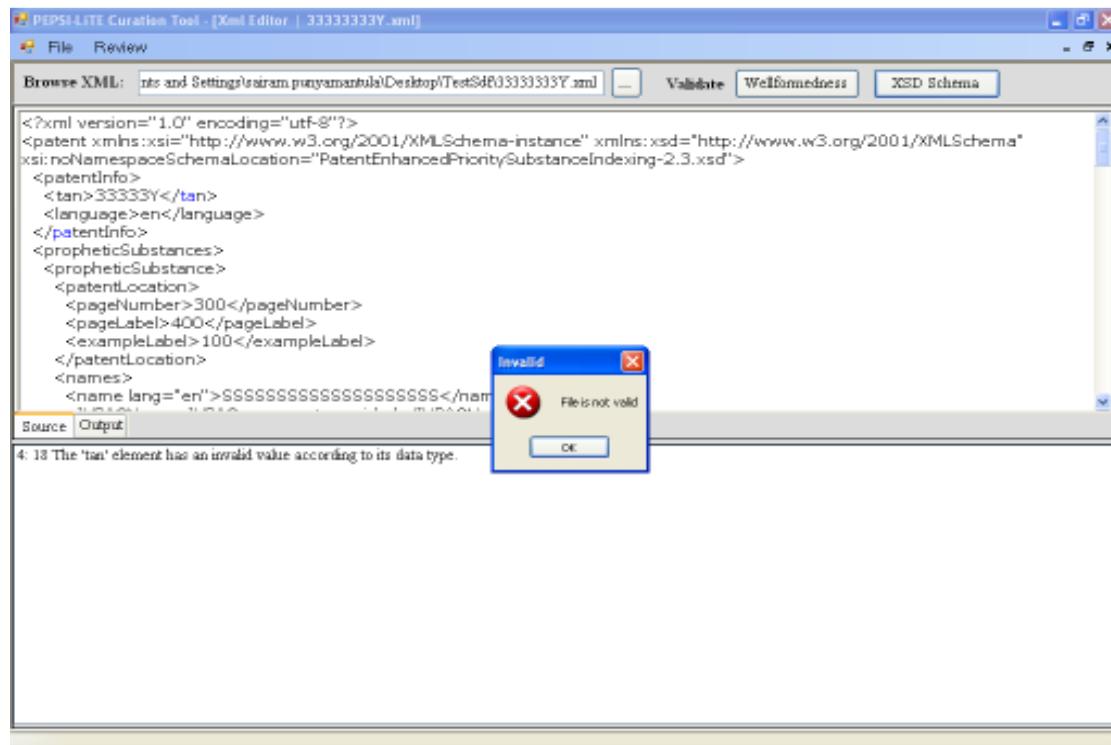
**Screen No: 18**

Once the errors are rectified and click on Wellformedness button valid status will be displayed to the user as an message, as shown in **Screen No:19**.



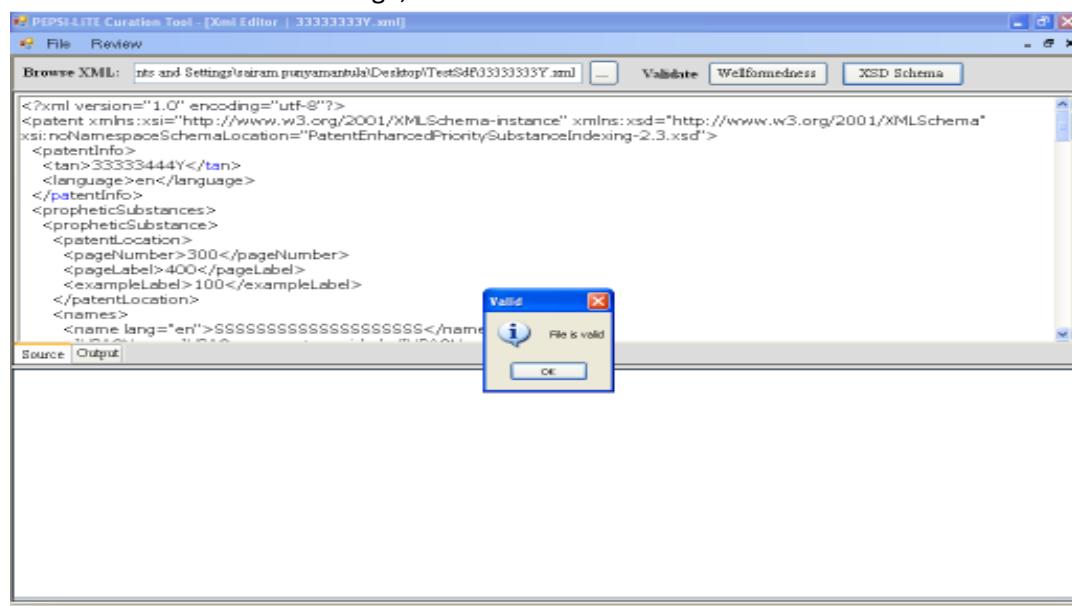
**Screen No:19**

When the xml file is wellformeded xml file, then click on XSD Schema button to validate against the given .xsd schema. If the xml file is not valid according to the rules specified in the .xsd schema, valid status will be displayed to the user as an message, as shown in **Screen No:20**.



**Screen No:20**

Once the errors are corrected in the xml file, click on XSD Schema button, the valid status will be shown to the user as an message, as shown in **Screen No:21**.



**Screen No:21**