**Steps of Extraction, Transformation and Consolidation.**

# **Steps of Data Extraction:**

1. We use Visual Web Ripper to extract data from targeted website. It is a powerful web scrapping tool that is used to scrape data from web according to our requirements. Here are the steps followed while extracting a site for recipes and its data:
2. Copy and paste the URL of targeted website’s recipe section into Visual Web Ripper.
3. Load the webpage in Ripper.
4. Select recipe category as page area.
5. Select recipe category as link (to open a recipe category).
6. Select recipes in a category as page area.
7. Select recipes in a category as link (to open a recipe as a whole).
8. Select and save the attributes as elements in content tab. We use a recipe schema to follow the naming conventions for different attributes required by us in extraction. The schema along with all the naming conventions of attributes and their descriptions can be found at <https://schema.org/Recipe>
9. Select and save collection the attributes (ingredients, instructions, nutrition, and equipment/tool) as page area in templates tab.
10. Use page areas to select and store half the collection of the attributes (ingredients, instructions) as a whole unified content.
11. Use page areas to select and store half the collection of the attributes (nutrition, and equipment/tools) as a separate unified collection of content.
12. If Subheadings are found in ingredients and instructions, we use a separate template to extract these values (group and howToSection).
13. Once the whole page and all possible variations are selected for extraction, we save the project as **.rip** file and run the project and it starts to extract recipe data from the website.
14. The extracted data is generated and the exported in 3 formats XML, Excel Worksheet and database file for transformation into SQL.
15. Following are some major recipe attributes and their description that are required to extract:

|  |  |
| --- | --- |
| RECIPE PROPERTY | DESCRIPTION |
| **url** | a recipe url that tells the address of the page where the recipe lies |
| **name** | a recipe’s name identifies it |
| **image** | a url that links us to the image showing what’s its final outlook like |
| **author** | a person or organization, who produces the content of the recipe |
| **cookTime** | Cook Time is the one that tells the duration it takes to cook a recipe |
| **prepTime** | Prep Time would be the time to do all the initial preparations as per directions |
| **totalTime** | Total Time is the time it took to perform all the instructions and complete the recipe |
| **marinateTime** | Marinate time, is used in the recipes where meat takes time to get marinated |
| **recipeCuisine** | Recipe Cuisine is the cuisine of the recipe, for example, Italian pasta |
| **recipeYield** | Recipe Yield denotes how many people does a recipe serve |
| **yield** | Yield to indicate the quantity it will produce |
| **cookingMethod** | Cooking Method tells the procedure we undertake to cook the recipe i.e frying, steaming, baking, grilling, etc. |
| **commentCount** | There is a Comment Count indicating the number of comments the recipe has received so far |
| **dateCreated** | Date Created is the date on which recipe was created |
| **datePublished** | Date Published is the date when the recipe was published and introduced to the public |
| **dateModified** | Date Modified is the recent date when the recipe was edited |
| **keywords** | Keywords are the tags that are used to search |
| **aggregateRating** | Aggregate Rating represents the overall ratings of a recipe, it is calculated in the backend by taking an average of the ratings collected so far |
| **contentRating** | Content rating, the official rating the recipe has gotten |
| **nutrition** | Nutrition indicates the nutrient content a recipe consists of like calories, fat content, protein content, fiber content, and vitamins |
| **estimatedCost** | Estimated cost shows the cost of raw materials or supplies |
| **recipeIngredients** | Recipe Ingredients is the one representing a list of all the ingredients used to make the recipe |
| **recipeIngredient** | Recipe Ingredient is the content showing a single ingredient from the ingredient list |
| **recipeInstructions** | Recipe Instructions are all the steps that make up the “how-to” part of the recipe, mainly the directions to follow |
| **step** | step is a single direction from the recipe instructions list |

# **Steps of Data Transformation:**

1. Once the data has been successfully extracted, and exported, we convert the data into a SQL database file using SQLite Studio. The steps are as follows:
2. Select the file named “internal\_websitename”, which is database file from the output folder of the website’s Visual Web Ripper Project.
3. Open SQLite Studio.
4. Load the internal file into SQLite Studio using the add database option.
5. Once Loaded, use the export database option to the same output folder as an SQL file of database with extension “.**sql**”.
6. Once the .sql file of the project is successfully exported, we apply 5 changes in that SQL file to make it compatible and according to our data requirements of finalized database.
7. The five changes are done by using Find and Replace option of Notepad. The five changes are as follows:
   1. Replace **"** with **`**
   2. Replace **[** with **`**
   3. Replace **]** with **`**
   4. Replace **nvarchar(4000)** with text
   5. Replace **guid** with **char(36)**
   6. Also, Comment PRAGMA BEGIN at start and at end.
8. Once these changes are done, the data is successfully transformed into our desired SQL database, we upload it onto our SQL Server at the address: <http://121.52.153.204/phpmyadmin/index.php>

# **Steps for Data Consolidation:**

After the recipe data of different websites has been extracted and transformed and uploaded to the server. We consolidate all the data by using MySQL scripts and MySQL Workbench to move and cleanse all the data from separate databases to one unified database (Group1). The steps involved are:

1. Install and open MySQL Workbench.
2. Setup a database connection using login credentials of the database server.
3. Select the website database to use.
4. If there are multiple parts of a website’s database, then they are merged before consolidation.
5. Alter the recipe table by adding a column named “idid” after the “url” column.
6. Fill the idid with recipe IDs using its ID from Extraction Sheet and filling the column in sequence using the format of ID+000000 and respective values onwards from 1,2,3….
7. Create two tables named “temp” and “temp2” to separate duplicated and unique recipe records from recipe table.
8. The table “temp” will store the row\_id and url of the recipes that are duplicated in the database (once or even more than one) and the table “temp” will store the row\_id and url of the recipes that are unique and found only once in the recipe table of website’s database.
9. After that, we take up all the attributes from the recipe table of website’s database and consolidate into the recipe table of the finalized database (Group1.Recipes) using its “idid” to uniquely identify each recipe and website it belongs to.
10. After recipe table is successfully consolidated, we move on to consolidating other three tables (ingredients, instructions, nutrition).
11. Using separate MySQL queries for subheadings and without subheadings we consolidate all the values (idid, source index, groupName, and recipeIngredient) of ingredients table of website’s database into ingredients table of the finalized database (Group1.Ingredients) using its “idid” as unique identifier.
12. Just like ingredients, using separate MySQL queries for subheadings and without subheadings we consolidate all the values (idid, source index, howToStep, and recipeInstruction/step) of instructions table of website’s database into instructions table of the finalized database (Group1.Instructions) using its “idid” as unique identifier.
13. Nutrition data is consolidated as same as the recipe table if they are extracted inside the recipe table and it is done because of different values of different nutrition present in it. We use a separate query if the nutrition is extracted as a separate table.
14. We consolidate all the values (idid, nutrition name) of nutrition table of website’s database into instructions table of the finalized database (Group1.Nutrition) using its “idid” as unique identifier.
15. Once all the tables are filled, a website’s consolidation is completed.

