

In general:

We need a **C#**.net assembly (version 2.0 or higher, with full source code) to scrape Wikipedia page in English and in Hebrew (Hebrew is right to left language).

Two main use cases:

1. Convert a specific wikipedia page to a predefined xml format.
2. Convert the home page's special features (Today's featured article, in the news etc.) to a predefined xml format.

Use case: Convert a specific definition page

The assembly should generate a predefined xml from a plain text definition by scraping the Wikipedia page corresponding this plain text definition.

The definition is a free text definition, so there are two scenarios for a given definition:

1. There is an exact Wikipedia definition matching the free text (for example "Lance Armstrong"), in that case the result is the definition page formatted in xml (http://en.wikipedia.org/wiki/Lance_armstrong)
2. There is no exact Wikipedia definition matching the free text, in that case the result is the page of the "closest" definition formatted in xml (see this page: <http://en.wikipedia.org/wiki/Bib>)

Use case: Convert the home page's special features

Converts a specific language's home page into XML.

The relevant parts are the just the "features" like "In the news", "Today's featured article" etc.

Interface:

The assembly's will implement this interface:

Public WikiDefinitionPage GetDefinitionPage(string definition, Languages language) – Should find the corresponding Wikipedia page for the definition and scrape it into wiki definition

Public WikiDefinitionPage GetHomePage(Languages language) – Should scrape the language's home page it into wiki definition

Public WikiDefinitionPage GetDefinitionPage(string wikipediaPageURL) – Should scrape the wikipediaPageURL into wiki definition

XML format:

This is an example xml:

```

<document Title='title' Definition='Definition' URL='http://xxx'>
- <abstract>
  - <paragraph>
    - <line>
      <format type='bold' size='12'>Title</format>
    </line>
    - <line>
      Cnidarians get their name from
      <a href='http://en.wikipedia.org/wiki/Cnidocyte'>cnidocytes</a>
      , which are specialized cells that carry
    </line>
    <line>Body shape of a cnidarian consists of a sac containing a gastrovascular cavity with a single</line>
  </paragraph>
  - <paragraph>
    <line>Traditionally the hydrozoans were considered to be the most primitive</line>
    <line type='bullet'>First issue is</line>
    <line type='bullet'>Second issue is</line>
    <line type='bullet'>Third issue is</line>
  </paragraph>
</abstract>
- <TOC>
  - <paragraph>
    - <line type='number' value='1'>
      <a href='http://en.wikipedia.org/wiki/Cnidaria#Nutrition'>Nutrition</a>
    </line>
    - <line type='number' value='2'>
      <a href='http://en.wikipedia.org/wiki/Cnidaria#Color'>Color</a>
    </line>
  </paragraph>
</TOC>
- <section name='Nutrition'>
  - <paragraph>
    - <line>
      <format type='bold' size='12'>Nutrition</format>
    </line>
    <image src='http://pic.com/gds234' alt='Puppet' width='80%' height='20%' />
    <line type='tab'>The big fish</line>
    <line>Always looks at the sea</line>
    <line type='horizontal_line' />
  </paragraph>
</section>
- <section name='Color'>
  - <paragraph>
    - <line>
      <format type='bold' size='12'>Color</format>
    </line>
    <line type='tab'>The colors</line>
    <line>Green</line>
    <line>Yellow</line>
    <line type='horizontal_line' />
  </paragraph>
</section>
</document>

```

The xml supports the following features:

1. Document <Document Title="title" Definition="Definition" URL="http://xxx"> - Container for the entire XML
2. Area
 - a. Abstract Area <Abstract> - The abstract area

- b. TOC area <TOC> - Table of content of the definitions with links to internal part in the document
- c. Section Area <Section name="XX"> - Sections for the definition, this is the definitions body
- d. External links <ExternalLinks>
3. Paragraph <Paragraph> - Each area (section 2) is combined from a few paragraphs
4. Line – Each paragraph contains a few lines
5. Paragraph line types
 - a. Regular - <Line>Text</Line>
 - b. Bullet - <Line type="bullet">text text</Line>
 - c. Number - <Line type="number" value="1">text text</Line>
 - d. Tab - <Line type="tab">text text</Line>
 - e. Horizontal line - <Line type="horizontal_line"/>
6. Text format
 - a. Bold <Format bold="true">text</ Format >
 - b. Size <Format size="13">text</ Format >
7. Image - The percentage value is important, give the width/height in relative to the screen
8. Link text - When linking inside the page the link is like this: <http://en.wikipedia.org/wiki/Cnidaria#Nutrition> - # separates the link and the page part

The *Languages* enum supports only English and Hebrew.

Special issues:

1. The assembly should support both English and Hebrew languages.