Session1

Main Agenda

1. Scraping / translation basics
2. Templating & xml generation

Expectations/Requirements

+ Images in/out wiki commons

+ Translators wrt context

+ Merging data from various sources

DATA

- Scraping
  - BeautifulSoup
  - Selenium
  - SPARQL for wikidata
  - Pdf reader
- Cleaning
  - Sweetviz
  - Logical / factual errors
    - Clean it with a rule based script
- Merging
  - Identify the primary/foreign key
  - Pandas

Translation/Transliteration

- Most Important
  - Structured data
  - Limit the data to word/phrases
- ML Tools
  - DeepTranslit
  - IndicBart
  - IndicTrans
  - indic_nlp_tools