[Domain](#_ks7guwt0lse)

Team

Data Collection

Sources/ Sites

Tools used for Data collection

Images

Data Storing

Data Cleaning

Cases taken care of

Version Control

Sample Article

Sections

Jinja template creation

Edge cases

Categories,

References Infobox

Translation/ Transliteration

Libraries

Common Issues while translation, transliteration

XML Generation

Quality Review

Github Structure

# 

# **Domain**

**RELIGIOUS PLACES**

The domain we worked on was *Religious Places* across the world, which involved the generation of 20K+ Wikipedia articles in Regional Language (in our case it was Telugu).

**Subdomains**

Our domain includes many subdomains that we had to work on individually. The subdomains are:

1. Mosques
2. Churches
3. Temples
4. Religious Buildings
5. Gurudwaras
6. Jain Temples

# **Team**

| **Name** | **Email Id** |
| --- | --- |
| Tharun Gade | tharungade2001@gmail.com |
| Geethika | srigeethika2002@gmail.com  [19311a1212@sreenidhi.edu.in](mailto:19311a1212@sreenidhi.edu.in) |

**Data Collection**

## **Sources/ Sites**

We explored many resources during the starting days of the internship to find valuable data, but couldn't find any that have many attributes and have the sparseness that we wanted. So we moved to English Wikipedia articles, we have only considered the religious places that have infoboxes and have generic attributes in each of the subdomains.

Our main resource for scrapping the data involved Wikipedia, as Wikipedia is a huge resource we could gather so much information for our domain of religious places across the world. We also tried scraping data from other resources such as templenet, etemple, dinamalar, etc.

We were able to gather many subdomains of data. We mainly focused on

religious buildings, churches, and gurudwaras.

**References**:

**Wikipedia links:**

Religious Buildings:

<https://en.wikipedia.org/wiki/Special:WhatLinksHere/Template%3AInfobox%20religious%20building?hidelinks=1&hideimages=1>

Churches:

<https://en.wikipedia.org/wiki/Special:WhatLinksHere/Template%3AInfobox%20church?hidelinks=1&hideimages=1>

Hindu Temples:

<https://en.wikipedia.org/wiki/Special:WhatLinksHere/Template%3AInfobox%20Hindu%20temple?hidelinks=1&hideimages=1>

Gurudwaras:

<http://etemple.com/temple-type/gurudwara/>

**Attributes of Religious Buildings:**

1. name
2. imageSrc
3. Affiliation
4. location
5. Deity
6. Rite
7. festival
8. Sect
9. Municipality
10. District
11. State
12. Province
13. Country
14. governingBody
15. Leadership
16. Architect
17. Type
18. Style
19. Founder
20. Creator
21. consecrationYear
22. Date established
23. Groundbreaking
24. completed
25. Construction cost
26. Capacity
27. length
28. Width
29. heightMax
30. interior area
31. Minaret
32. dome
33. Minaretheight
34. Materials
35. elevation
36. website
37. LatitubeLongitude

**Attributes of Churches:**

1. Name
2. imageSrc
3. Location
4. District
5. Country
6. Province
7. Tradition
8. Denomination
9. Membership
10. Weekly attendance
11. Website
12. Founded
13. Founders
14. Dedication
15. consecrated
16. Events
17. Architects
18. Architectural type
19. Style,
20. Year built,
21. Groundbreaking,
22. Completed,
23. Construction cost
24. Capacity
25. Length
26. Width
27. Height
28. Chancellor
29. Priest
30. Minister
31. Pastor
32. latitude longitude

**Attributes of gurudwaras:**

1. Name
2. Location
3. Nearbygurudwaras
4. Type
5. Affiliation
6. Country
7. Deity
8. Prasad
9. Festival
10. State
11. Community

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Tools used for Data collection**

### **BeautifulSoup**

Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree. It is used to scrape static web pages. Easy to use and fewer errors to handle. Most of the data was scraped using this module.

### **Selenium**

It is used to scrape data dynamically ( involving clicking HTML elements etc.. ). We used this module to scrap religious places' names and their Wikipedia links as we couldn't observe any similarities in Wikipedia links where we had to scrape from.

## Images

* Source: Wikimedia commons Infoboxes image attribute

## **Data Storing**

We stored all the data in excel and google sheets as we found it easy to access them using python. As we were two members in the team and it was easier for us to access the datasets from Google drive.

# **Data Cleaning**

We had to clean the data as the data we scraped was inconsistent, some data points we had to clean manually which were mainly non - translated values. We used anuvaad and bing translate for translating the values from English to Telugu.

**Key Takeaways:**

- Ensure while scraping data from Wikipedia citation text ( ex: [1] ) are replaced or deleted.

- Ensure the data scraped shouldn’t contain any special characters ( like \n \t ) that text generally contains.

- trim ( remove leading and trailing spaces ) the data before adding it to the dataset

- Ensure that the values in the final data set are completely in telugu and not in any other language.

**Version Control**

* For version control of files used in the project, we used the IIIT GitHub repository, which was especially useful when tasks associated with data cleaning and editing the jinja template had to be performed when multiple people were working on various versions of the template/data which were sequentially updated.

# 

# 

# 

# 

# 

# 

# 

# 

# **Sample Article**

To review the sample templates that we wrote, we published a few sample articles on the tewiki website to look at the presentation and size of the article. It was a nice way of interpreting the data that we had collected. Here are some of the sample articles present in the tewiki website:

[Temple sample article](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%95%E0%B0%BE%E0%B0%AE%E0%B0%BE%E0%B0%95%E0%B1%8D%E0%B0%B7%E0%B0%BF_%E0%B0%85%E0%B0%AE%E0%B1%8D%E0%B0%AE%E0%B0%A8%E0%B1%8D_%E0%B0%86%E0%B0%B2%E0%B0%AF%E0%B0%82#.E0.B0.AA.E0.B0.82.E0.B0.A1.E0.B1.81.E0.B0.97.E0.B0.B2.E0.B1.81)

The above link leads to the article on Kamakshi Amman temple.

[Mosque sample article](https://tewiki.iiit.ac.in/index.php?title=%E0%B0%9C%E0%B0%BE%E0%B0%AE%E0%B0%BE_%E0%B0%AE%E0%B0%B8%E0%B1%80%E0%B0%A6%E0%B1%81,_%E0%B0%A2%E0%B0%BF%E0%B0%B2%E0%B1%8D%E0%B0%B2%E0%B1%80&venotify=created)

The above link leads to the article on Jama Masjid, Delhi.

Writing the sample articles was helpful as we got to know what the final output of our work looked like.

**Sections**

All the subdomains have 3 to 4 sections. We made sure that all the sections contained sentences that were relevant to the section.

**Sections of religious buildings:**

1. **introduction (పరిచయం)**

Attributes contained: name, type, country, state, district, affiliation, municipality, province, deity, rite

1. **History (చరిత్ర)**

Attributes contained: name, type, groundbreaking, consecrationYear, date established,

1. **Construction (నిర్మాణం)**

Attributes contained: name,type,style,architect,creator,Constructioncost, materials

1. **Description (వివరణ)**

Attributes contained: name, type, capacity, length, width, interior area,

minaret, minaret height, province, elevation, festival, dome

**Sections of churches:**

1. **Introduction (పరిచయం)**

Attributes contained: name,location,district,province,country,denomination

1. **History (చరిత్ర)**

Attributes contained: name, founded, founders, dedication, consecrated,

events, tradition

1. **Construction (నిర్మాణం)**

Attributes contained: architects, style, years built, groundbreaking, completed, construction cost, capacity, height, length, width

1. **Current scenario ( ప్రస్తుత దృశ్యం )**

Attributes contained: chancellor, priest, minister, membership, pastor, weekly attendance, website

**Sections of Gurudwaras:**

1. **Introduction (పరిచయం):** Contained the introduction part of the gurudwara, its name, location, and deity.
2. **Prasad(ప్రసాదం):** It mainly focused on the importance of langar. Which is a main part of the sikh tradition.
3. **Specialities (ప్రత్యేకతలు):** Specialities included the importance of the place, festivities and traditions.

# 

# 

# 

# 

# 

# 

# 

# **Jinja template creation**

## **Edge cases**

- Ensure that Nan / Null values are not rendered in the template.

- If a sentence has more than one attributes make sure you handle the cases

where all attributes don't occur simultaneously.

- always pass a dictionary to the template as accessing the dictionary in the

Template is more easy and more intuitive (makes the template more readable )

## **Categories and References**

-Scrapped references, external links, and categories that we rendered in respective sections.

- If a Wikipedia article has multiple links for references or external links scraped a fixed ( 5-7 links ) number of links.

- all the categories are included in its section if present in Wikipedia article.

## **Infobox**

We have written our own custom Infobox for each of the subdomains and are able to render them based on NaN values.

**Jinja template of Religious buidings Infobox:**

{{'{{Infobox '}}

| bodystyle = width:25em;text-align:center;

| abovestyle = background-color:#cfc;

| headerstyle = background-color:#ccf;text-align:center;

| labelstyle = background-color:#ddf; width:20%;text-align:left;

| title = {{name}}

| label1 = చిత్రం

| data1 = {{imageSrc}}

| label2 = అనుబంధం

| data2 = {{affiliation}}

| label3 = స్థానం

| data3 = {{location}}

| label4 = దైవం

| data4 = {{deity}}

| label5 = ఆచారం

| data5 = {{rite}}

| label6 = పండుగ(లు)

| data6 = {{festival}}

| label7 = శాఖ

| data7 = {{sect}}

| label8 = మున్సిపాలిటీ

| data8 = {{municipality}}

| label9 = జిల్లా

| data9 = {{district}}

| label10 = రాష్ట్రం

| data10 = {{state}}

| label11 = పరిపాలనా విభాగం

| data11 = {{province}}

| label12 = దేశం

| data12 = {{country}}

| label13 = ముడుపు సంవత్సరం

| data13 = {{consecrationYear}}

| label14 = పరిపాలన సంస్థ

| data14 = {{governingBody}}

| label15 = నాయకత్వం

| data15 = {{leadership}}

| label16 = వాస్తుశిల్పి

| data16 = {{architect}}

| label17 = రకం

| data17 = {{type}}

| label18 = శైలి

| data18 = {{style}}

| label19 = స్థాపకుడు

| data19 = {{founder}}

| label20 = సృష్టికర్త

| data20 = {{creator}}

| label21 = స్థాపించబడింది తేదీ

| data21 = {{dateestablished}}

| label22 = పునాది తేదీ

| data22 = {{groundbreaking}}

| label23 = పూర్తయిన తేదీ

| data23 = {{completed}}

| label24 = నిర్మాణ వ్యయం

| data24 = {{constructioncost}}

| label25 = సామర్థ్యం

| data25 = {{capacity}}

| label26 = పొడవు

| data26 = {{length}}

| label27 = వెడల్పు

| data27 = {{width}}

| label28 = గరిష్ట ఎత్తు

| data28 = {{heightMax}}

| label29 = అంతర్గత ప్రాంతం

| data29 = {{interiorArea}}

| label30 = మినార్(లు)

| data30 = {{minaret}}

| label31 = డోమ్(లు)

| data31 = {{dome}}

| label32 = మినార్(లు) ఎత్తు

| data32 = {{minaretheight}}

| label33 = నిర్మాణ సామగ్రి

| data33 = {{materials}}

| label34 = ఎత్తు

| data34 = {{elevation}}

| label35 = వెబ్సైట్

| data35 = {{website}}

| label36 = అక్షాంశరేఖాంశం

| data36 = {{LatitubeLongitude}}

{{'}}\n'}}

{%- endmacro -%}

**Jinja Template of Churches:**

{{'{{Infobox '}}

| bodystyle = width:25em;text-align:center;

| abovestyle = background-color:#cfc;

| headerstyle = background-color:#ccf;text-align:center;

| labelstyle = background-color:#ddf; width:20%;text-align:left;

| title = {{name}}

| image = {{imageSrc}}

| label1 = స్థానం

| data1 = {{location}}

| label2 = జిల్లా

| data2 = {{district}}

| label3 = దేశం

| data3 = {{country}}

| label4 = ప్రావిన్స్

| data4 = {{province}}

| label5 = సంప్రదాయం

| data5 = {{tradition}}

| label6 =

| data6 = {{denomination}}

| label7 = సభ్యత్వం

| data7 = {{membership}}

| label8 = వారంవారీ హాజరు

| data8 = {{weeklyattendance}}

| label9 = వెబ్సైట్

| data9 = {{website}}

| label10 = స్థాపించారు

| data10 = {{founded}}

| label11 = వ్యవస్థాపకులు

| data11 = {{founders}}

| label12 = అంకితం

| data12 = {{dedication}}

| label13 = పవిత్రమైన

| data13 = {{consecrated}}

| label14 = సంఘటనలు

| data14 = {{events}}

| label15 = వాస్తుశిల్పులు

| data15 = {{architects}}

| label16 = నిర్మాణ రకం

| data16 = {{architecturaltype}}

| label17 = శైలి

| data17= {{style}}

| label18 = సంవత్సరాలు నిర్మించబడింది

| data18 = {{yearsbuilt}}

| label19 = భూస్థాపితం

| data19 = {{groundbreaking}}

| label20 = పూర్తయింది

| data20 = {{completed}}

| label21 = నిర్మాణ వ్యయం

| data21 = {{constructioncost}}

| label22 = సామర్థ్యం

| data22 = {{capacity}}

| label23 = పొడవు

| data23 = {{length}}

| label24 = వెడల్పు

| data24 = {{width}}

| label25 = ఎత్తు

| data25 = {{height}}

| label26 = ఛాన్సలర్

| data26 = {{chancellor}}

| label27 = పూజారి

| data27 = {{priest}}

| label28 = మంత్రి

| data28 = {{minister}}

| label29 = పాస్టర్

| data29 = {{pastor}}

| label30 = అక్షాంశరేఖాంశం

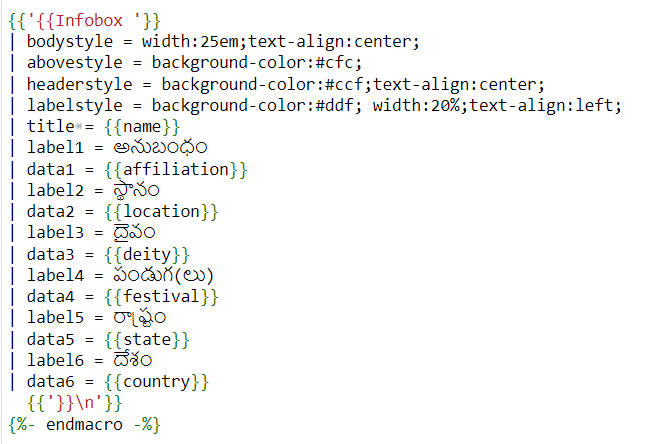
| data30 = {{LatitudeLongitude}}

{{'}}\n'}}

{%- endmacro -%}

# 

**Jinja template of Gurudwaras Infobox:**

****

# **Translation/ Transliteration**

* **Bing Translate**
* Most of the translation is done using bing translate.
* We have written a selenium code which reads the dataset and translate it to telugu using bing translate website (<https://www.bing.com/translator> )
* **Anuvaad**

State of the art open-source translation models for Indic languages.

* **Google Translate**

We have used this for cleaning. After the data translation some values are not able to translate by bing translate so we used google translate.If google translate also can’t translate we manually translate using google input tools.

**Issues with Translation and Transliteration**

The main issues we faced during translation using anuvaad was it was very slow while executing and was time taking. Also after translating the values, some values remained in English, which we needed to manually translate it to Telugu using google input tools.

# 

# **XML Generation**

This is the final step, it involved the generation of XML files. We have generated the XML articles according to the size i.e, we fixed size ( say 2000 ) and every XML file contains 2000 Religious places.

**New learnt and Errors**:

* In XML generation two jupyter notebook are used one ‘main.ipynd’ and other ‘genXML.ipynb’ so we had to use some functions in genXML.ipynb in main.ipynb so we used these libraries to to that:

**importlib**

**Import\_ipynb**

**import importlib**

**import import\_ipynb**

**import genXMLReligiousBuildings**

**importlib.reload(genXMLReligiousBuildings)**

**from genXMLReligiousBuildings import tewiki**

**from genXMLReligiousBuildings import writePage**

**importing Jupyter notebook from genXMLReligiousBuildings.ipynb**

**importing Jupyter notebook from genXMLReligiousBuildings.ipynb**

* Our XML involved rendering characters like ‘&’ (ampersand) which is a special character in XML and it was giving an error, so we had to use ‘&amp;’ in place of ‘&’.

# 

# Quality Review

* We made sure sentences with the NaN values are not rendered into the final article so that the article is more readable and complete.
* We have included randomization such that sentences don't look similar and each file would look different from the others in terms of sentence formation.
* We only considered the rows/data that have more than a certain number of attributes to maintain the size of the article consistently, so that articles would not be stubbed.

# Github Structure

Our Github repository [link](https://github.com/indicwiki-iiit/Religious-Places). Please follow this link to gain access to all the code we used for scraping, scraped dataset etc.