Taxonomies, Ontologies and Knowledge Graphs

Vasudeva Varma

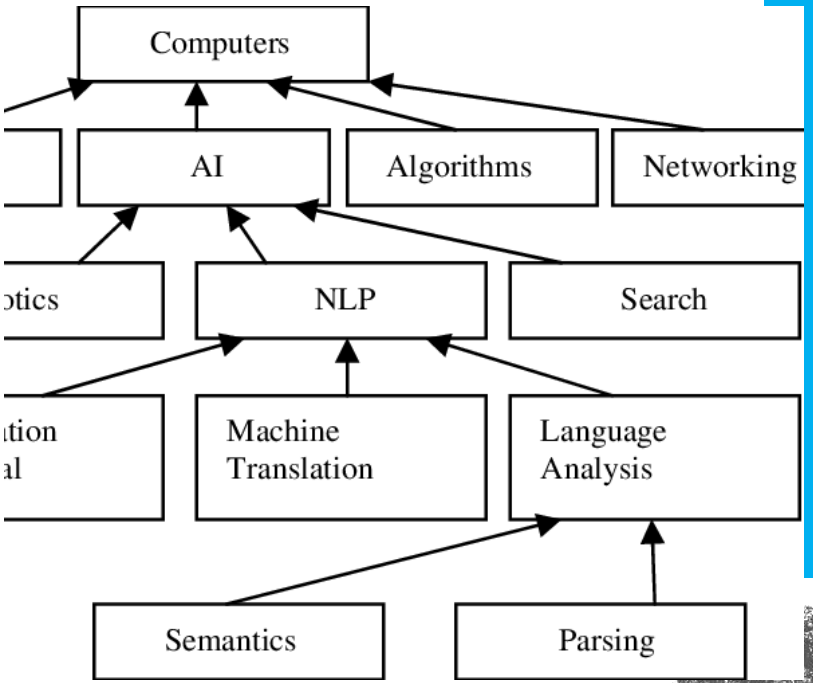
External Knowledge

- Integrating external knowledge augments NLP/IR technologies
- Key questions for our discussion in this module:
  - What is knowledge? Vs. fact vs. truth?
  - How to represent the knowledge?
    - Taxonomy, ontology, … Knowledge Graphs
    - Knowledge representation choices
  - How to extract the knowledge?
    - Text to knowledge and knowledge to text
    - Systems like NELL, TextRunner
- Key Semantic databases/knowledge Bases
  - Wikidata

Taxonomy

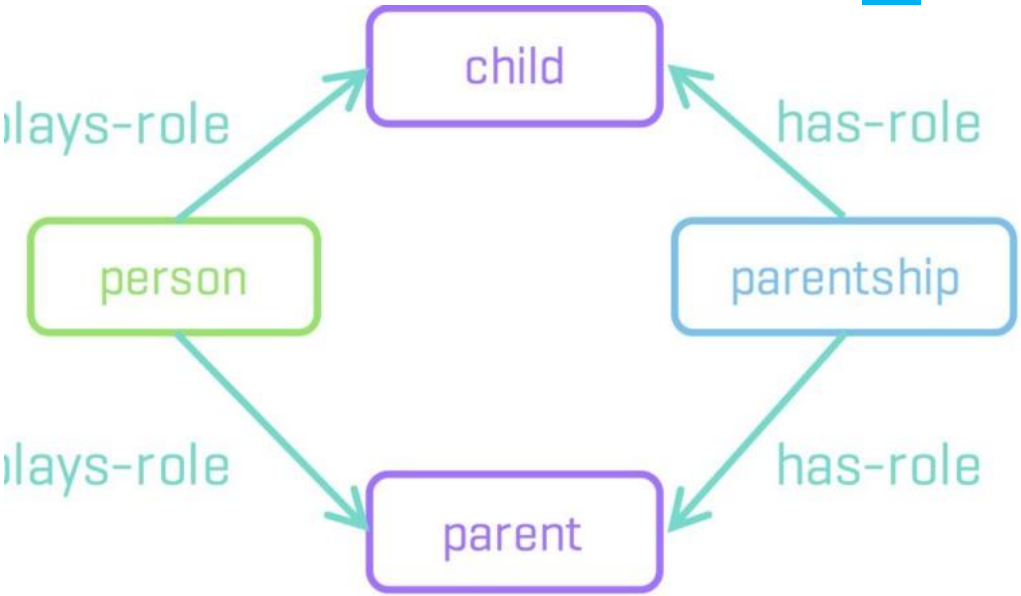
More than an agreed vocabulary

The science of categorization, or classification, of things based on a predetermined system.

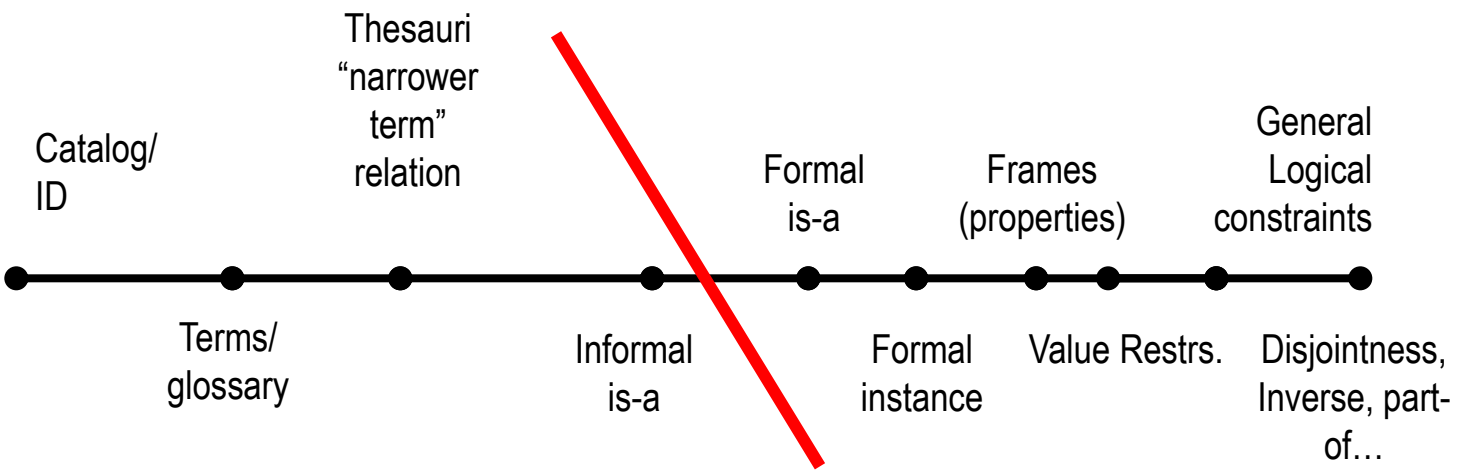


Ontology

- More than a taxonomy or classification of terms
- What does Ontology do?
  - Captures knowledge
  - Creates a shared understanding – between humans and for computers
  - Makes knowledge machine processable
  - Makes meaning explicit – by definition and context



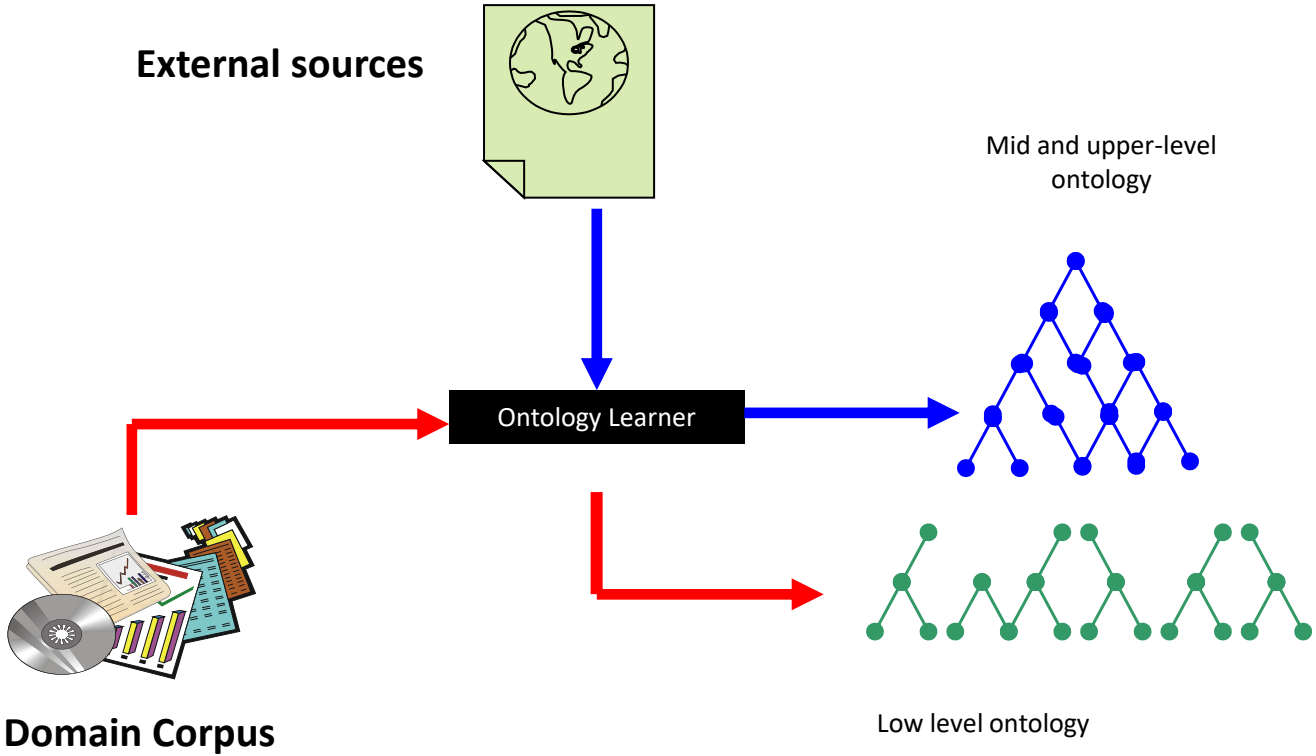
What is an Ontology?



Components of an Ontology

- Concepts: Class of individuals
- Relationships between concepts
  - *Is a kind of* relationship forms a taxonomy
  - Other relationships give further structure – *is a part of*
- *Axioms – Disjoint-ness, covering, equivalence,…*

Using external sources



Ontologies -> Knowledge graph
Where Ontologies end, KG begins

- Size of nature of data debate

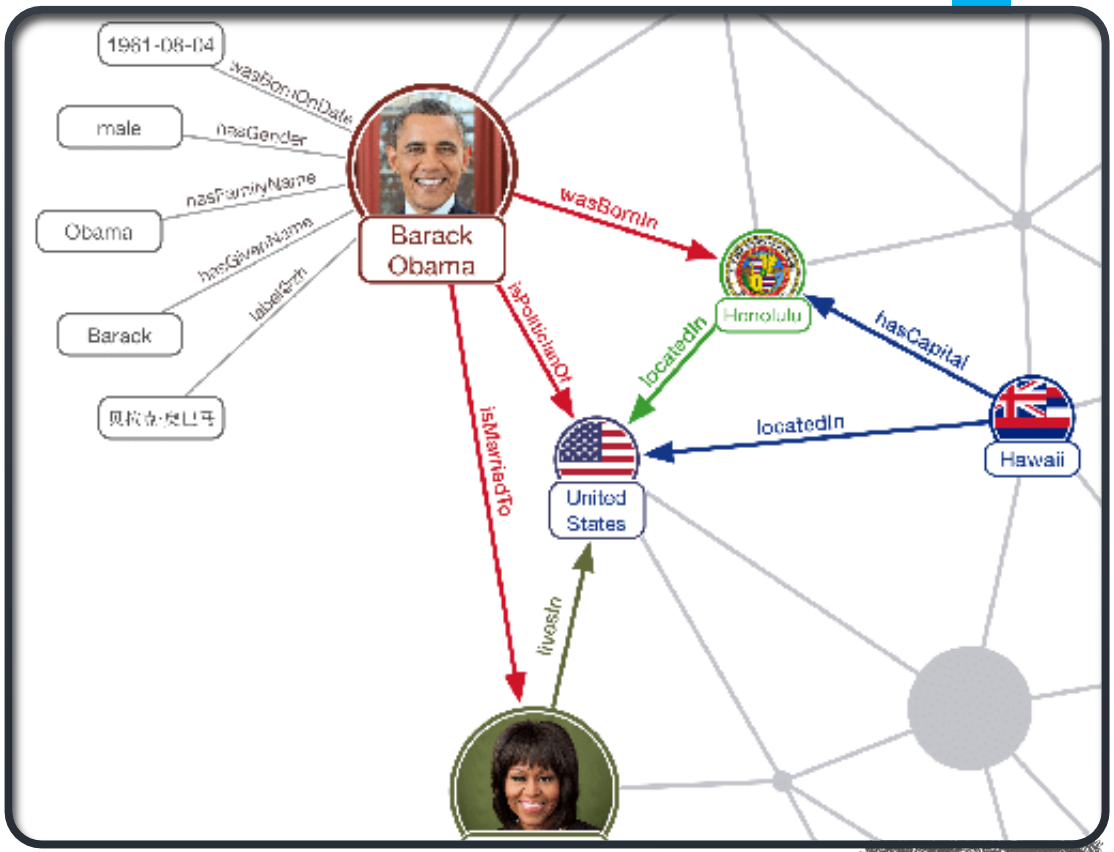
- Ontologies are generally regarded as
  - Smaller collections of assertions that are hand-curated,
  - Usually for solving a domain-specific problem.

- By comparison, knowledge graphs can include literally billions of assertions, just as often domain-specific as they are cross-domain.

What is a Knowledge Graph?

Knowledge in graphs form!

- Nodes (entities)
- Labels (Attributes)
- Relationships (Typed edges)



Focus on instance aspect of knowledge, not the schema aspect.

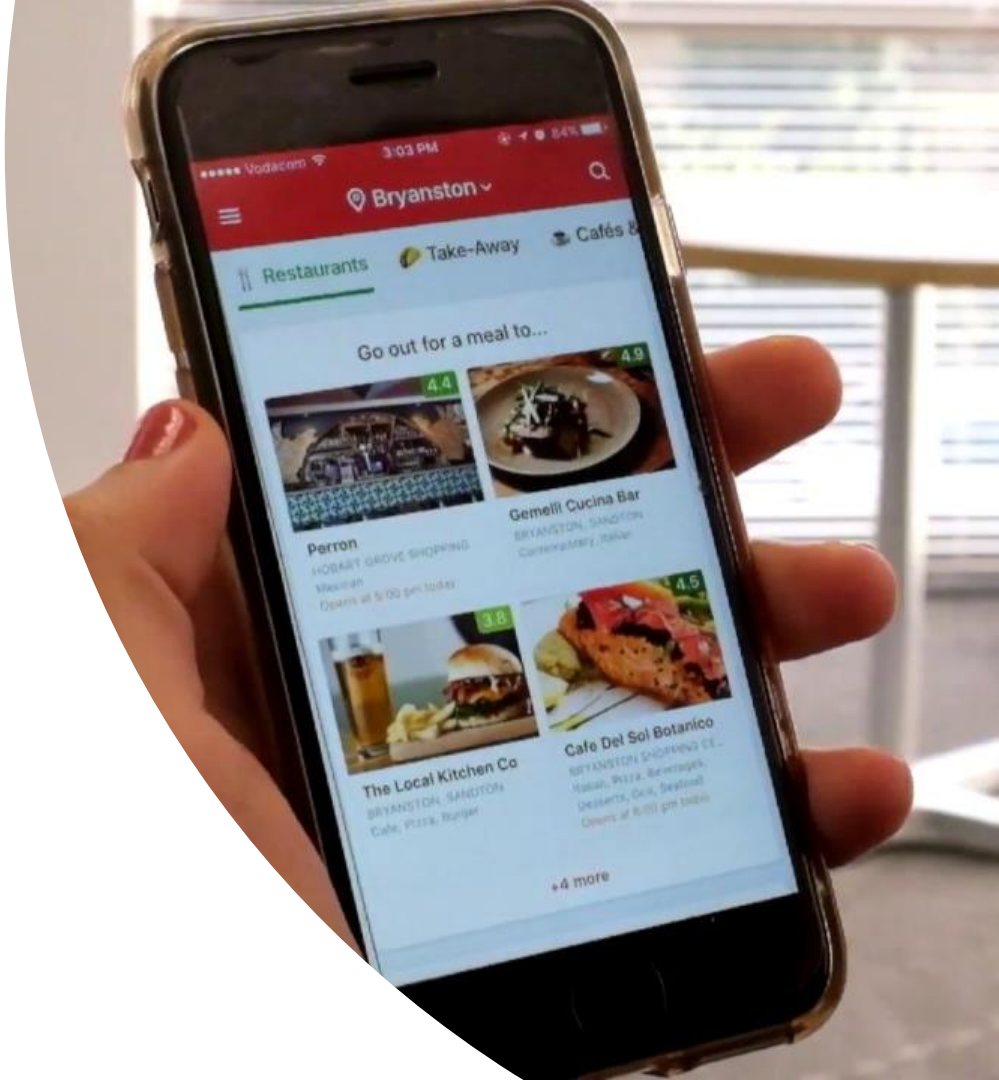
Applications of KGs

Turning Complex Information into Helpful Answers

- Useful for Humans
  - Addressing information overload
  - Helps in knowledge driven tasks
  - Navigation through “knowledge structures”
- Useful for AI systems
  - From Data to human semantics
  - Builds on Graphs Theory – hence inherits all the mathematical infrastructure
  - Fundamental building block for many AI tasks

Knowledge Graph Powers Search Engines and beyond

- Richer Data for Entity Pane, weather pane, Sports pane, Carousel, and Facts Across Segments
- Knowledge graph serves NL fact answers
- knowledge in answers (Query: Area size of India)
- Knowledge powered Q&A
- Knowledge-powered Conversation



State of Art KGs

- Popular KGs
  - Cyc and Open Cyc
  - Freebase
  - DBpedia
  - YAGO
  - NELL
  - ConceptNet
  - OpenIE
  - Wikidata
- Large production KGs
  - Google Knowledge Vault - Google KG
  - Microsoft Satori KG
  - Yandex Object Answer
  - IBM Watson
- Vertical KGs
  - Facebook (social network) Graph API
  - LinkedIn (people graph)
  - Amazon (product graph)
- Misc
  - Diffbot, GraphIQ, Maana, ParseHub, Reactor Labs, SpazioDati

Schema.org
Datacommons.org

https://www.youtube.com/watch?v=Oips1aW738Q
https://figshare.com/articles/Documenting_and_preserving_programming_languages_and_software_in_Wikidata/7388297
https://link.springer.com/chapter/10.1007%2F978-3-319-70863-8_16
https://link.springer.com/chapter/10.1007%2F978-3-319-98932-7_12

Wikidata

- Structured data
- One common database
  - For 280 editions of Wikipedia and countless websites outside Wikipedia
  - Empowers Infoboxes (e.g. Population of a country)
  - Enables Wikipedia lists
- Multilingual
  - Enables interwiki links
- Machine readable
- Collaboratively edited by the community
  - Community decides whether data is useful or not

Wikidata

Modeling Wikidata - Statements

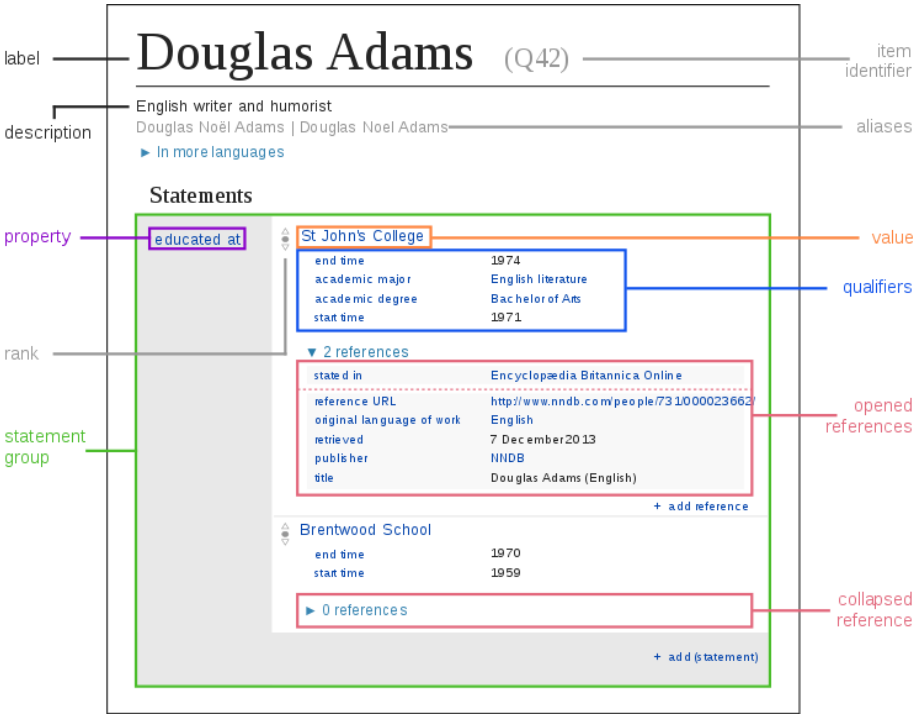
- Mount Everest is the highest mountain in the world

  Earth (Q2) (item) → highest point (P610) (property) → Mount Everest (Q513) (value)

- Wiki Data also holds a statement about the item Mount Everest (indicating it is a mountain):

  Mount Everest (Q513) (item) → instance of (P31) (property) → mountain (Q8502) (value)

- A statement is composed of an item and a property-value pair
- An item can be viewed as the subject part of a triplet
- The property represents a triplet's predicate;
- A value is used to express the object of a triplet.


label
Douglas Adams
(Q42)
item identifier
description
English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
aliases
► In more languages
Statements
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
+ add reference
Brentwood School
end time 1970
start time 1959
► 0 references
collapsed reference
+ add (statement)

Mahatma Gandhi

From Wikipedia, the free encyclopedia

"Gandhi" redirects here. For the third prime minister of India, see Indira Gandhi. For other uses, see Gandhi (disambiguation).

Mohandas Karamchand Gandhi (/ˈgɑːndi, ˈgændi/;[2] 2 October 1869 – 30 January 1948) was an Indian lawyer,[3] anti-colonial nationalist,[4] and political ethicist,[5] who employed nonviolent resistance to lead the successful campaign for India's independence from British Rule,[6] and in turn inspire movements for civil rights and freedom across the world. The honorific **Mahātmā** (Sanskrit: "great-souled", "venerable"),[7] first applied to him in 1914 in South Africa,[8] is now used throughout the world.

Born and raised in a Hindu family in coastal Gujarat, western India, Gandhi was trained in law at the Inner Temple, London, and called to the bar at age 22 in June 1891. After two uncertain years in India, where he was unable to start a successful law practice, he moved to South Africa in 1893 to represent an Indian merchant in a lawsuit. He went on to stay for 21 years. It was in South Africa that Gandhi raised a family, and first employed nonviolent resistance in a campaign for civil rights. In 1915, aged 45, he returned to India. He set about organising peasants, farmers, and urban labourers to protest against excessive land-tax and discrimination. Assuming leadership of the Indian National Congress in 1921, Gandhi led nationwide campaigns for easing poverty, expanding women's rights, building religious and ethnic amity, ending untouchability, and above all for achieving *Swaraj* or self-rule.[9]

The same year Gandhi adopted the Indian loincloth, or short *dhoti* and, in the winter, a shawl, both woven with yarn hand-spun on a traditional Indian spinning wheel, or *charkha*, as a mark of identification with India's rural poor. Thereafter, he lived modestly in a self-sufficient residential community, ate simple vegetarian food, and undertook long fasts as a means of self-purification and political protest. Bringing anti-colonial nationalism to the common Indians, Gandhi led them in challenging the British-imposed salt tax with the 400 km (250 mi) Dandi Salt March in 1930, and later in calling for the British to Quit India in 1942. He was imprisoned for many years, upon many occasions, in both South Africa and India

Mahātmā
Mohandas Karamchand Gandhi

Born	Mohandas Karamchand Gandhi 2 October 1869 Porbandar, Kathiawar Agency, British-ruled India
Died	30 January 1948 (aged 78)

Main page
Community portal
Project chat
Create a new Item
Create a new Lexeme
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Mohandas Karamchand Gandhi (Q1001)

pre-eminent leader of Indian nationalism during the British Raj

Mahatma Gandhi | Mahatma Mohandas Karamchand Gandhi | M. K. Gandhi | Mohandas K Gandhi | M K Gandhi | Mohandas Gandhi | Bapu | Gandhi | Mohandas K. Gandhi | Gandhiji | Gandhji | Bapuji

In more languages

Configure

Language	Label	Description	Also known as
English	Mohandas Karamchand Gandhi	pre-eminent leader of Indian nationalism during the British Raj	Mahatma Gandhi Mahatma Mohandas Karamcha… M. K. Gandhi Mohandas K Gandhi M K Gandhi Mohandas Gandhi Bapu Gandhi Mohandas K. Gandhi Gandhiji Gandhji Bapuji
Telugu	మహాత్మా గాంధీ	భారతదేశ జాతిపిత, స్వాతంత్ర్య సమర యోధులు.	మోహన్దాస్ కరంచంద్ గాంధీ మోహన్దాస్ గాంధీ బాపు గాంధీ గాంధీజి
Hindi	महात्मा गांधी	भारतीय स्वतंत्रता आंदोलन के एक प्रमुख राजनैतिक एवं	मोहनदास करमचंद गांधी

sex or gender	male 6 references
country of citizenship	British Raj start time end time 1 reference Dominion of India start time end time 1 reference
name in native language	મોહનદાસ ગાંધી (Gujarati) 0 references
birth name	Mohandas Karamchand Gandh 1 reference

date of death	30 January 1948 9 references
place of death	Gandhi Smriti located in the administrative territorial entity: Delhi country: Dominion of India 2 references New Delhi 1 reference
manner of death	homicide statement is subject of: assassination of Mahatma Ga 1 reference
cause of death	ballistic trauma quantity: 3 has immediate cause: assault 2 references

Assignment … Last Date to Submit: Nov 2nd

- How to create Wikipedia pages from Wikidata?

- … Study Wikidata
- … Study SPARQL for Wikidata
- … Study Wikidata of your mother tongue

- … Create one Wikipedia page automatically in your mother tongue
- … Submit a report + submit the Wiki page.

IndicWiki Project: Human Assisted Machine Generated Wikipedia in Indian languages

For each domain:

- Choosing seed datasets
- Expanding datasets
- Attributed Identification
- Attribute Classification
- Translation/transliteration of Metadata (attributes) and data (values)
- Template creation
- Text Generation

Data/Knowledge to Text

Data Collection and pre-processing step

Problem: Create an end-to-end data collection and preprocessing pipeline for a specific domain

Input

Name of the domain

A seed set of structured sources OR

A seed set of unstructured sources

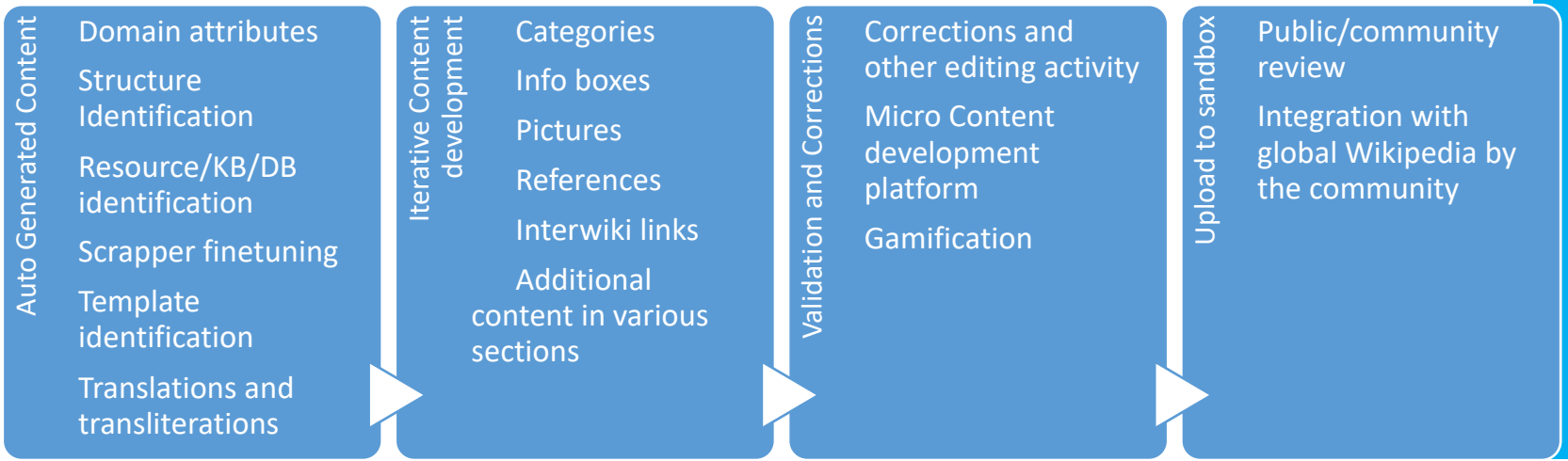
Output

A structured data source of collected/refined data (as a JSON file)

Process

- Understand the domain, seed sources: Domain List
- Add more structured/unstructured sources from open domain/Internal sources
- Download/crawl/collect data from all the sources
- Convert data from original sources (Webpages, pdf files, CSV files, …) to structured data fields
- Data cleaning/pre-processing as needed
- Strategy to enhance the data with crowd sourcing methods

Content development - Human-Bot Collaboration Flow

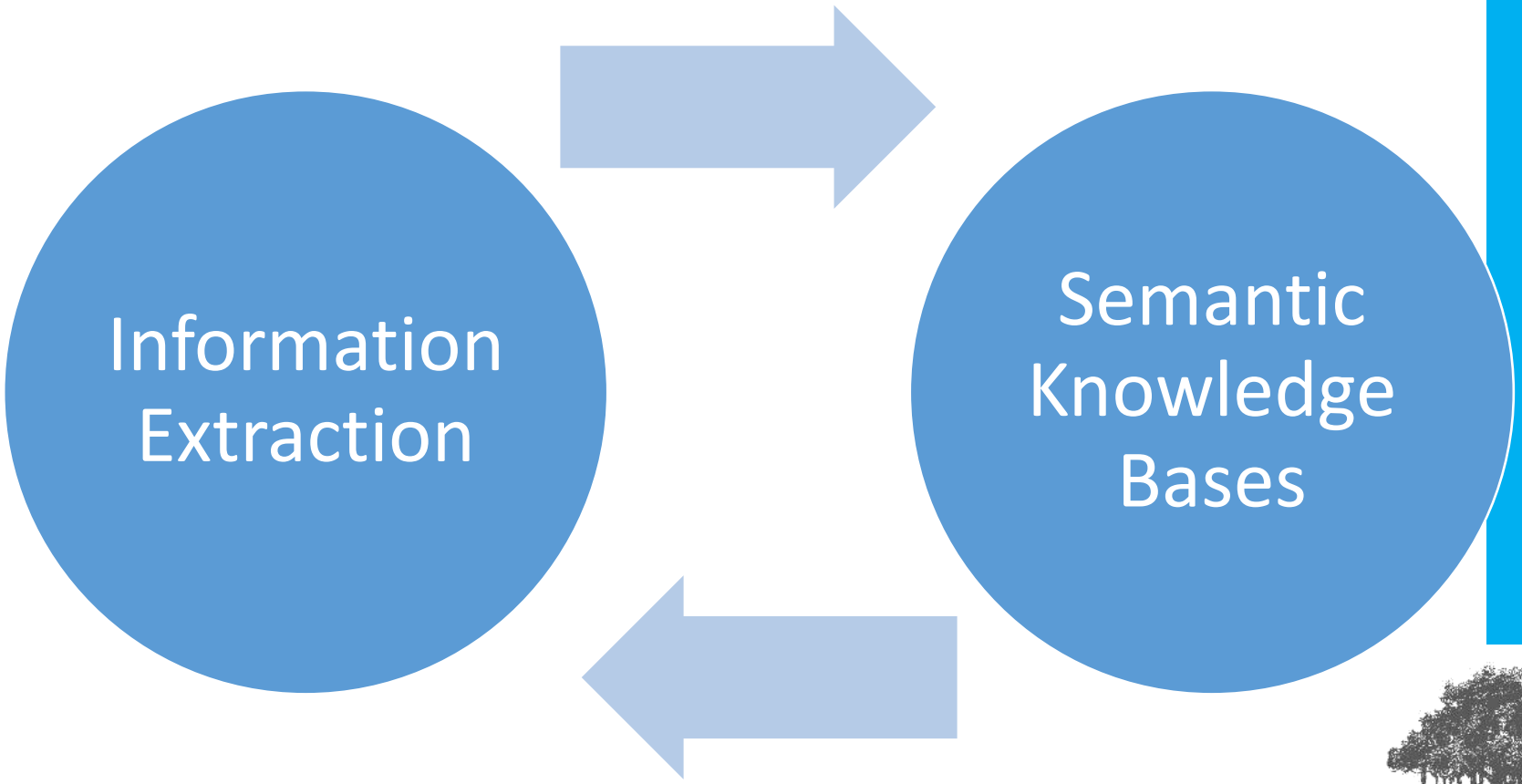


New things in Wikidata

- Wiki Functions (Wikilambda) https://arxiv.org/abs/2004.04733
- Abstract Wikipedia https://research.google/pubs/pub48057/

Challenges of scaled KGs

- Building a small KG is easy - building a vast system is a huge challenge
- Conflicting Goals of KGs
  - Coverage : Have we got the information we need?
  - Freshness: Is information up to date?
  - Correctness Is our information accurate?
- Example: Will Smith: Single entity, 108K facts assembled from 41 web sites. There are 200 Will Smiths on Wikipedia alone.


Information Extraction
Semantic Knowledge Bases

Where do KGs come from?

- Structured text: Databases, tables, social nets, Wikipedia infoboxes, …
- Unstructured text: WWW, social media, news, reference articles
- Images
- Videos: YouTube, video feeds

Process of building KG

IE for KB Generation Process

- Input: Text/Embeddings
- Phase I: Identify entity mentions
- Phase II Identify NEs
- Phase III: Identify attributes of the NEs
- Phase IV: Identify relations (N-ary)
- Output: Knowledge Graph

Two Perspectives: Knowledge construction and Graph construction

Who are the entities (nodes) in the graph?

- Named Entity Recognition
- Entity Coreference

What are their attributes and types (labels)?

- Named Entity Recognition

How are they related (edges)?

- Relation Extraction
- Semantic Role Labeling

Who are the entities (nodes) in the graph?

- Entity Linking
- Entity Resolution

What are their attributes and types (labels)?

Collective classification

How are they related (edges)?

- Link Prediction

Knowledge extraction (IE Based)

- Built on the foundation of NLP techniques
  - Part-of-speech tagging, dependency parsing, named entity recognition, coreference resolution…
  - Challenging problems with very useful outputs
- Information extraction techniques use NLP to:
  - define the domain
  - extract entities and relations
  - score candidate outputs
- Trade-off between manual & automatic methods

Entity linking and KBP task

Loving County is a county in the U.S. state of Texas, and is the least populous county in the entire United States.[1][2] Its seat, and only community, is Mentone. The nearest sizable towns are Pecos, Texas, and Carlsbad, New Mexico. In 2000, its population was 67.

Part of the Haley Ranch, founded by the father of Texas historian J. Evetts Haley, is in Loving County, with another portion in neighboring Winkler County.

Contents [hide]

1 History
1.1 Exploration and incorporation
1.2 Politics
1.3 Takeover attempt by the "Free Town Project"
2 Geography
2.1 Highways
2.2 Adjacent counties
3 Demographics
4 Economy
5 Education
6 Popular culture
7 References
8 External links

History [edit]

Exploration and incorporation [edit]

Prehistorically, the area had many springs with drinkable water that supported wildlife and nomadic hunters. Antonio de Espejo visited the area in 1583 and crossed the Pecos River. Having surveyed the area in 1854 for a railroad company, John Pope returned in 1855 to start a camp in northwestern Loving County and establish artesian wells in the area, but the venture was unsuccessful and was abandoned in 1861.

From 1837 to 1874 the area of modern Loving County was part of the Bexar land district. In 1874 it was separated from Bexar County, becoming a part of Tom Green County.

Loving County is named for Oliver Loving, a cattle rancher and pioneer of the cattle drive who together with Charles Goodnight developed the Goodnight-Loving Trail. He was mortally wounded by Comanches while on a cattle drive in 1867 in the vicinity of the county.

Loving is the only county in Texas to be incorporated twice, first in 1893 and then once more in 1931. Its initial organization was effected by a canal company founded in Denver, Colorado, and appears to have been based upon fraud and willful misrepresentations made by the founders

Loving County, Texas

Map

Location in the state of Texas

Texas's location in the U.S.

Statistics

Founded	1931
Seat	Mentone
Area	
- Total	677 sq mi (1,753 km²)
- Land	673 sq mi (1,743 km²)
- Water	4 sq mi (10 km²), 0.56%
Population	
- (2000)	67
- Density	0.0996/sq mi (0.0384/km²)

Comments 0 | Recommend 0

Print
RSS
MY YAHOO!
Yahoo! Buzz
Share

Loving County, population 55, is Texas' richest

Collin comes in third, and Dallas is fifth wealthiest

03:37 PM CDT on Friday, April 25, 2008

By BRENDAN M. CASE / The Dallas Morning News
bcase@dallasnews.com

There's nothing like an oil boom to boost incomes in West Texas.

Tiny Loving County, population 55, is the state's richest county, according to an income study released Thursday by the U.S. Commerce Department.

Based on 2006 data, Loving had a per capita income of $83,569. That may be a statistical curiosity, given the county's teensy population (58 in 2006) and its perch atop the oil-rich Permian Basin. But another West Texas county – Midland – was second, rising from fifth in 2005. It had an average income of $48,644.

"There's so much oil and gas wealth out there right now," Waco economist Ray Perryman said, referring to West Texas. "The joke out there is that the oil business is so good, they can afford to be in the cattle business again."

Collin County came in third, at

CLICK IMAGE TO ENLARGE

Where the moneyed live

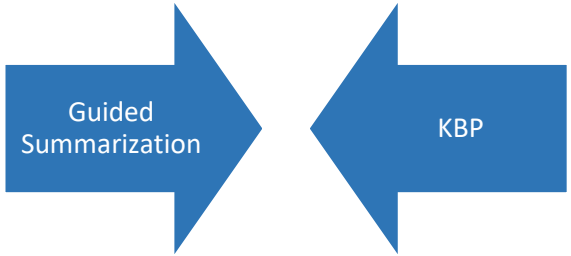
A map of 2006 per capita income by county shows the traditional clusters of wealth around the state's major cities. But

- Inconsistency
- Incompleteness
- Accuracy of facts
- Novel information
- Cost of Manual efforts

Solution: **Automatically updating information of the entities in knowledge bases**

Knowledge Base Population

- Knowledge Base Population can be fundamentally broken down into two sub problems
- *Entity Linking* : Linking entity mentions in documents to Knowledge Base nodes
- *Slot Filling* : Extracting attribute information for query entities



Summarization and KBP are complementary tasks

Summaries help in filling the slot values more effectively

Slot values enhance the quality of guided summaries

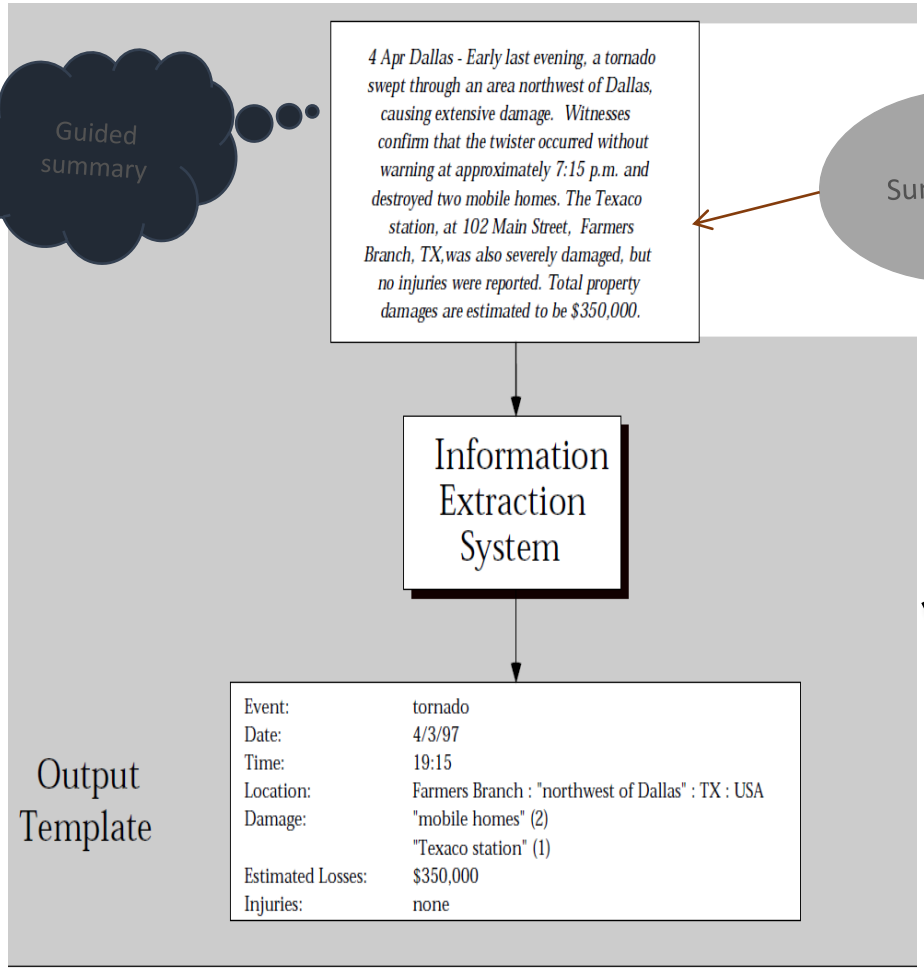


Figure 1. An Information-Extraction System in the Domain of Natural Disasters.

Possibilities… Turning the web into a KB

Is it possible to…

- Know what Wikipedia knows?
- Know everything that is machine readable?
- Know collection of all entities, classes, relations, and facts?

Thank You!

Vasudeva Varma

vv@iiit.ac.in

Twitter: **devvarma**

Web: www.iiit.ac.in/~vv