



BUILDING A KNOWLEDGE GRAPH FOR SCIENTISTS USING WIKIDATA

Shivam Chaturvedi (2023201019)

Manas Kumar Biswas (2023201016)

Niraj Gupta (2023201013)

Shalu Kumari (2023201031)



Introduction

Objective:

The main objective of this project is to build a knowledge graph using data from Wikidata to organize detailed information about scientists. The knowledge graph is enhanced with:

- Translation to Hindi for accessibility to non-English speakers.
- Gender-specific sentence generation for grammatically correct and culturally appropriate descriptions.
- Template-based sentence generation, providing structured, automated narratives for each scientist.

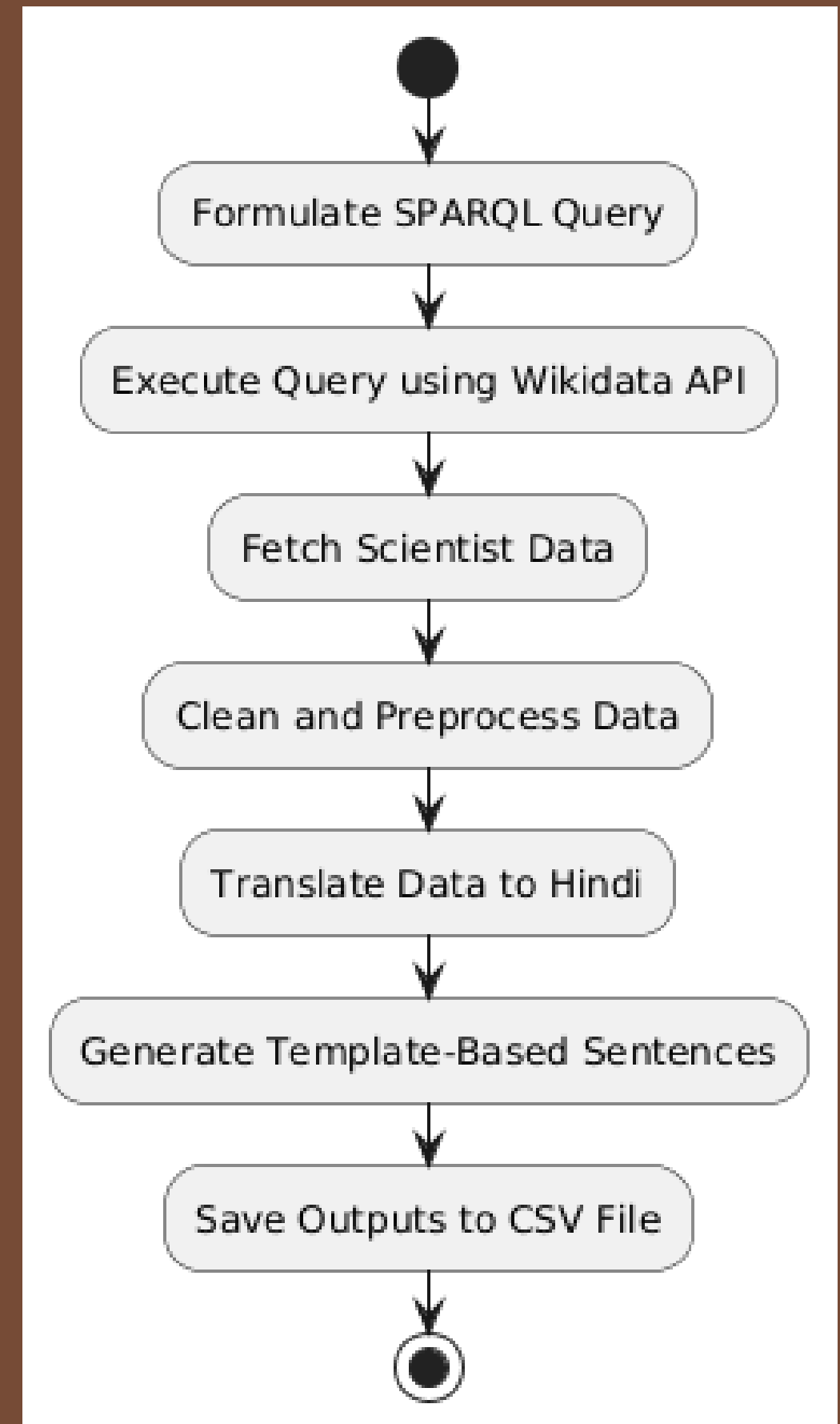


MOTIVATION

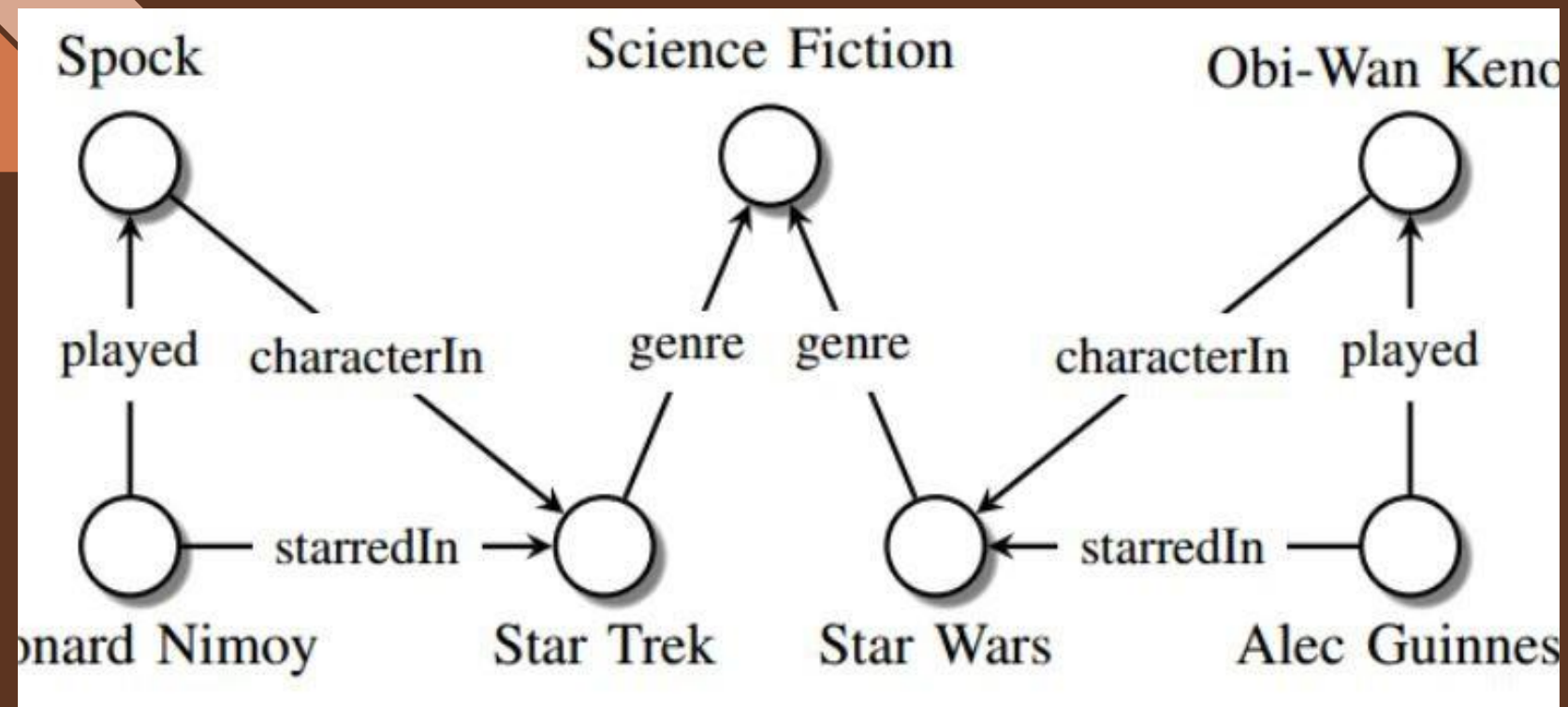
- **Accessibility of Knowledge:** A significant portion of India's population speaks Hindi as their first language. By translating key scientific information into Hindi, this project aims to make knowledge about scientists accessible to a wider audience, especially students and educators.
- **Automation of Knowledge Representation:** Manual entry of data into knowledge systems is time-consuming. Automating this process using NLP techniques and knowledge graphs makes large-scale data processing feasible.

PROCESS OVERVIEW

- SPARQL Query Execution to extract scientist data.
- Fetch additional details using Wikidata API.
- Clean and translate data to Hindi.
- Generate template-based sentences.
- Save outputs in CSV files.

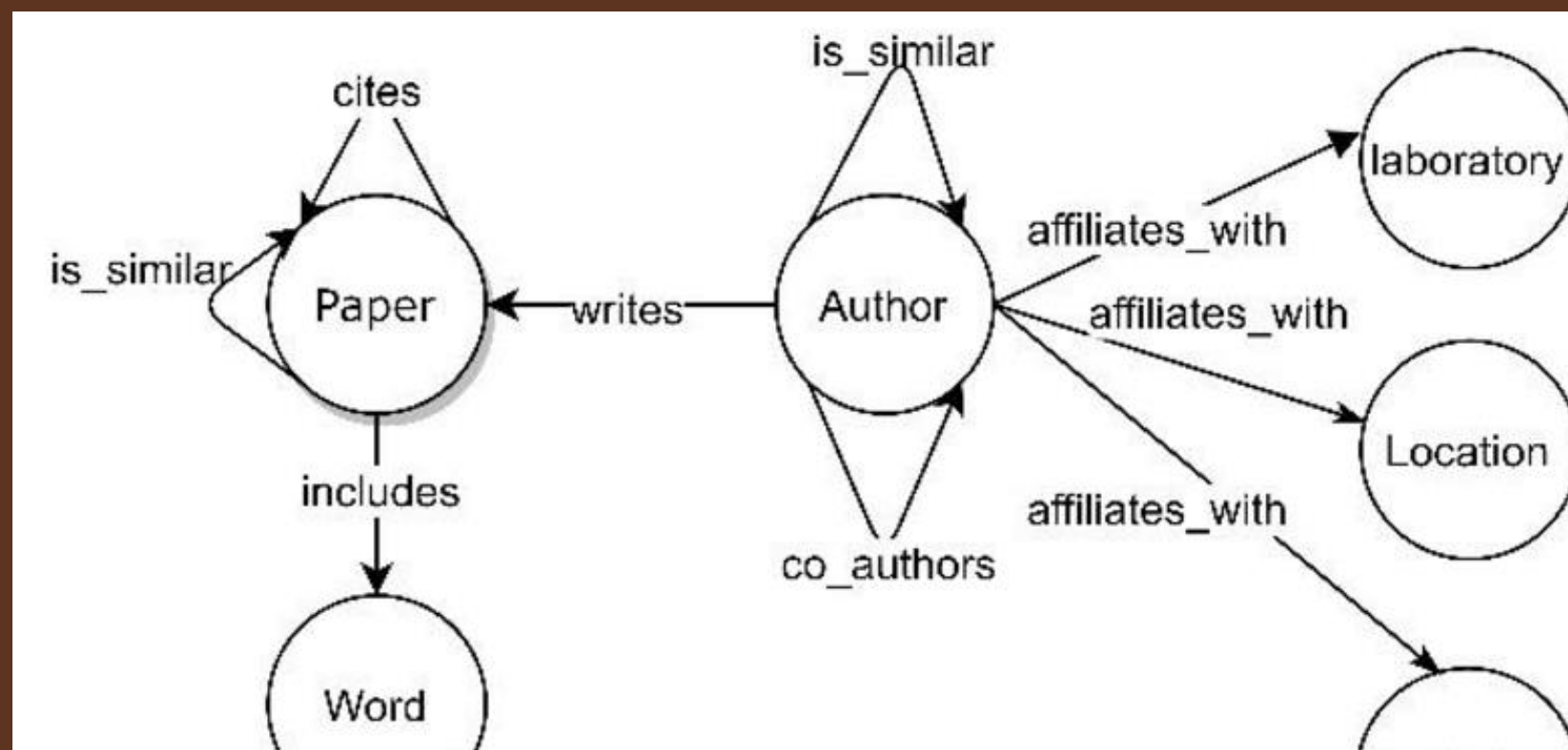


KNOWLEDGE GRAPHS



A knowledge graph (KG) represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts – where:

1. Descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner.
2. Entity descriptions contribute to one another, forming a network, where each entity represents part of the description of the entities, related to it.





WHAT IS WIKIDATA?

- A collaboratively edited knowledge base hosted by the Wikimedia Foundation
- Common source of open data for Wikimedia projects and anyone else, under a public domain license
- Primary data storage: JSON blobs in an SQL database, using specific IDs as the base of Wikidata
- Entity ID structure:
 - Items: Prefixed with Q. The core entities in Wikidata, representing concepts, people, places, etc., identified by unique QIDs (e.g., Albert Einstein (Q937))
 - Properties: Prefixed with P. Define relationships between items, identified by PIDs (e.g., instance of (P31))
 - Lexemes: Prefixed with L. Represent linguistic entities like words or phrases, identified by LIDs, capturing language data such as grammar. (e.g., L1)
- Wikidata Query Service: Uses an RDF triple store to allow SPARQL queries against current data

label

Douglas Adams (Q42)

item
identifier

description

English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
► In more languages

aliases

property

educated at

St John's College

value

end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971

qualifiers

rank

▼ 2 references

stated in Encyclopædia Britannica Online
reference URL http://www.nndb.com/people/731/000023662/
original language of work English
retrieved 7 December 2013
publisher NNDB
title Douglas Adams (English)

opened
references

statement
group

Brentwood School

end time 1970
start time 1959

► 0 references

collapsed
reference

+ add (statement)

SPARQL Query Example:

```
SELECT ?human ?label
WHERE
[
?human wdt:P31(instance of)
wd:Q15632617(fictional being);
rdfs:label ?label.
FILTER(LANG(?label) =
"en").
FILTER(STRSTARTS(?label,
"Mr. ")).
]
```



USE CASE

- Educational Tools: The knowledge graph can be used to automatically generate scientific profiles in Hindi, assisting teachers and students in accessing accurate, structured knowledge.
- Intelligent Systems: This work can be extended to chatbots, question-answering systems, or voice assistants that offer information in Hindi, making them useful for wider audiences.
- Research and Data Insights: The graph provides a structured format that can be extended to track research contributions, collaborations, and educational lineages (e.g., doctoral advisors and students).



DATA EXTRACTION


Data extraction involves retrieving structured data from Wikidata using SPARQL queries. This allows us to gather information such as scientists' names, birth dates, birth places, occupations, awards, and academic relationships.



DATA EXTRACTION

Key Points:

- SPARQL: A query language for accessing and retrieving structured data from Wikidata.
- The query extracts:
 - Name of the scientist.
 - Birthdate and birthplace.
 - Occupation (e.g., physicist, biologist).
 - Awards
 - Educational Institutions
 - Death dates
 - Descriptions
 - Occupations
 - Labels



```
def main():
    query = """
    SELECT ?scientist ?scientistLabel ?birthDate ?deathDate ?birthPlaceLabel
           (GROUP_CONCAT(DISTINCT ?awardLabel; SEPARATOR="; ") AS ?awards)
           (GROUP_CONCAT(DISTINCT ?educationInstitutionLabel; SEPARATOR="; ") AS ?educationInstitutions)
    WHERE {
        ?scientist wdt:P31 wd:Q5 ;
                  wdt:P106 wd:Q901 ;
                  wdt:P569 ?birthDate .

        OPTIONAL { ?scientist wdt:P570 ?deathDate. }
        OPTIONAL { ?scientist wdt:P19 ?birthPlace. }
        OPTIONAL { ?scientist wdt:P166 ?award. ?award rdfs:label ?awardLabel. FILTER(LANG(?awardLabel) = "en") }
        OPTIONAL { ?scientist wdt:P69 ?educationInstitution. ?educationInstitution rdfs:label ?educationInstitutionLabel }
        SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
    }
    GROUP BY ?scientist ?scientistLabel ?birthDate ?deathDate ?birthPlaceLabel
    LIMIT 10
    """
```

```
# Create a nested dictionary for each scientist
scientist_info = {
    'QID': qid,
    'Name': name_hindi,
    'BirthDate': birth_date,
    'DeathDate': death_date,
    'BirthPlace': birth_place_hindi,
    'Occupation': occupation_hindi,
    'Description': description_hindi,
    'Aliases': aliases_hindi,
    'Awards': awards_hindi,
    'EducationalInstitutions': education_institutions_hindi
}
```



Data Preprocessing & Cleaning

- Data preprocessing involves structuring the extracted data into key-value pairs for further processing. It ensures that the data is consistent, properly translated, and ready for sentence generation.
- Why Clean Data?
Raw data often contains timestamps, unknown values, or inconsistencies.
Example: Convert 1940-05-20T00:00:00Z to 1940 (Hindi: १९४०).
- STEPS:
 1. KEY-VALUE PAIR EXTRACTION:
 - Extracted fields include the scientist's name, birthdate, birthplace, occupation, awards, and academic lineage.
 2. TRANSLATION TO HINDI:
 - Automatic translation of extracted data into Hindi using the Google Translate API.



STEP 1: KEY-VALUE PAIR EXTRACTION

```
scientist_info = {  
    "Name": "Albert Einstein",  
    "BirthDate": "1879-03-14",  
    "BirthPlace": "Ulm",  
    "Occupation": "Physicist",  
    "Awards": ["Nobel Prize in Physics"],  
    "Doctoral Advisors": ["Alfred Kleiner"],  
    "Doctoral Students": ["Ernst G. Straus"]  
}
```

STEP 2: TRANSLATION TO HINDI:

```
name_hindi = translator.translate("Albert Einstein", dest="hi").text
```



USING TF-IDF TO PRIORITIZE KEY-VALUE PAIRS

What is TF-IDF?

A statistical method to measure the relevance of words in a document.

Combines:

Term Frequency (TF): Frequency of a term in a document.

Inverse Document Frequency (IDF): How unique a term is across documents.

Why TF-IDF?

Helps identify the most relevant attributes (key-value pairs) for each scientist.

Ensures important details like doctoral advisor, awards, and fields of work are included.

How it Works:

Calculated for each key-value pair:

High TF-IDF score → Highly relevant to the scientist.

Low TF-IDF score → Less relevant or common across scientists.

Example:

”Doctoral Advisor” appears rarely but has high relevance.

Template Sentence Generation

- Template sentence generation involves filling predefined sentence structures with the extracted key-value pairs to produce meaningful narratives about each scientist in both English and Hindi.

Example Template Sentence:

Template in English:

```
"{Scientist} was born on {BirthDate} in {BirthPlace} and is a famous {Occupation}."
```

Template in Hindi:

```
"{Scientist} का जन्म {BirthDate} को {BirthPlace} में हुआ था और वह एक प्रसिद्ध {Occupation} थे।"
```

Steps:

- Use the key-value pairs extracted earlier to fill in the placeholders in the sentence.
- Translate the final sentences into Hindi while respecting grammatical rules, sentence structures change based on the gender of the subject (e.g., gender-specific endings).

English Template → Filled with Key-Values → Translated Sentence in Hindi.



Feature Addition

DEFINITION:

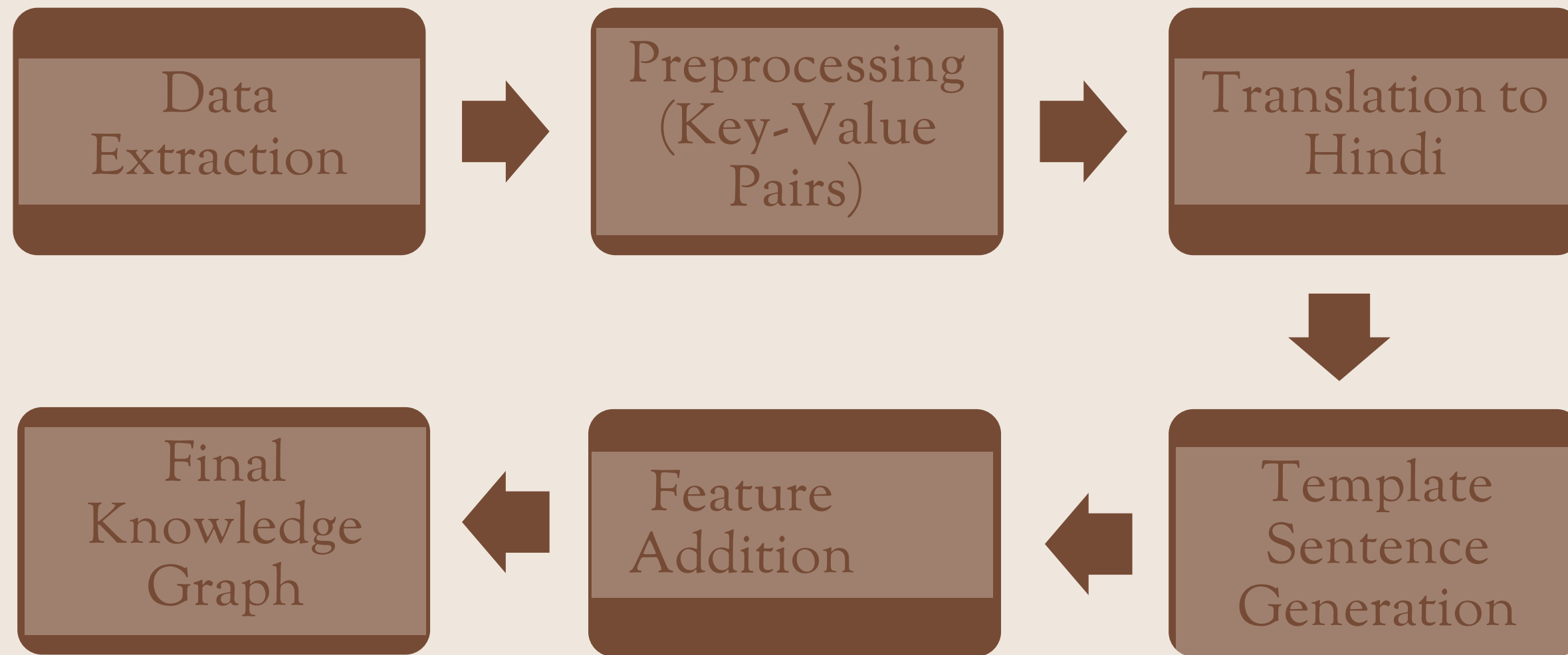
Additional features such as awards, degrees, doctoral advisors, and doctoral students were added to enrich the knowledge graph and provide more context about each scientist.

EXAMPLE OF ADDED FEATURES:

```
scientist_info = {  
    "Name": "Albert Einstein",  
    "Occupation": "Physicist",  
    "Awards": ["Nobel Prize in Physics"],  
    "Doctoral Advisors": ["Alfred Kleiner"],  
    "Doctoral Students": ["Ernst G. Straus"]  
}
```



WORKFLOW DIAGRAM



OUTPUT FILES

1 to 10 of 19 entries Filter

QID	Name	BirthDate	DeathDate	BirthPlace	Occupation	Description	Aliases	Awards	EducationalInstitutions
Q192688	इवर गियावर	१९२९	Unknown	बर्गन	भौतिक विज्ञानी	नॉर्वेजियन भौतिक विज्ञानी	['I. Giaever', 'में giaever', 'जियावर', 'Giaever i', 'Giaever I.']	गनरस मेडल;ऑनसैगर मेडल;गुगेनहाइम फैलोशिप;ओलिवर ई। बकले संघनित पदार्थ पुरस्कार;भौतिकी में नोबेल पुरस्कार	नॉर्वेजियन इंस्टीट्यूट ऑफ टेक्नोलॉजी;हमार कैथेड्रल स्कूल;क्लेयर हॉल;रेंससेलर पॉलिटेक्निक संस्थान
Q194108	मिहेल रोको	१९५०	Unknown	Unknown	वैज्ञानिक	अमेरिकी वैज्ञानिक	[]		पोलिटेनिका यूनिवर्सिटी ऑफ बुखारेस्ट
Q199640	अदीब जाटीन	१९२९	२०१४	ज़ापुरी	वैज्ञानिक	ब्राजील के विश्वविद्यालय के प्रोफेसर, वैज्ञानिक और थोरेसिक सर्जन (1929-2014)	['अदीब डोमिंगोस जाटीन']	Anísio Teixeira अवार्ड;वैज्ञानिक योग्यता के राष्ट्रीय आदेश का महान क्रॉस	यूनिवर्सिटी ऑफ साओ पाउलो
Q201803	बेंजामिन रॉबिन्स	१७०७	१७५१	नहाना	अभियंता	ब्रिटिश इंजीनियर	[]	रॉयल सोसाइटी के फेलो;कोपले मेडल	
Q203243	लियोनार्ड सुसकिंड	१९४०	Unknown	न्यूयॉर्क शहर	भौतिक विज्ञानी	अमेरिकी भौतिक विज्ञानी	['लेनी सुसकिंड', 'एल। सुस्किंड', 'सुस्किंद']	ICTP DIRAC पदक;विज्ञान लेखन पुरस्कार;पोमेरनचुक पुरस्कार;सकुराई प्राइज़	न्यूयॉर्क के सिटी कॉलेज;कॉर्नेल विश्वविद्यालय
Q203769	जीन-पर्टीन मोंटुक्ला	१७२५	१७९९	ल्यों	गणितज्ञ	फ्रांसीसी गणितज्ञ (*1725 -) 1799)	['जीन-इटीन मोंटुक्ला']		
Q205981	अलेक्जेंडर फ्रीडमैन	१८८८	१९२५	सेंट पीटर्सबर्ग	गणितज्ञ	रूसी गणितज्ञ (1888-1925)	['अलेक्जेंडर अलेक्जेंड्रोविच फ्रीडमैन', 'अलेक्जेंडर अलेक्जेंड्रोविच फ्रीडमैन', 'फ्राइडमेन', 'अलेक्जेंड्र अलेक्जेंड्रोविच फ्रीडमैन']	तलवार और धनुष के साथ सेंट व्लादिमीर 4 वीं कक्षा का आदेश;सेंट जॉर्ज 4 वीं कक्षा का क्रॉस;सेंट व्लादिमीर, 4 वीं कक्षा का आदेश;बहादुरी के लिए सोने की तलवार;ऑर्डर ऑफ सेंट जॉर्ज	दूसरा सेंट पीटर्सबर्ग जिमनैजियम;सेंट पीटर्सबर्ग विश्वविद्यालय के भौतिकी और गणित संकाय
Q206679	जोसेफ एप्स ब्राउन	१९२०	२०००	रिजफील्ड	Unknown	अमेरिकी अमेरिकी परंपराओं के अमेरिकी विद्वान	[]		
Q207325	क्रिश्चियन गोल्डबैक	१६९०	१७६४	कोनिग्सबर्ग	गणितज्ञ	जर्मन गणितज्ञ	[]		यूनिवर्सिटी ऑफ कोनिग्सबर्ग

SENTENCE GENERATION OUTPUT

1 to 10 of 16 entries

Filter

Name	TemplateSentences
पियरे पॉल डेहरिन	["पियरे पॉल डेहरिन का जन्म 1830-04-19T00:00:00Z को पेरिस का पहला अरेनडिसमेंट में हुआ था। पियरे पॉल डेहरिन एक प्रसिद्ध रसायनज्ञ थे। पियरे पॉल डेहरिन के बारे में कहा जाता है: फ्रांसीसी संयंत्र फिजियोलॉजिस्ट और कृषि रसायनज्ञ (1830-1902)। उन्हें सम्मान की सेना का अधिकारी से सम्मानित किया गया। पियरे पॉल डेहरिन ने चैप्टल हाई स्कूल;प्रशासन स्कूल;नेशनल म्युज़ियम ऑफ नेचुरल हिस्ट्री से अपनी शिक्षा प्राप्त की। पियरे पॉल डेहरिन को अन्य नामों से भी जाना जाता है, जैसे पी। पी। डेहरिन, पियरे-पॉल डेहरिन। उनका निधन 1902-12-07T00:00:00Z को हुआ।"]
नाथन मिर्वॉल्ड	["नाथन मिर्वॉल्ड का जन्म 1959-08-03T00:00:00Z को सिएटल में हुआ था। नाथन मिर्वॉल्ड के बारे में कहा जाता है: Microsoft में पूर्व CTO। उन्हें जेम्स मैडिसन मेडल से सम्मानित किया गया। नाथन मिर्वॉल्ड ने मिरमैन स्कूल;कैलिफोर्निया विश्वविद्यालय, लॉस एंजिल्स;प्रिंसटन यूनिवर्सिटी से अपनी शिक्षा प्राप्त की। नाथन मिर्वॉल्ड को अन्य नामों से भी जाना जाता है, जैसे नाथन पी मिर्वॉल्ड।"]
एडोल्फ वॉन वेरेड	["एडोल्फ वॉन वेरेड का जन्म 1807-10-14T00:00:00Z को मंस्टर में हुआ था। एडोल्फ वॉन वेरेड के बारे में कहा जाता है: जर्मन एक्सप्लोरर (1807-1863)। उन्हें से सम्मानित किया गया। एडोल्फ वॉन वेरेड ने से अपनी शिक्षा प्राप्त की। उनका निधन 1863-03-15T00:00:00Z को हुआ।"]
जेम्स मोरियारन टान्नर	["जेम्स मोरियारन टान्नर का जन्म 1920-08-01T00:00:00Z को केंबर्ली में हुआ था। जेम्स मोरियारन टान्नर के बारे में कहा जाता है: ब्रिटिश बाल रोग विशेषज्ञ (1920–2010)। उन्हें से सम्मानित किया गया। जेम्स मोरियारन टान्नर ने पेंसिल्वेनिया विश्वविद्यालय में पेरेलमैन स्कूल ऑफ मेडिसिन;मार्लबोरो कॉलेज;यूनिवर्सिटी ऑफ एक्सेटर से अपनी शिक्षा प्राप्त की। उनका निधन 2010-08-11T00:00:00Z को हुआ।"]
पॉल-एमिल लेकोक ऑफ बोइसबॉइन	["पॉल-एमिल लेकोक ऑफ बोइसबॉइन का जन्म 1838-04-18T00:00:00Z को कॉर्नेक में हुआ था। पॉल-एमिल लेकोक ऑफ बोइसबॉइन एक प्रसिद्ध रसायनज्ञ थे। पॉल-एमिल लेकोक ऑफ बोइसबॉइन के बारे में कहा जाता है: फ्रांसीसी केमिस्ट (1838-1912)। उन्हें शूरवीर ऑफ द लीजन ऑफ ऑनर;डेवी मेडल से सम्मानित किया गया। पॉल-एमिल लेकोक ऑफ बोइसबॉइन ने पॉलिटेक्निक स्कूल से अपनी शिक्षा प्राप्त की। पॉल-एमिल लेकोक ऑफ बोइसबॉइन को अन्य नामों से भी जाना जाता है, जैसे फ्रांस्वा लेकोक डे बोइसबॉइन। उनका निधन 1912-05-28T00:00:00Z को हुआ।"]
जोसेफ मारेक	["जोसेफ मारेक का जन्म 1868-03-18T00:00:00Z को हॉर्न स्ट्रैंडा में हुआ था। जोसेफ मारेक एक प्रसिद्ध वैज्ञानिक थे। जोसेफ मारेक के बारे में कहा जाता है: हंगेरियन साइंटिस्ट (1868-1952)। उन्हें कोसुथ प्राइज़ से सम्मानित किया गया। जोसेफ मारेक ने वेटरनरी मेडिसिन बडापेस्ट विश्वविद्यालय;यूनिवर्सिटी ऑफ बर्न से अपनी शिक्षा प्राप्त की। जोसेफ मारेक को अन्य नामों से भी जाना जाता है, जैसे जोजसेफ मारेक। उनका निधन 1952-09-07T00:00:00Z को हुआ।"]
सर्गेई झुक	["सर्गेई झुक का जन्म 1892-04-04T00:00:00Z को कीव में हुआ था। सर्गेई झुक एक प्रसिद्ध अभियंता थे। सर्गेई झुक के बारे में कहा जाता है: सोवियत इंजीनियर (1892-1957)। उन्हें आंतरिक मामलों के मंत्रालय के सम्मानित कार्यकर्ता;पदक "बैटल मेरिट के लिए";पदक "महान देशभक्ति युद्ध में बहादुर श्रम के लिए 1941-1945";पदक "लेबर वेलोर के लिए";स्टालिन पुरस्कार;लाल बैनर का आदेश;समाजवादी श्रम का नायक;रेड स्टार का आदेश;लेनिन का आदेश;श्रम के लाल बैनर का आदेश से सम्मानित किया गया। सर्गेई झुक ने सेंट पीटर्सबर्ग स्टेट ट्रांसपोर्ट यूनिवर्सिटी;सम्राट अलेक्जेंडर रेलवे इंजीनियर्स संस्थान से अपनी शिक्षा प्राप्त की। सर्गेई झुक को अन्य नामों से भी जाना जाता है, जैसे सर्गेई याकोवलेविच झुक, सर्गेई याकोवलेविच जुक। उनका निधन 1957-03-01T00:00:00Z को हुआ।"]
दुर्ग	["दुर्ग का जन्म 1817-09-28T00:00:00Z को Iste pee में हुआ था। दुर्ग के बारे में कहा जाता है: रूसी सिनोलॉजिस्ट (1817-1878)। उन्हें सेंट अन्ना का आदेश, तीसरा वर्ग;सेंट व्लादिमीर का आदेश, तीसरा वर्ग;ऑर्डर ऑफ सेंट अन्ना, प्रथम श्रेणी से सम्मानित किया गया। दुर्ग ने सेंट पीटर्सबर्ग थियोलॉजिकल एकेडमी से अपनी शिक्षा प्राप्त की। दुर्ग को अन्य नामों से भी जाना जाता है, जैसे पीटर इवानोविच काफ़रोव, पल्लादियस (काफ़रोव)। उनका निधन 1878-12-18T00:00:00Z को हुआ।"]
वाल्टर कनिंघम	["वाल्टर कनिंघम का जन्म 1932-03-16T00:00:00Z को क्रेस्टन में हुआ था। वाल्टर कनिंघम के बारे में कहा जाता है: अमेरिकी अंतरिक्ष यात्री (1932–2023)। उन्हें अंतर्राष्ट्रीय अंतरिक्ष हॉल ऑफ फेम;नासा प्रतिष्ठित सेवा पदक;यूनाइटेड स्टेट्स एस्ट्रोनॉट हॉल ऑफ फेम से सम्मानित किया गया। वाल्टर कनिंघम ने वेनिस हाई स्कूल;सांता मोनिका कॉलेज;कैलिफोर्निया विश्वविद्यालय, लॉस एंजिल्स;हार्वर्ड बिज़नेस स्कूल से अपनी शिक्षा प्राप्त की। वाल्टर कनिंघम को अन्य नामों से भी जाना जाता है, जैसे रॉनी वाल्टर "वॉल्ट" कनिंघम, वॉल्ट कनिंघम। उनका निधन 2023-01-03T00:00:00Z को हुआ।"]
प्रति टेओडोर क्लेव	["प्रति टेओडोर क्लेव का जन्म 1840-02-10T00:00:00Z को द ग्रेट चर्च असेंबली में हुआ था। प्रति टेओडोर क्लेव एक प्रसिद्ध रसायनज्ञ थे। प्रति टेओडोर क्लेव के बारे में कहा जाता है: स्वीडिश केमिस्ट जिन्होंने होल्मियम और थुलियम की खोज की (1840-1905)। उन्हें डेवी मेडल से सम्मानित किया गया। प्रति टेओडोर क्लेव ने उप्साला यूनिवर्सिटी से अपनी शिक्षा प्राप्त की। प्रति टेओडोर क्लेव को अन्य नामों से भी जाना जाता है, जैसे प्रति टेओडोर क्लेव, प्रति क्लेव, क्लीव, प्रति टी। क्लेव। उनका निधन 1905-06-18T00:00:00Z को हुआ।"]



Challenges and Solutions

- DATA EXTRACTION:
 - Handling incomplete or missing data from Wikidata.
 - Solution: Implement optional queries and fallback defaults (e.g., "Unknown")
- TRANSLATION ACCURACY:
 - The challenge of maintaining context when translating scientific terms.
 - Solution: Post-processing and validation with manual checks for critical translations



Github link for Code and PPT

<https://github.com/shaluKm/Automatic-Hindi-Wikipedia-Generation>



References

- <https://aclanthology.org/2023.ranlp-1.2.pdf>
- https://www.researchgate.net/publication/340694153_Neural_Machine_Translation_English_to_Hindi
- <https://jena.apache.org/tutorials/sparql.html>
- <https://en.wikipedia.org/wiki/Wikipedia:Wikidata>
- https://www.wikidata.org/wiki/Help:Linking_Wikipedia_pages



THANK YOU!