IndicWiki Summer Internship

Rivers



[Domain](#_Domain) [Team](#_Team)

Data Collection [Sources/ Sites](#_Sources/_Sites)

[Tools used for Data collection](#_Tools_used_for)

Data Cleaning

Sample Article

Jinja template creation

Translation/ Transliteration

XML Generation

Final data

# Domain

The domain we worked on was **“Rivers”**, the aim of the project being generating comprehensive articles for Telugu Wikipedia on 27000+ rivers, comprising all possible details on a river

# Team

|  |  |
| --- | --- |
| **Member** | **Email** |
| Annareddy Divya | reddydivya170@gmail.com |
| Yadavalli Varshitha | 19wh1a0556@bvrithyderabad.edu.in |
| Sreeshanth Ganduri | sreeshanthmpsganduri@gmail.com |
| Velumuri Aravinda Aparna | vaaparna20@gmail.com |

Data Collection

## Sources/ Sites

We searched for data in various websites, and we found a website which is solely dedicated in providing rivers and its details to its viewers and allows to query its database –Dbpedia. Apart from this website, we scraped data from Wikipedia info boxes as well.

* Dbpedia
  + Link: <https://dbpedia.org/sparql/>
  + Format of data available – Database
  + Tools used – Sparqlwrapper, python
  + Attributes found
* River name, length, mouth elevation, mouth mountain, mouth place, river mouth, geometry, mouth location, other names, source, abstract, subdivision, thumbnail, latitude, longitude, source location, discharge 1 location, discharge 2 location, left tributaries, right tributaries.
* Wikipedia
  + Link: <https://www.wikipedia.com/>
  + Format of data available - Web Pages (HTML)
  + Tools used - Beautiful Soup (Python Libraries)
  + Attributes found
    - We have found around 290 attributes from the Wikipedia, but for a river on an average there are 20-30 attributes. So we removed at the remaining attributes during cleaning.

## Tools used for Data collection

### Beautiful Soup

* + This was used to navigate through HTML elements and obtain text/information from them, hence, **no major issues** were observed.
* **Python pandas**
  + A famous python library which is used to efficiently handle datasets present in csv or xlsx formats
* **Sparqlwrapper**
* This is a python module used to wrap python code and sparl code together

Data Cleaning

* Data cleaning has been done on the data collected and removed the attributes which are not necessary.
* Remove the rows with a smaller number of attributes.
* Cleaned the duplicate rows and corrected them.
* After cleaning, we created structured data.

Translation and Transliteration

* After cleaning the data , translation and transliteration is done.
* To translate the data, google translator is used. We translated each column by uploading the file to the Google’s translator website - <https://translate.google.co.in/?sl=auto&tl=te&op=translate>
* Apart from it, Bing Translator and Yandex are also used
* We also used anuvaad for transliteration.

Sample Article

* Generated sample articles for some rivers
  + Sample article for a river with more attributes article – [krishna river](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%95%E0%B1%83%E0%B0%B7%E0%B1%8D%E0%B0%A3%E0%B0%BE_%E0%B0%A8%E0%B0%A6%E0%B0%BF(%E0%B0%89%E0%B0%A6%E0%B0%BE%E0%B0%B9%E0%B0%B0%E0%B0%A3)&oldid=882017)
  + Sample article for a river with less attributes article – [umiray river](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%89%E0%B0%AE%E0%B0%BF%E0%B0%B0%E0%B1%87_%E0%B0%A8%E0%B0%A6%E0%B0%BF)

Jinja Template Generation

//yet to write

XML generation

// yet to write

Final Data:

Our final data set consists of both translated and English datasets

* English data set – Mastersheet.xlsx
* Final Dataset – Final\_translated.xlsx