Welcome to:

Applications of Natural Language Processing

Unit objectives

After completing this unit, you should be able to:

- Understand what is information retrieval and the concepts
- Learn about work with the steps in IR and perform IR
- Gain knowledge on information answering, the various types of QA, how to model a QA
- Understand the concepts of information extraction, basic ideas and operations in IE
- Learn about what is ontology construction, the types, categories and steps involved in OC

Information retrieval

- Storage and access to the information.
- “Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”.
- Information retrieval system ➔ Software to store and retrieve information.

Figure: Information Retrieval

Source: https://itexperttraining.com/core/courses/information-retrieval/

Information retrieval in NLP

- User inputs a Query.
- Process ➔ Natural language.
- Identify the relevant information.
- Process the query.

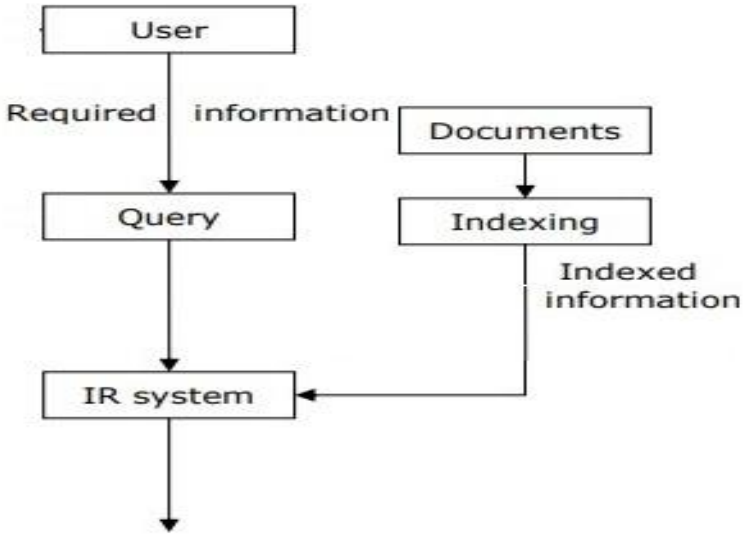

Figure: IR system

IR development (1 of 2)

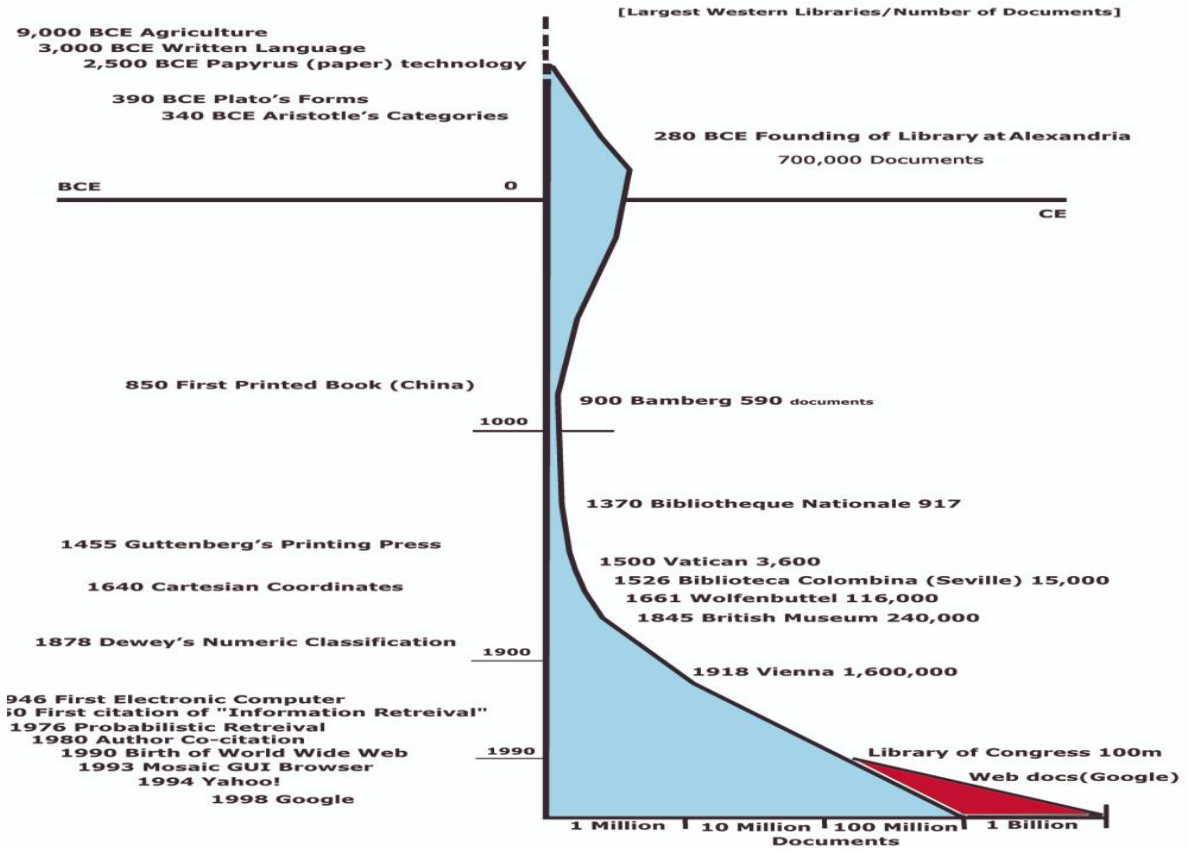

Figure: Availability of Data over timeline

Source: https://www.researchgate.net/figure/Rough-timeline-of-the-generations-of-information-retrieval-in-digital-libraries-The_fig1_14214241

IR development (2 of 2)

- 1920 to 1930 ➔ First document storage and retrieval system.
- 1950 ➔ Searching for information through selective process.
- 1970 ➔ Large-scale retrieval system.
- 1990 ➔ Internet and World Wide Web ➔ Information retrieval process.
- Search engines ➔ Retrieval systems ➔ Data within seconds.
- 21st century ➔ Large amount of data.

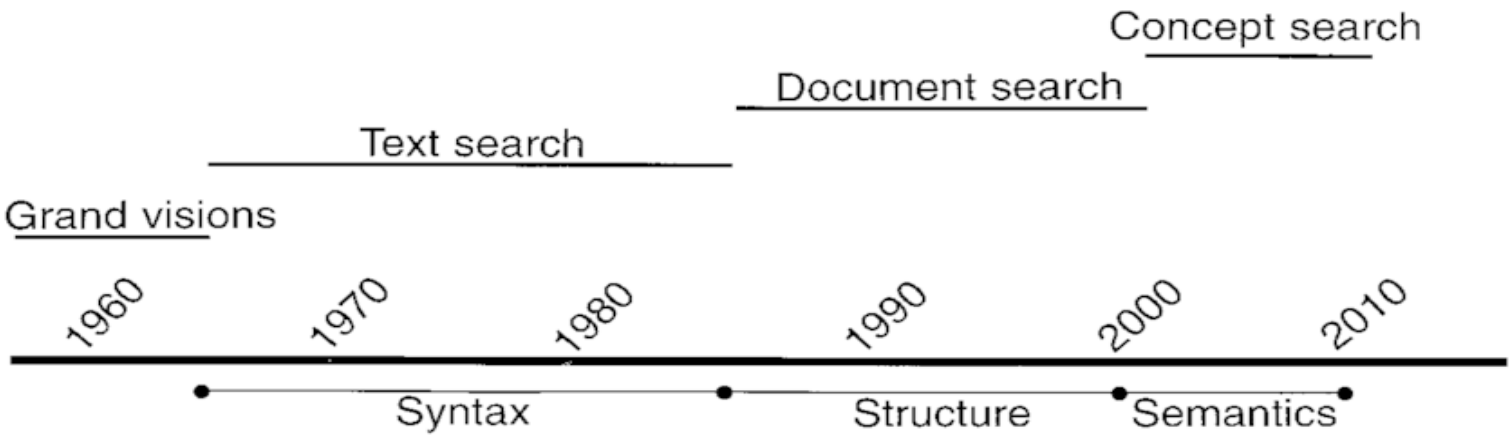

Figure: IR Timeline

Source: http://online.sfsu.edu/fielden/hist.htm

Model types

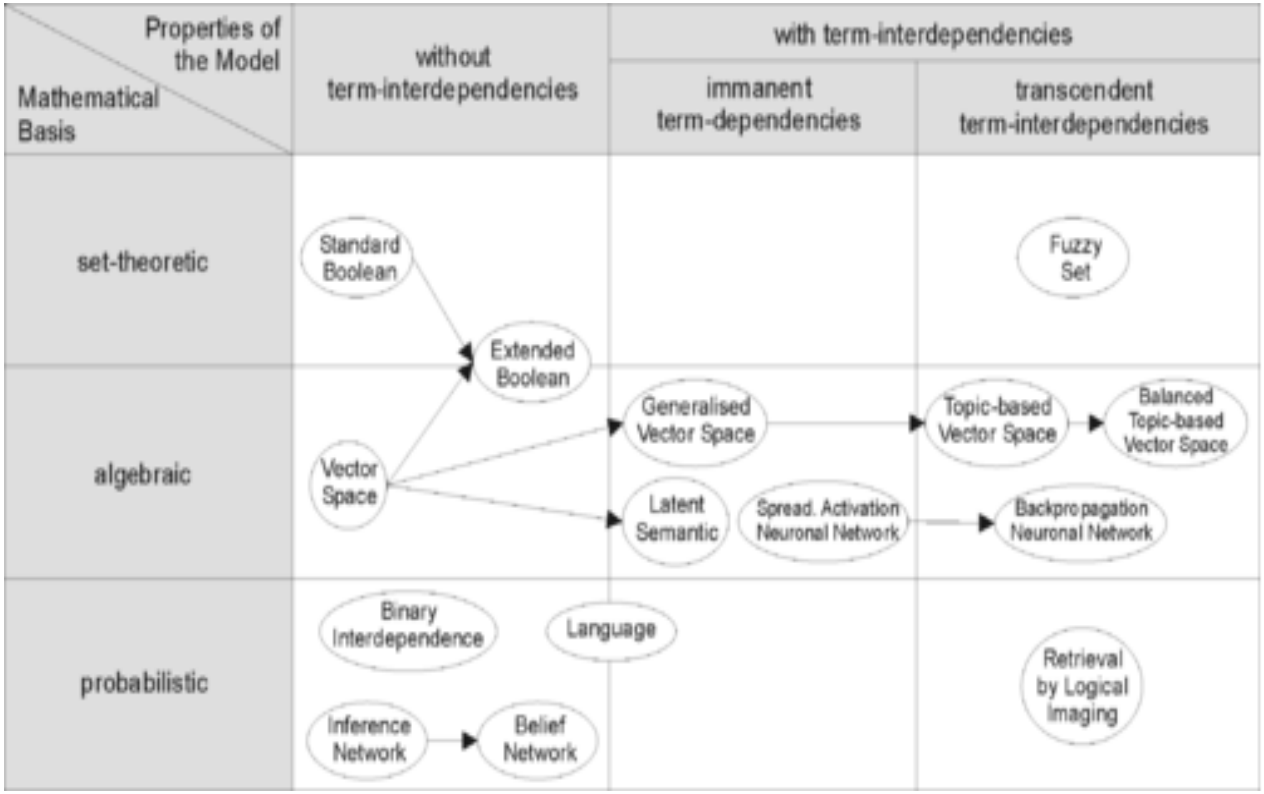

Figure: Models

Source: https://en.wikipedia.org/wiki/Information_retrieval#Model_types

Model types: Mathematical basis model

- Set-theoretic models:
  - Words ➔ Sets.

- Standard Boolean.
- Extended Boolean.
- Fuzzy retrieval.

- Algebraic models:
  - Documents ➔ Matrix, Vector or Tuples.
  - Similarity values ➔ Match document to query.

- Vector space.
- Generalized vector space.
- Topic-based vector space.
- Extended Boolean.
- Latent semantic indexing.

Problems with NLP in information retrieval

- Linguistic variation:
  - Different words ➔ Same concept.
  - Linguistic variation ➔ Document silence.
  - Document silence ➔ Omission of the relevant documents.
  - Different words ➔ Convey the idea.

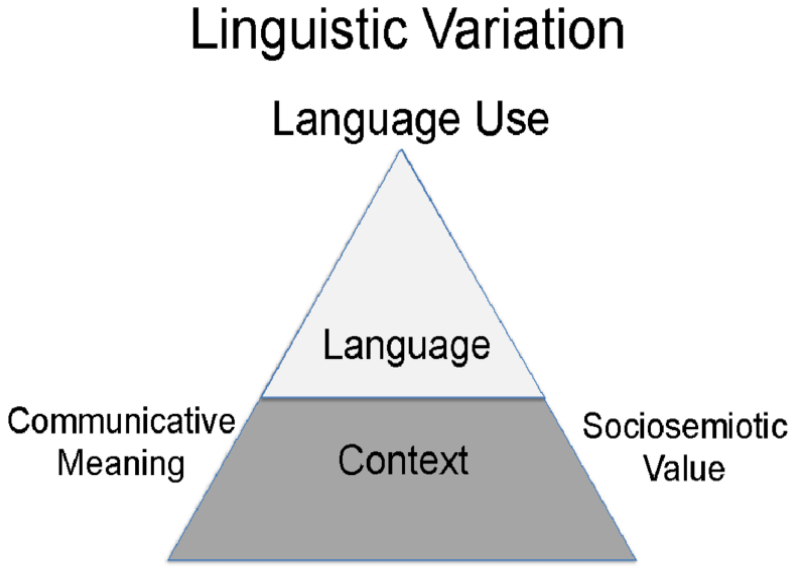

Figure: Linguistic variation

Source: https://www.researchgate.net/figure/Discourse-and-context-in-linguistic-variation_fig1_235951304

NLP in information retrieval

- Indexing.
- Query analysis.
- Comparison.
- Result processing.

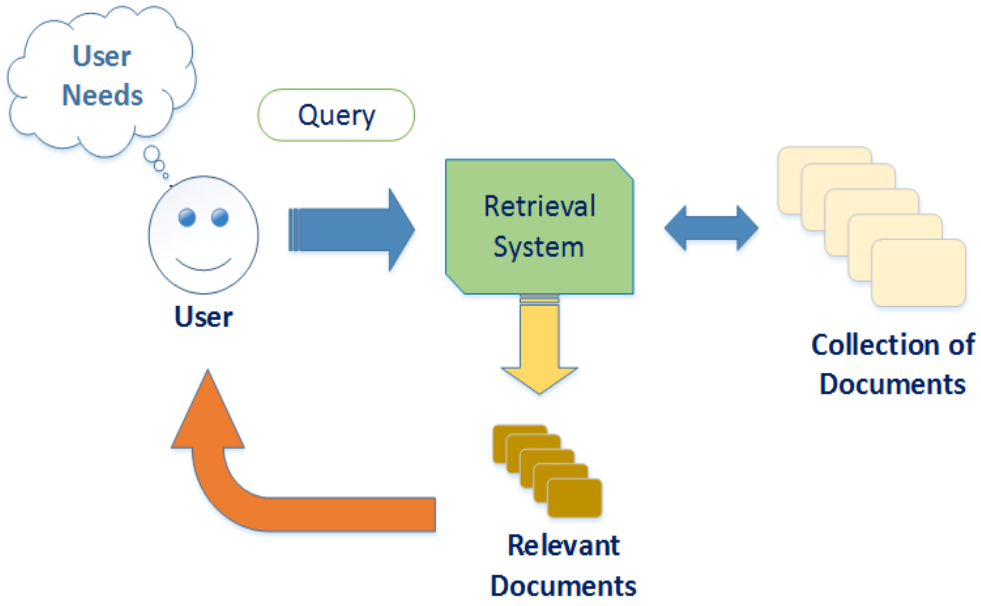

Figure: Information Retrieval system outline

Source: https://ir.cs.ui.ac.id/new/

IR evaluation metrics

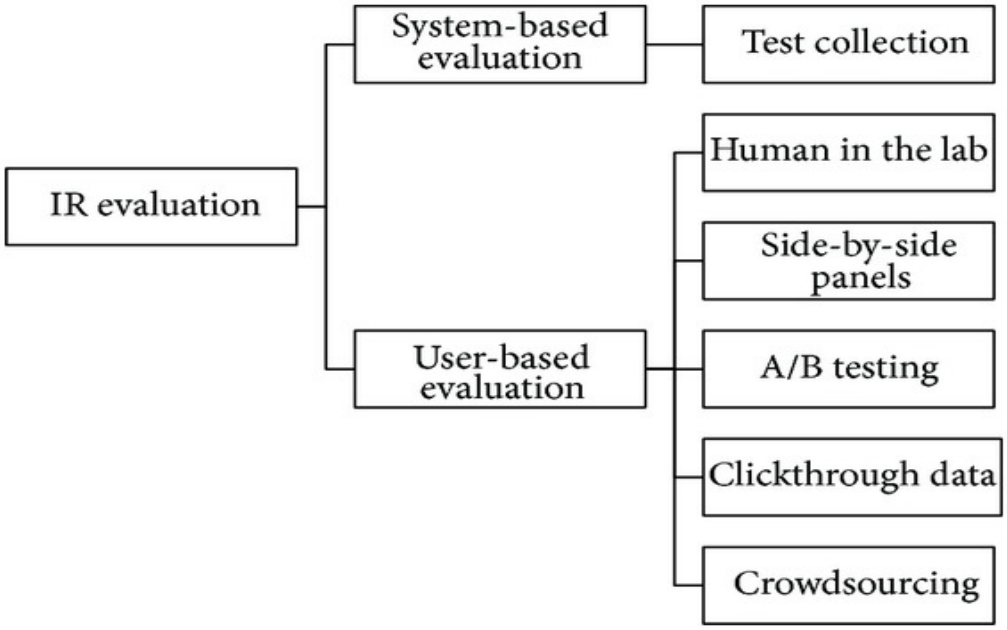

Figure: IR Evaluation Methods

Source: https://www.researchgate.net/figure/Classification-of-IR-evaluation-methods_fig1_263517579

Information Retrieval (IR) model and types

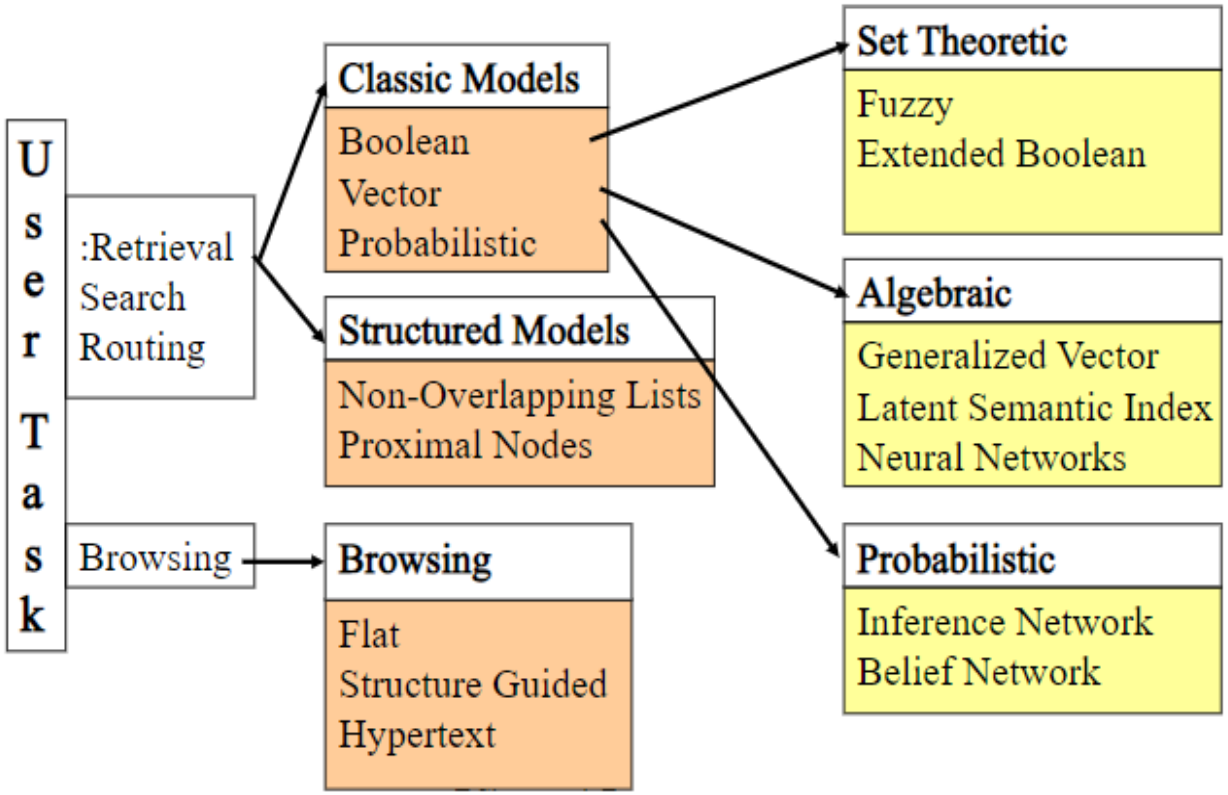

Figure: Information Retrieval Model

Source: https://slideplayer.com/slide/4905733/

Design features of IR systems (1 of 2)

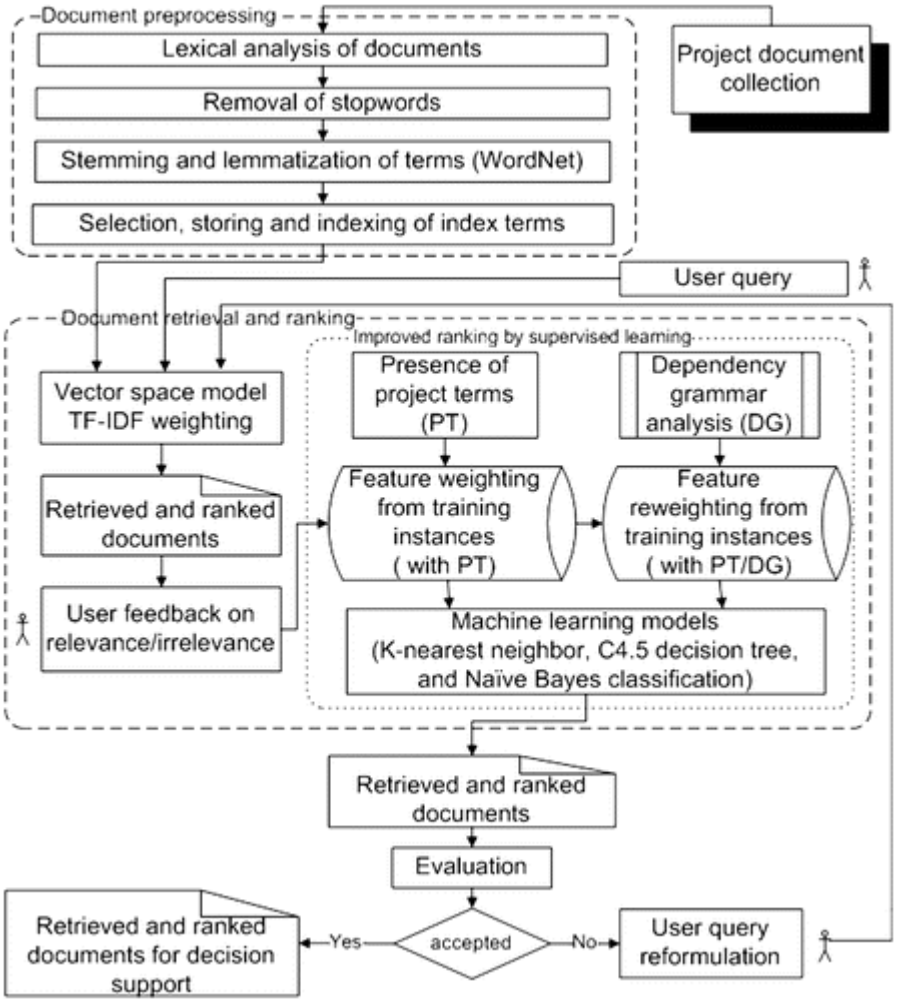

Figure: Design features of IR systems

Source: https://ascelibrary.org/doi/10.1061/%28ASCE%29ME.1943-5479.0000341

Design features of IR systems (2 of 2)

- NLP in IR ➔ Large computational cost.
- NLP and Information retrieval ➔ Same algorithm for better performance.
- Success rate ➔ Length of the queries.

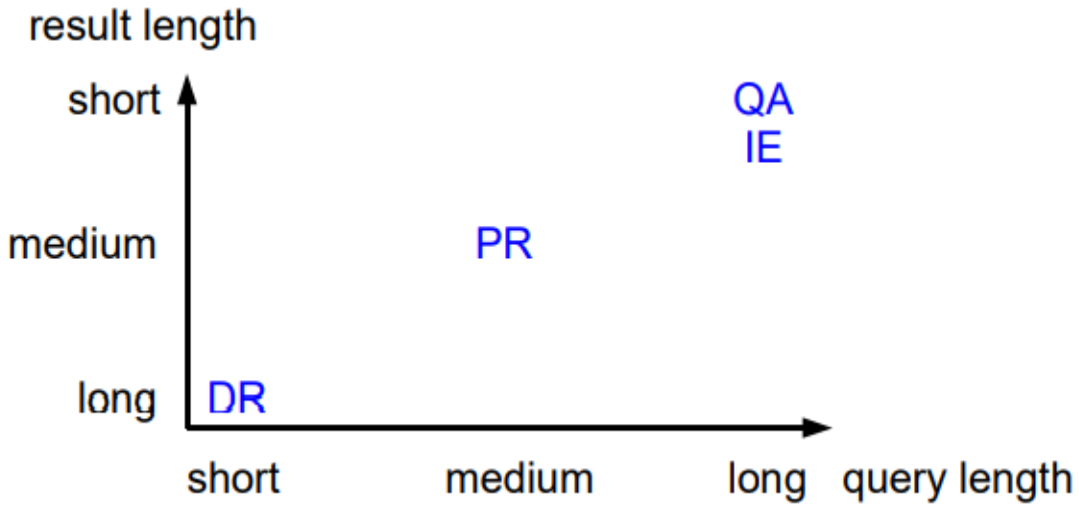

Figure: Classification of Document Retrieval, Passage Retrieval, Question Answering and Information Retrieval according to query length and result length

Source: https://pdfs.semanticscholar.org/8721/f2a087ff35318a056a5814ba287a37df0ec8.pdf

Question answering systems

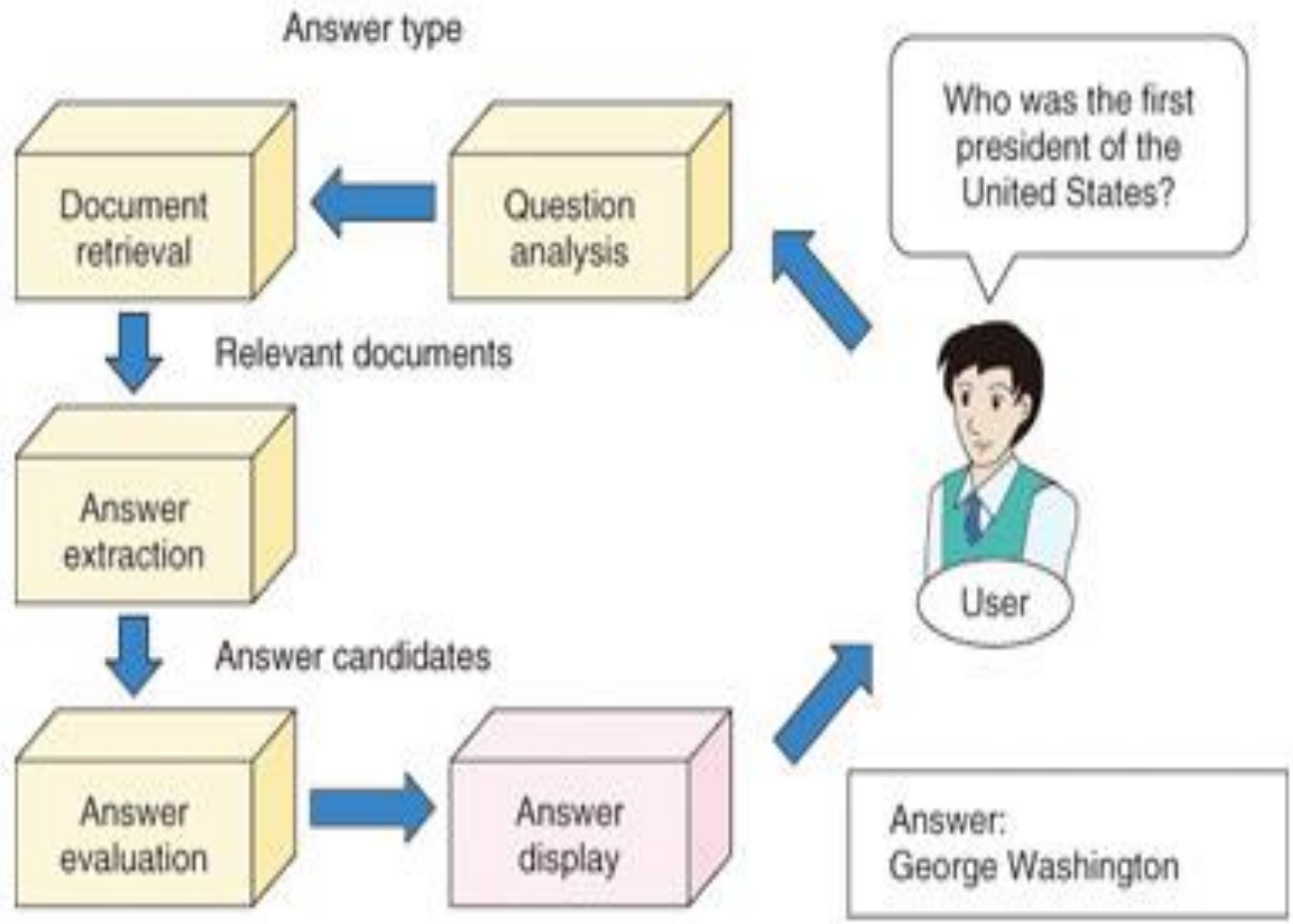

Figure: QA system Outline

Source: https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html

QA system architecture

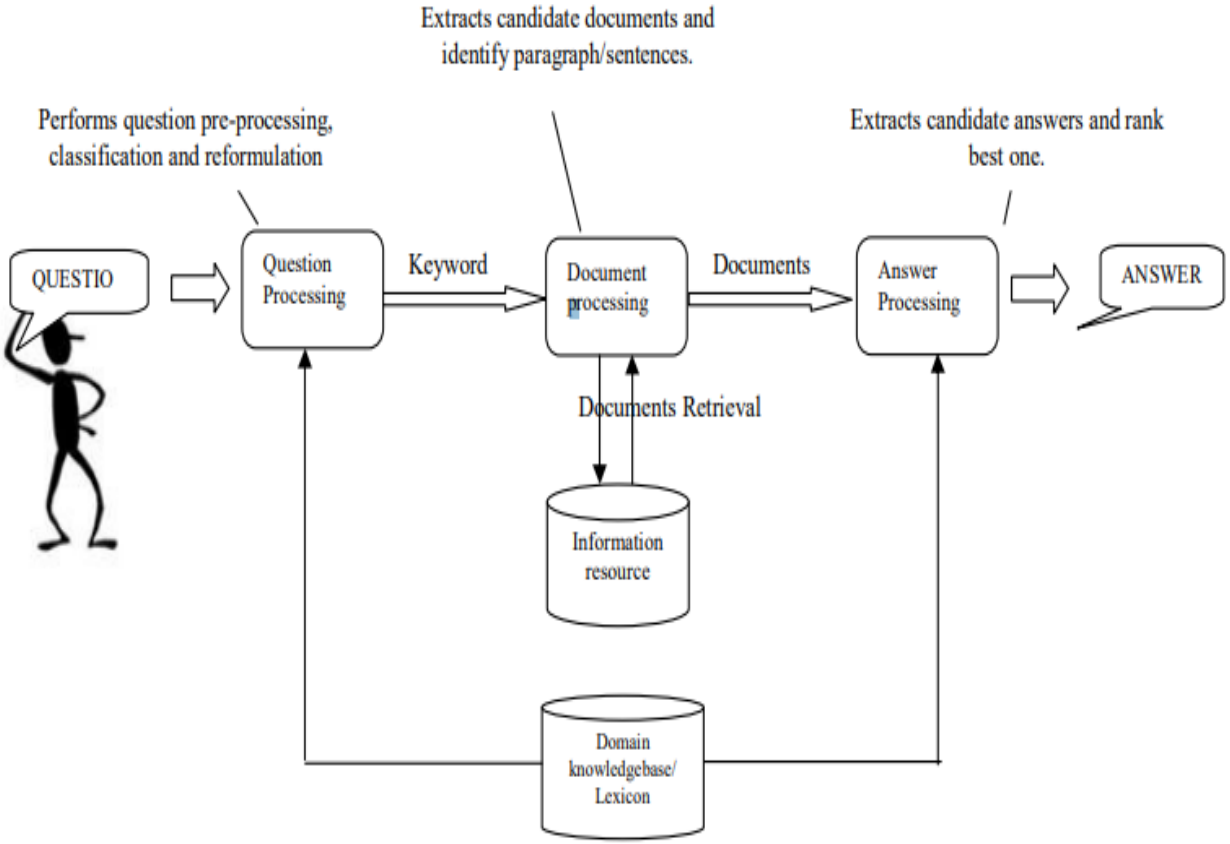

Figure: QA System Architecture

Source:
https://www.researchgate.net/publication/323729727_Different_Facets_of_Text_Based_Automated_Question_Answering_System/link/5aca58a20f7e9bcd5198adf1/download

QA system types

Figure: QA System Types

Source: https://www.researchgate.net/publication/320978810_An_Overview_of_Question_Answering_System

Text based QA systems

	Standard QA	Multilingual QA	Community QA	Interactive QA
Question type	Single sentence questions (Mostly factoid).	Single sentence questions in different accepted language	Multi sentence questions (Usually non-factoid).	Series of single sentence questions to have better understanding of the subject.
Question Reformulation	Automatic	Automatic	Manual	Automatic but user guided
Question Understanding	Depends on techniques (Shallow or deep linguistics) implemented	Depends on techniques implemented.	Depends on the understanding of community members responding to the asked question.	Good understanding as question representation is improved by real time interaction.
Answer Resource	Corpus, Knowledge base, Web documents	Corpus, Knowledge base, Web documents available for different languages	User (Expert) generated	Corpus, Knowledge base, Web documents
Answer representation	Short	Short	Long answers (or as required to the question)	Mixed answers
Answer reliability	Usually high	Average	Depends on potential experts	Average
Time lag	Immediate	Immediate	Have to wait until an answer is posted.	Real time response

Figure: Various QA System Representation

Source:
https://www.researchgate.net/publication/323729727_Different_Facets_of_Text_Based_Automated_Question_Answering_System/link/5aca58a20f7e9bcd5198adf1/download

Factoid question answering system

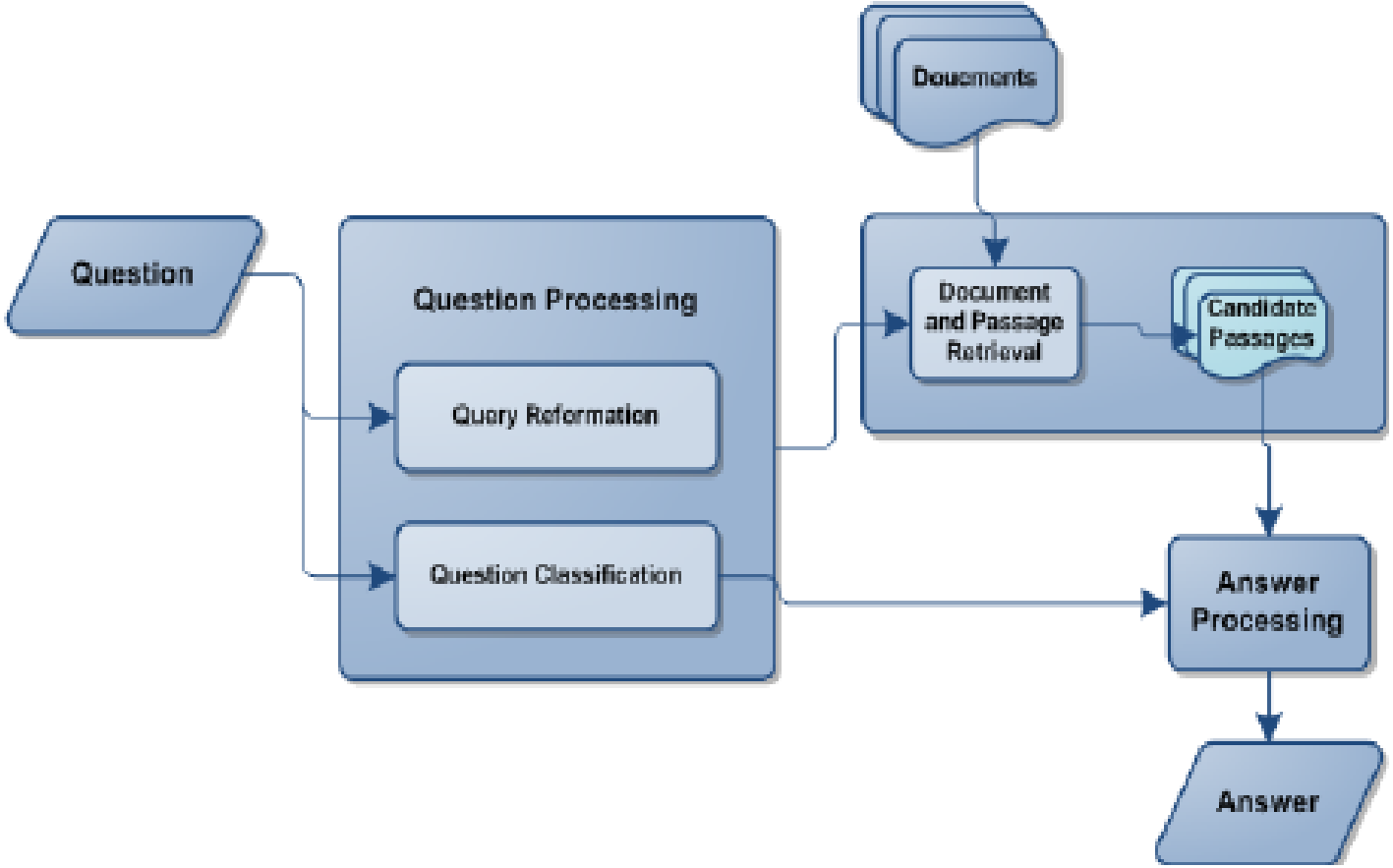

Figure: Factoid Question Answering system

Source: https://www.researchgate.net/figure/The-common-architecture-of-a-factoid-question-answering-system_fig1_241886726

Web based question answering system

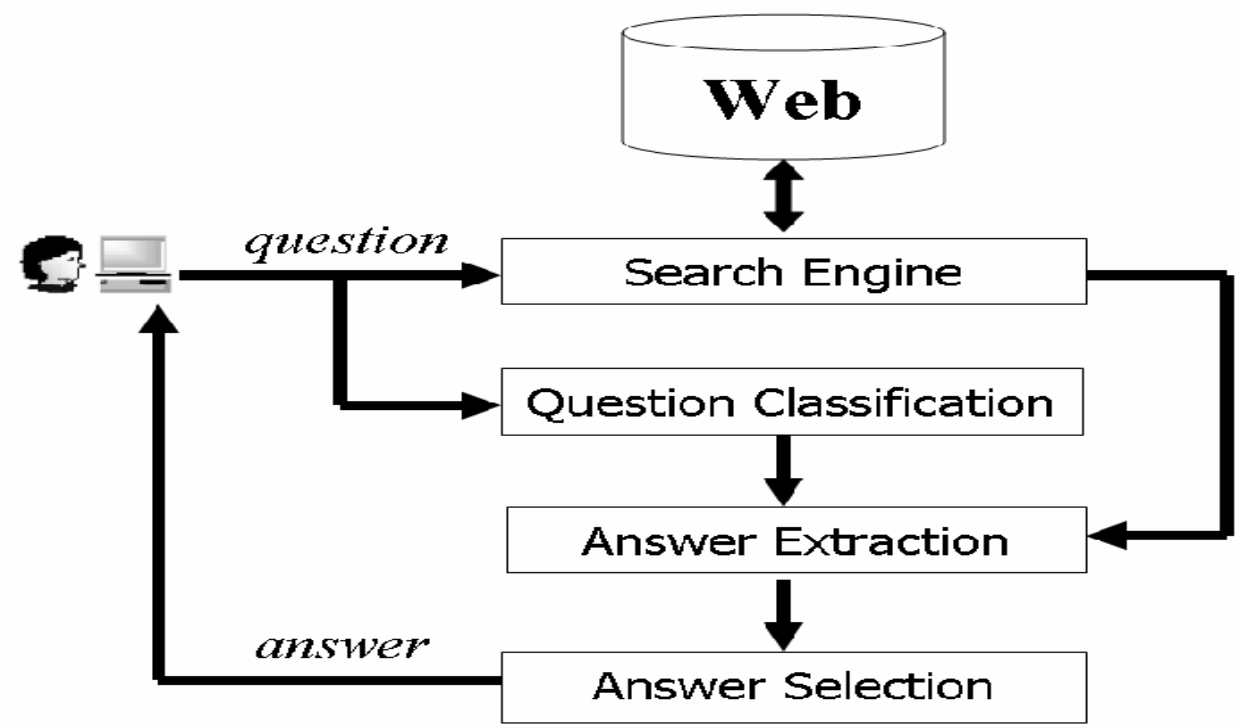

Figure: Web Based Question Answering System

Source: https://www.semanticscholar.org/paper/A-Web-based-Question-Answering-System-Zhang-Lee/ad23647c895d57668fc202259dccbf29edb9e683

Information retrieval or information extraction based QA systems

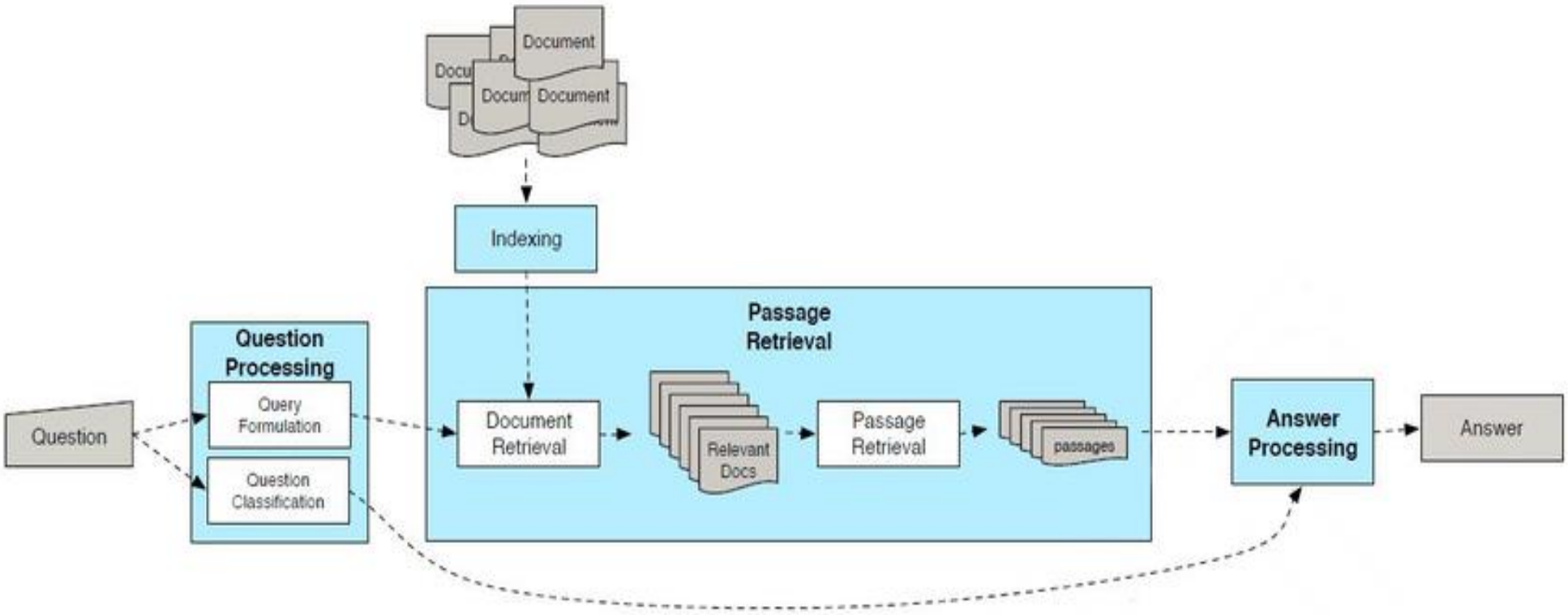

Figure: Information Retrieval or Information Extraction based QA systems

Source: https://www.researchgate.net/figure/Information-retrieval-based-QA-system-procedure-115_fig1_273122359

Restricted domain question answering

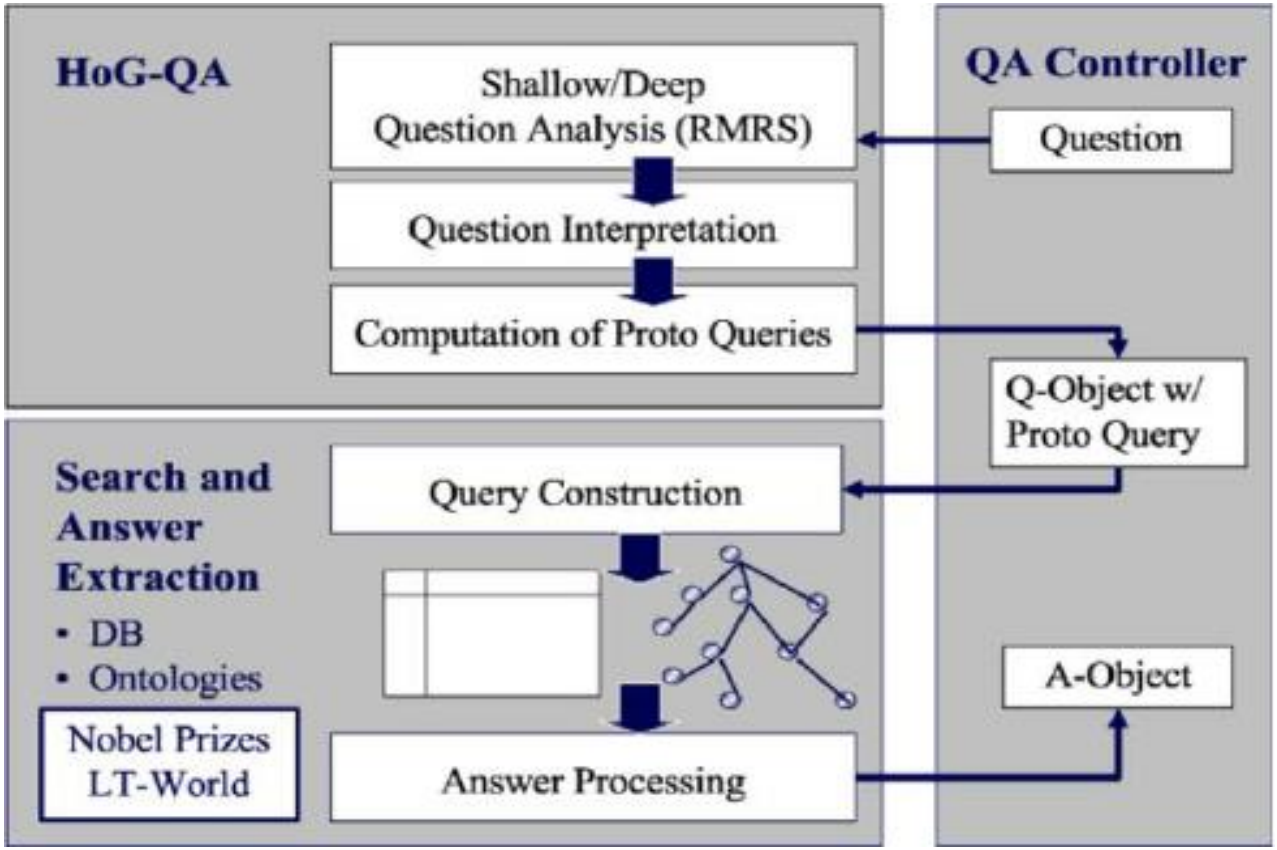

Figure: Restricted Domain Question Answering

Source: https://www.researchgate.net/figure/Architecture-of-Domain-Restricted-question-answering-system_fig3_258651905

Rule based question answering systems

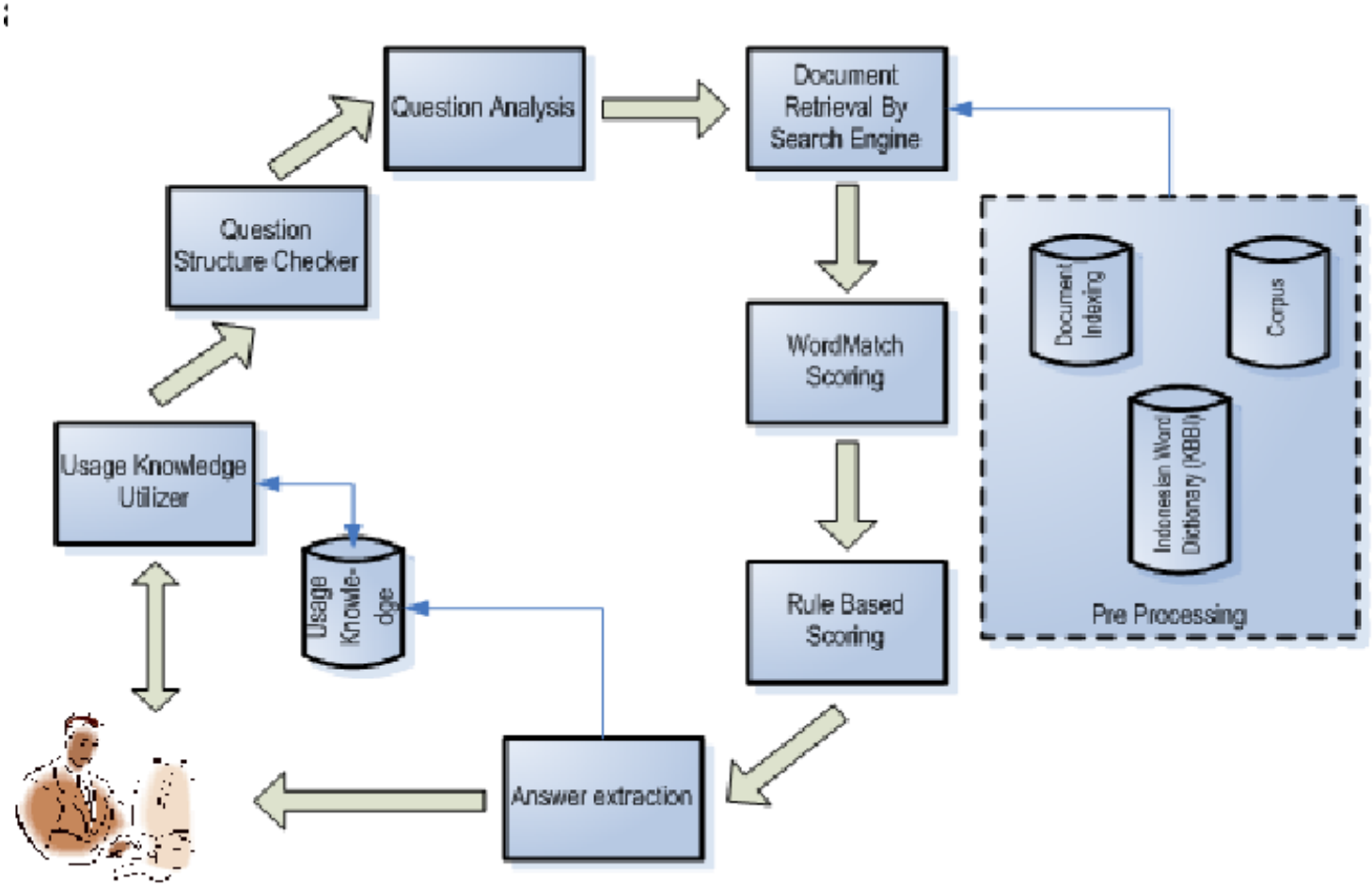

Figure: Rule Based Question Answering Systems

Source: https://www.semanticscholar.org/paper/A-rule-based-question-answering-system-on-relevant-Gusmita-Durachman/995e84a3c0e4c5877df0e214f7bfc8355d0c616f

Self evaluation: Exercise 15

- To continue with the training, after learning the concepts of Information Retrieval and Question Answering in Natural Language Text Processing, it is time to write code to work with IR in NLP using the earlier topics implementing POS tagging, Tokenization and use it in Information Retrieval Process. It is instructed to utilize the concepts of reading data from files Tokenization, Word Similarity, POS tags, Lemmatization, Word Embeddings and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 15: Build a recommendation system with text data and perform Information retrieval through queries.

Self evaluation: Exercise 16

- To continue with the training, after learning the concepts of Information Retrieval and Question Answering in Natural Language Text Processing, it is time to write code to work with IR in NLP using the earlier topics implementing POS tagging, Tokenization and use it in Information Retrieval Process. It is instructed to utilize the concepts of reading data from files Tokenization, Word Similarity, POS tags, Lemmatization, Word Embeddings and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 16: Create a Rule-Based Chat bot.

Information extraction (1 of 2)

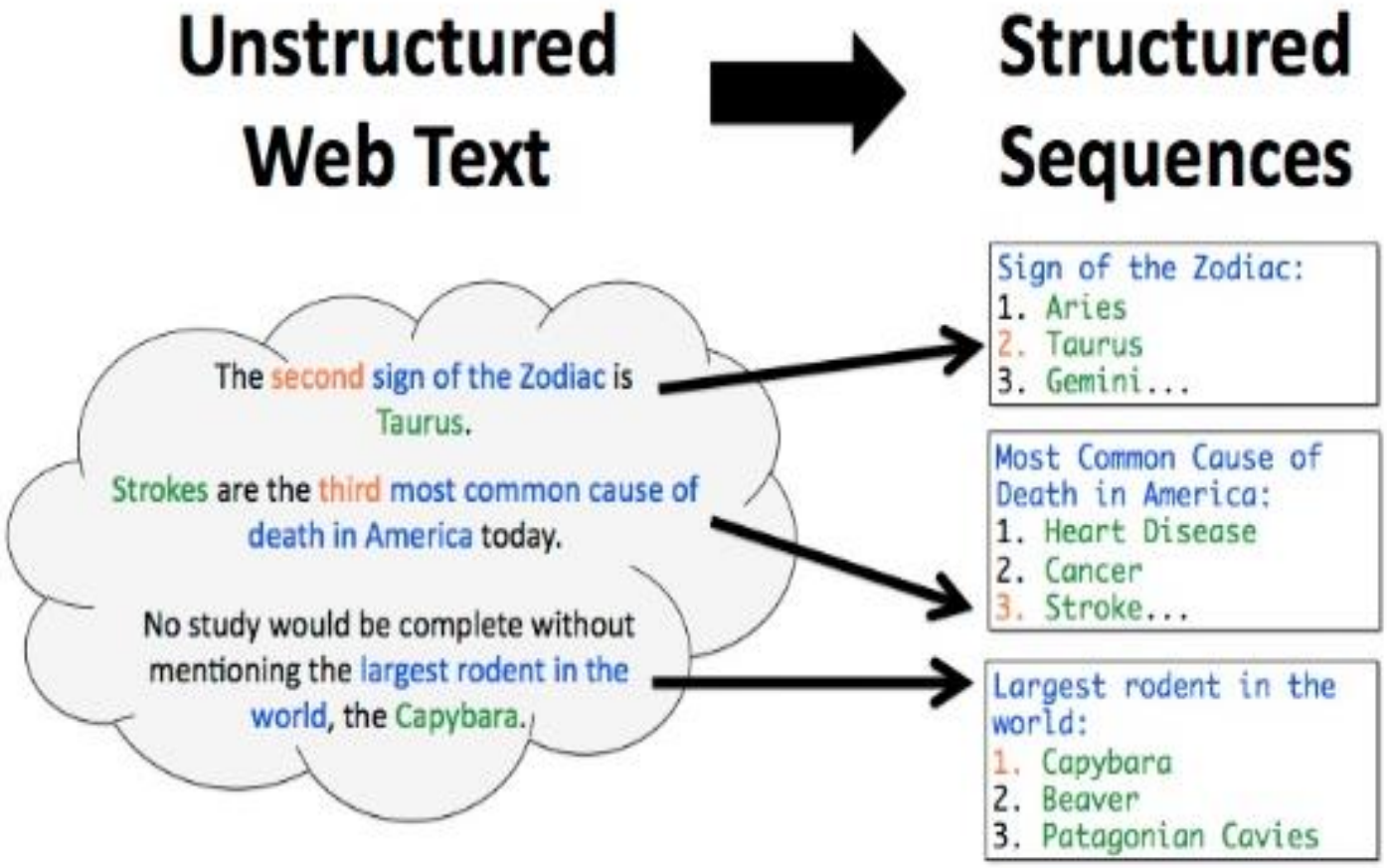

Figure: Information Extraction

Source: https://www.slideshare.net/rubenizquierdobevia/information-extraction-45392844

Information extraction (2 of 2)

Indian captain Virat Kohli was dismissed cheaply for just 2 in Wellington on Friday by debutant Kyle Jamieson extending a rare lull in the batsman's stellar career. Throughout the ongoing New Zealand tour, Kohli has managed to score just a single fifty across 8 innings in all 3 international formats.

Figure: Sample Document

Source: https://www.analyticsvidhya.com/blog/2020/06/nlp-project-information-extraction/

- The following information can be extracted from the text:

Country – India, Captain – Virat Kohli

Batsman – Virat Kohli, Runs – 2

Bowler – Kyle Jamieson

Match venue – Wellington

Match series – New Zealand

Series highlight – single fifty, 8 innings, 3 formats

Figure: Extracted Information

Working of information extraction

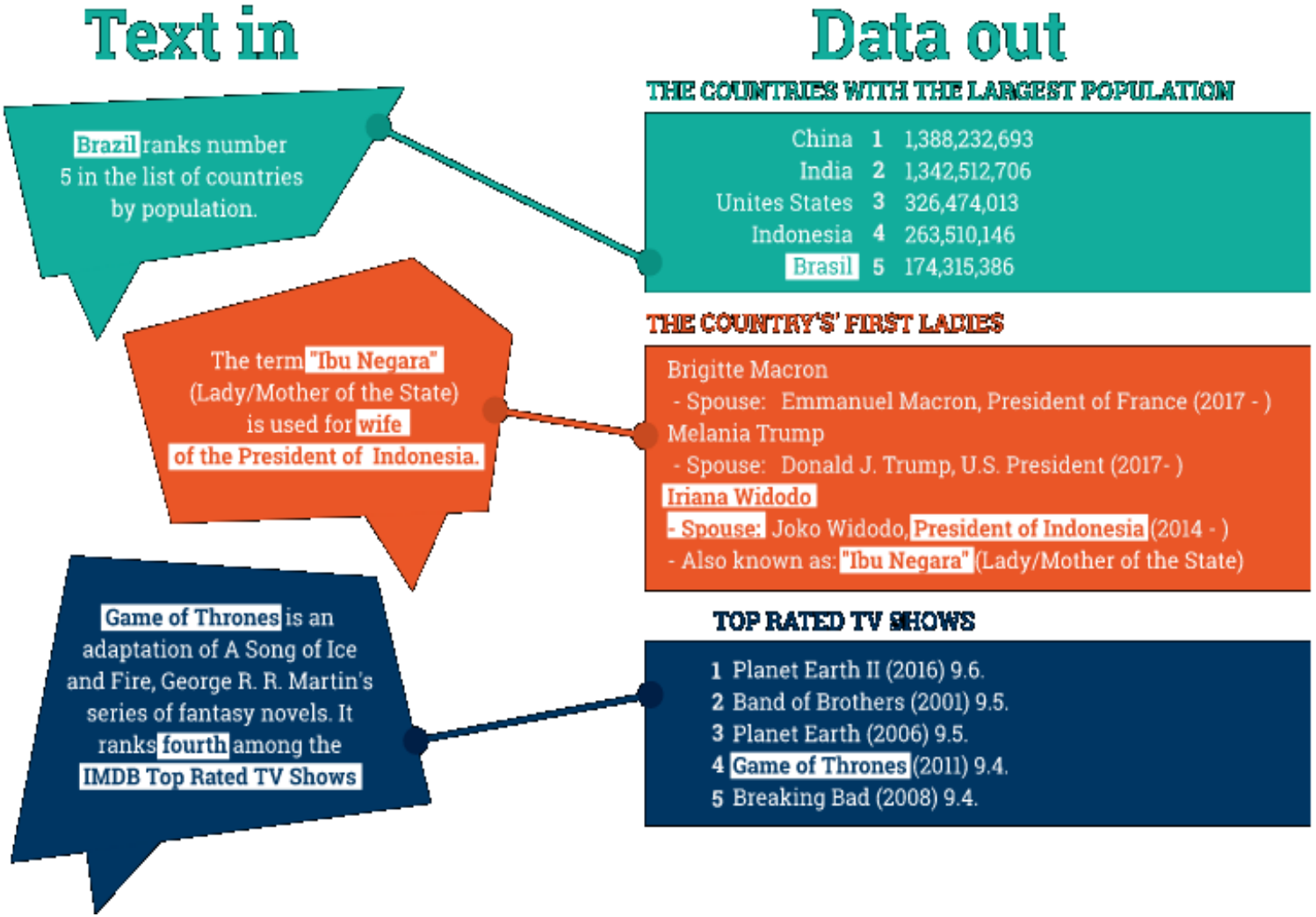

Figure: Simple IE Activity

Source: https://www.ontotext.com/knowledgehub/fundamentals/information-extraction/

Information extraction applications (1 of 2)

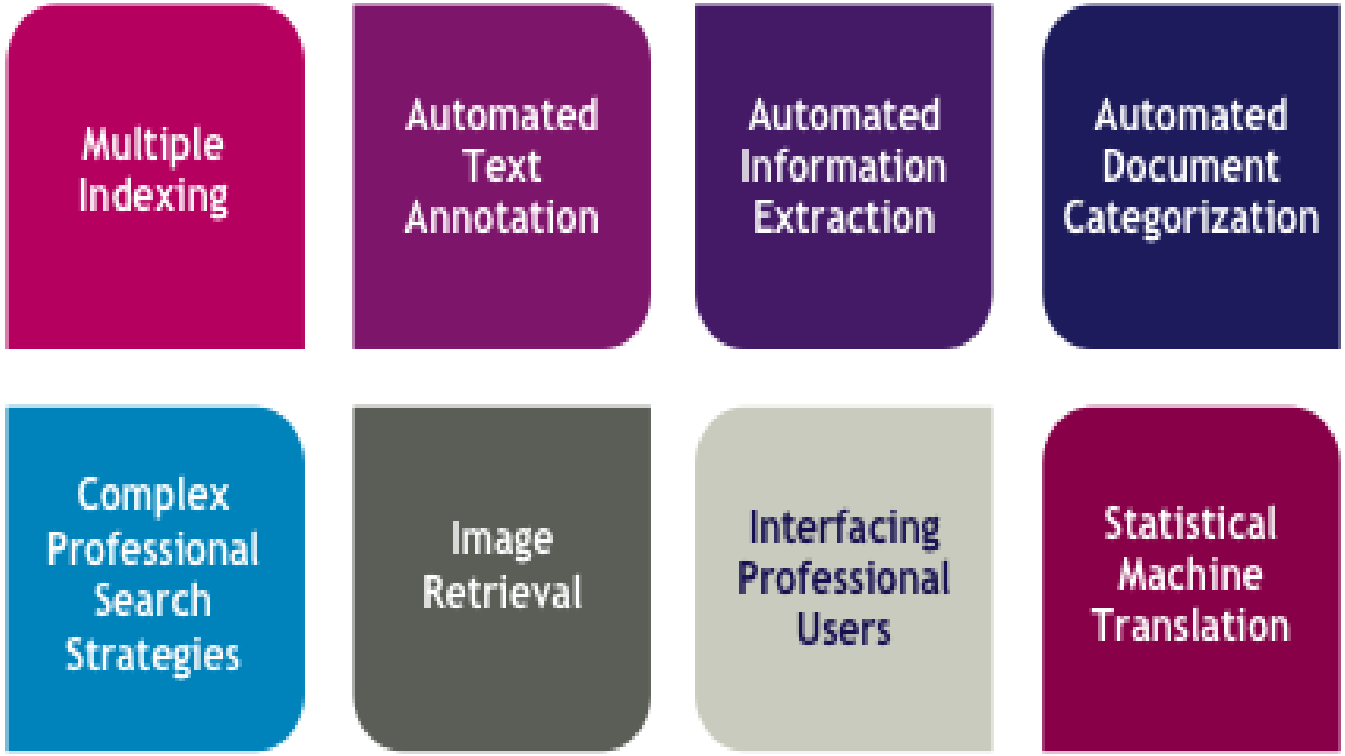

Figure: IE Sectors

Source: https://www.ir-facility.org/research-areas

Information extraction architecture (2 of 2)

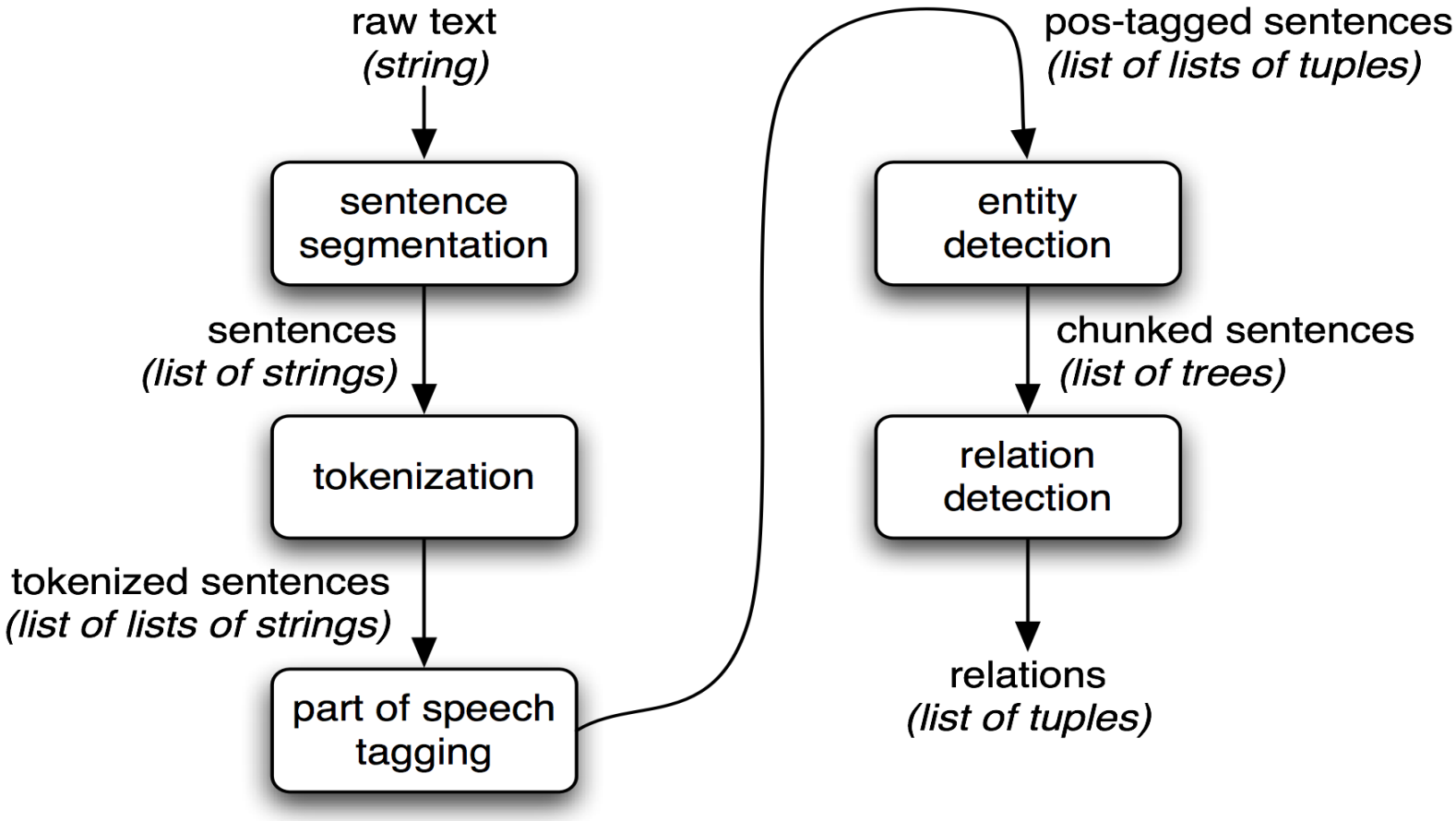

Figure: Information Extraction Architecture

Source: https://www.nltk.org/book/ch07.html

Chunking (1 of 2)

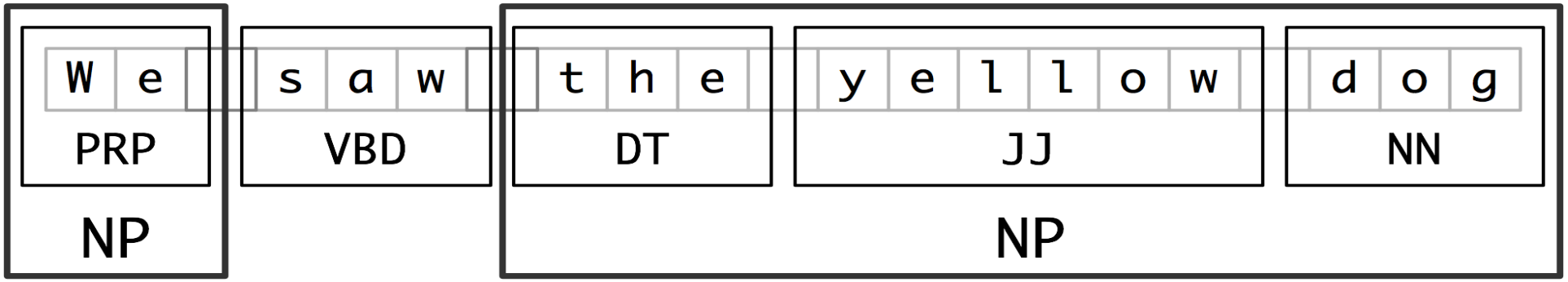

Figure: Simple Tokens and Chunks

Source: https://www.nltk.org/book/ch07.html#sec-ner

Chunking (2 of 2)

- Regular expression Chunking:

```
nouns = [("money", "NN"), ("market", "NN"), ("fund", "NN")]
grammar = "NP: {<NN><NN>} # Chunk two consecutive nouns"
nltk.RegexpParser(grammar)
```

- Structure: (S (NP money/NN market/NN) fund/NN).

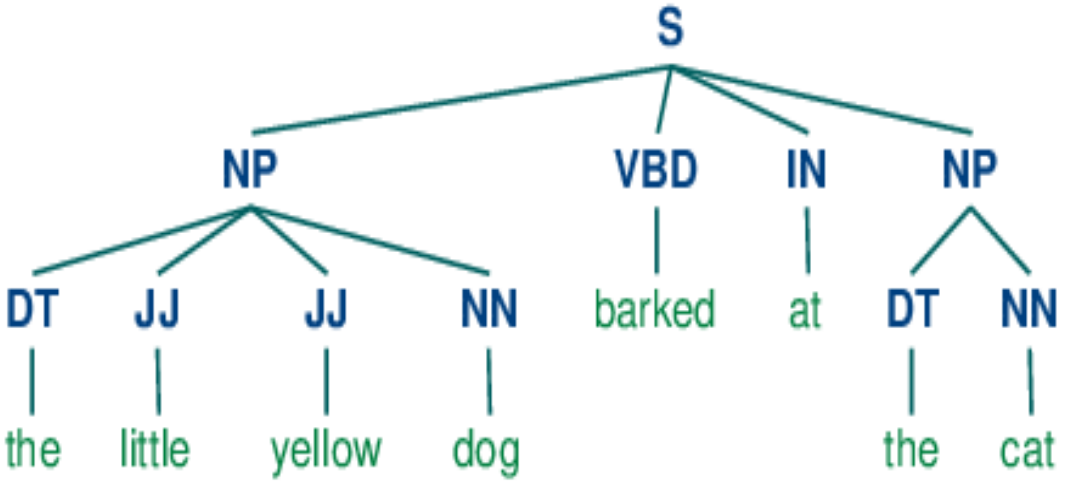

Figure: Parse Tree

Source: https://www.nltk.org/book/ch07.html#sec-ner

Representing chunks: Tags vs trees

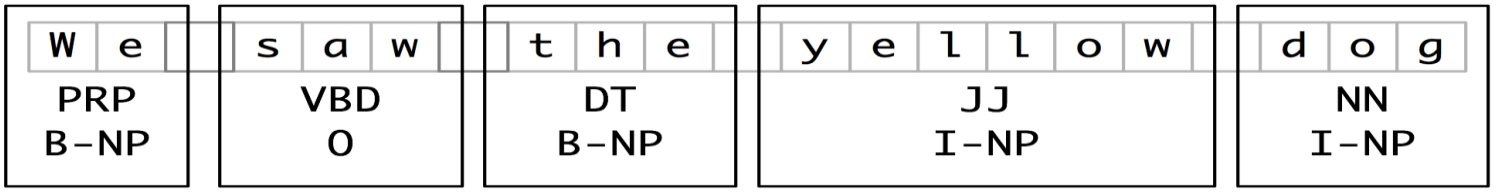

Figure: Chunks with Tags

Source: https://www.nltk.org/book/ch07.html#sec-ner

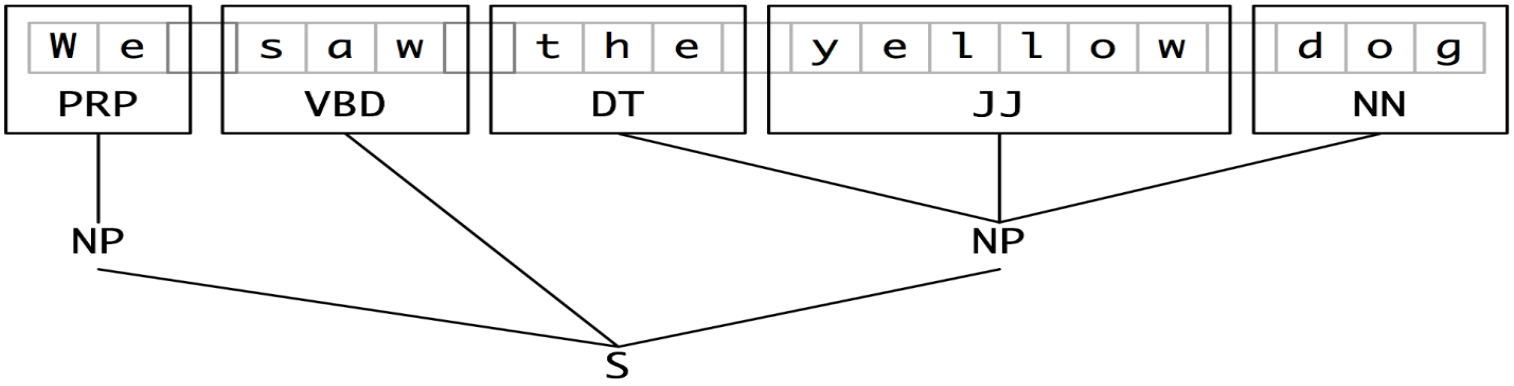

Figure: Chunks with Trees

Source: https://www.nltk.org/book/ch07.html#sec-ner

Self evaluation: Exercise 17

- To continue with the training, after learning the concepts of Information Extraction in Natural Language Text Processing, it is time to write code to work with IE in NLP using the earlier topics implementing POS tagging and use it to extract information from any text. It is instructed to utilize the concepts of POS tags and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 17: Extract Information from Text using Spacy's Rule-Based Matching.

Self evaluation: Exercise 18

- To continue with the training, after learning the concepts of Information Extraction in Natural Language Text Processing, it is time to write code to work with IE in NLP using the earlier topics implementing POS tagging and use it to extract information from any text. It is instructed to utilize the concepts of POS tags and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 18: Perform Relation Extraction using Subtree Matching for representing text in Active and Passive Voices.

Self evaluation: Exercise 19

- To continue with the training, after learning the concepts of Information Extraction in Natural Language Text Processing, it is time to write code to work with IE in NLP using the earlier topics implementing POS tagging and use it to extract information from any text. It is instructed to utilize the concepts of POS tags and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 19: Information Extraction from a Text based on specific Pattern matching using Spacy's Matcher class elements.

Report generation

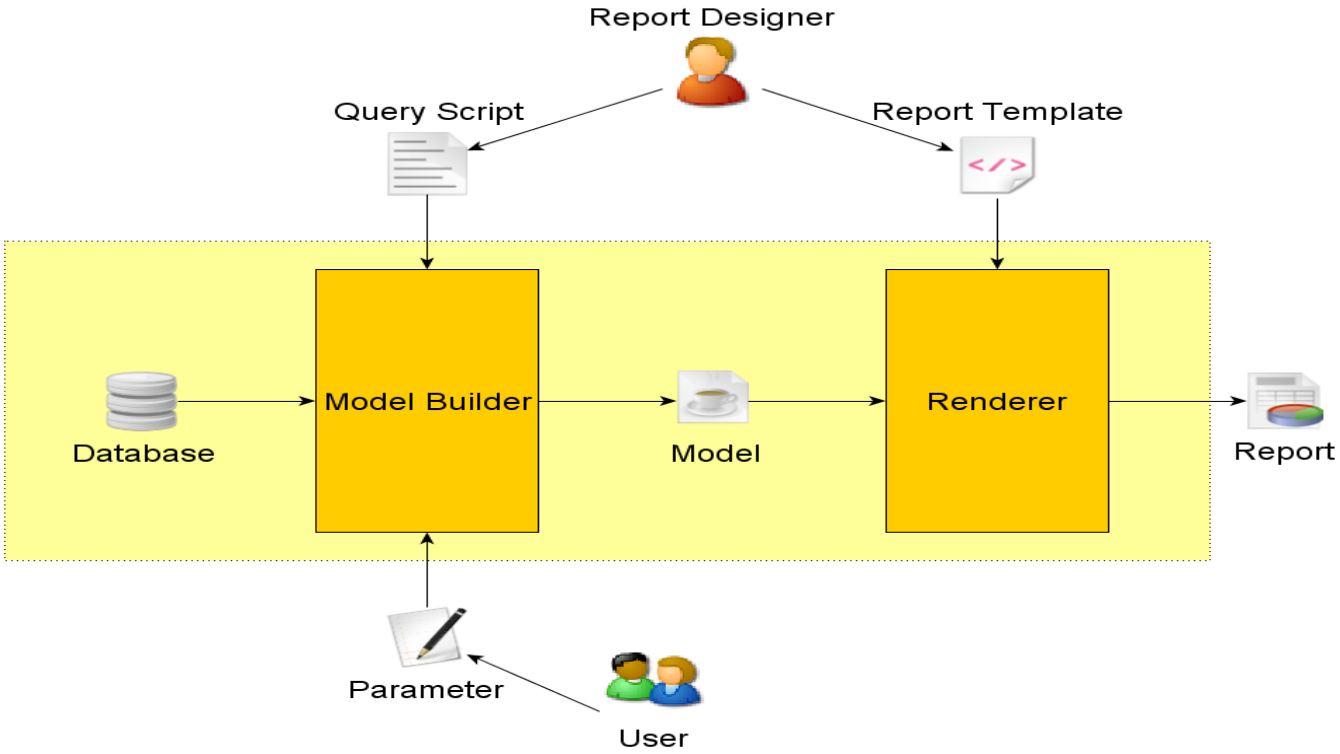

Figure: Report Generation Outline

Source: https://www.klaros-testmanagement.com/files/doc/html/User-Manual.CustomReport.html

Text report specifications

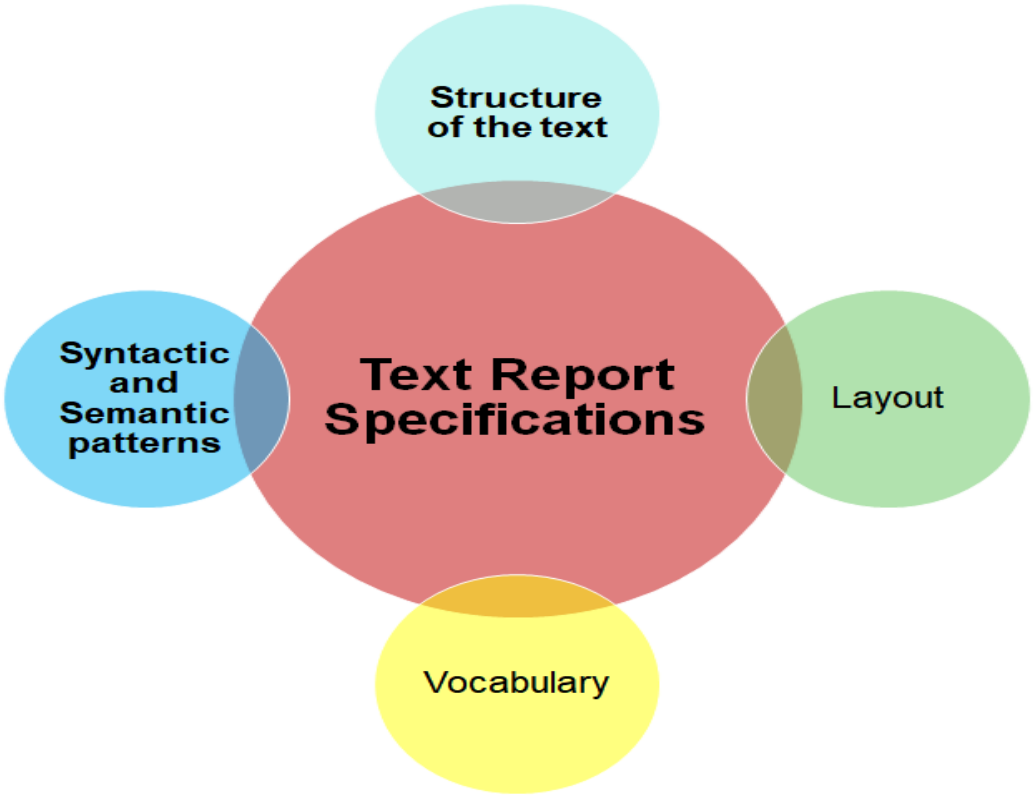

Figure: Specification for a Report

Features of reports

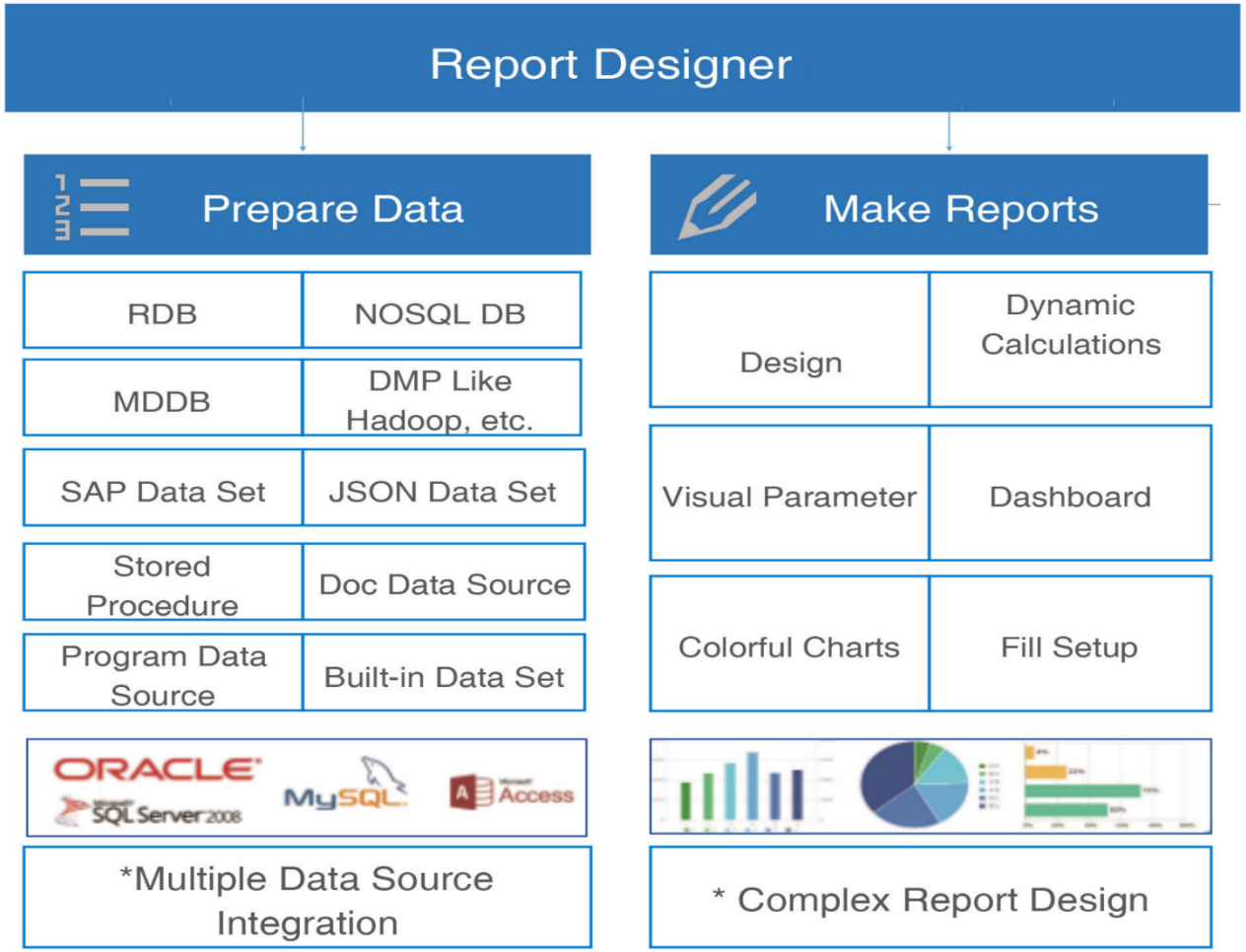

Figure: RG tool activities

Source: https://www.finereport.com/en/reporting-tools/report-generation.html

Report generation process

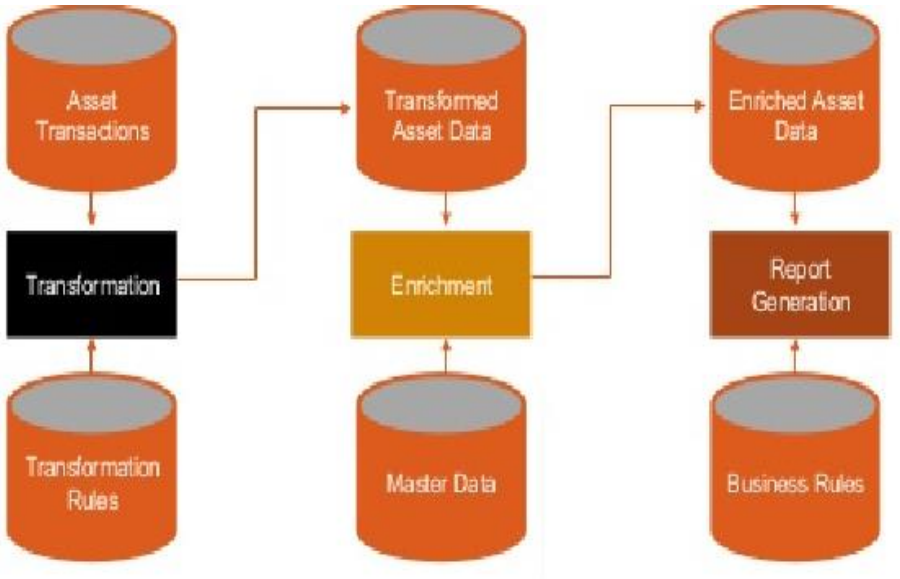

Figure: RG Process

Source: https://www.slideshare.net/SparkSummit/regulatory-reporting-of-asset-trading-using-apache-sparkdasgupta-rao

Usage of NLP text in report generation

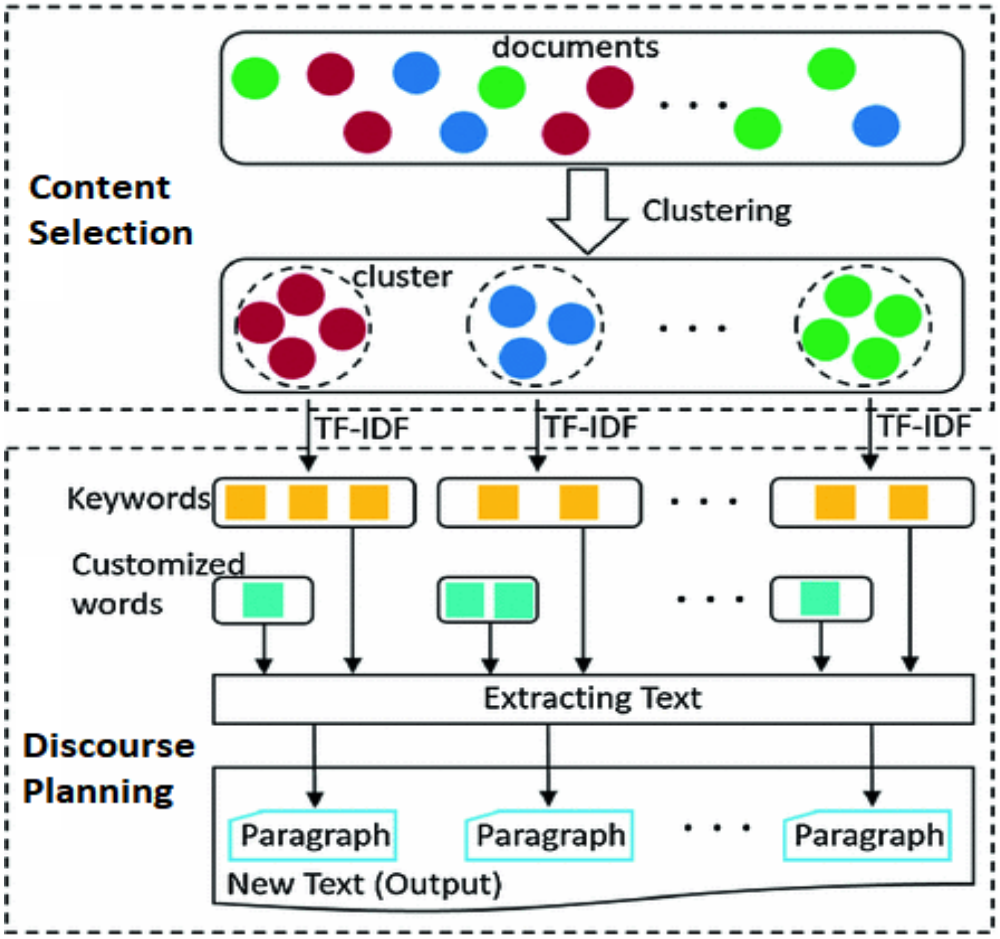

Figure: Usage of Text

Source: https://link.springer.com/chapter/10.1007/978-3-319-69781-9_23

Ontology construction

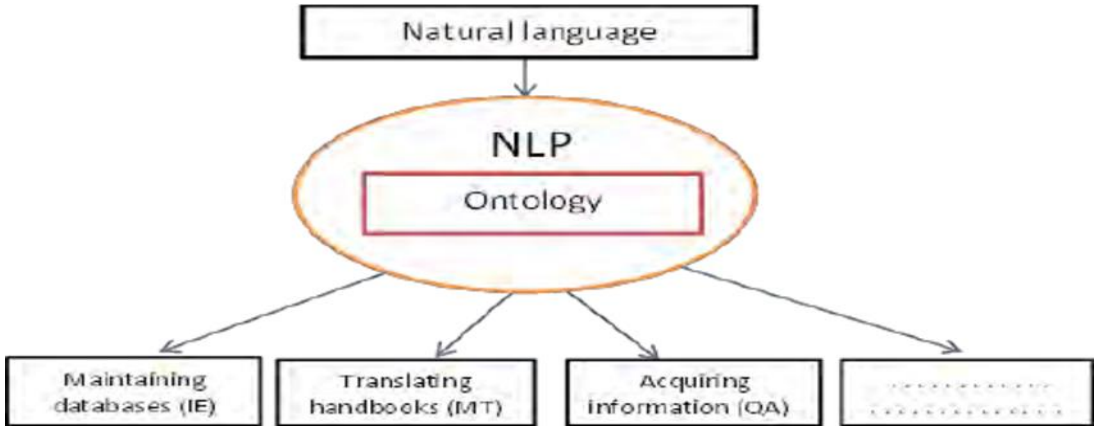

Figure: Ontology Representation

Source: https://www.slideshare.net/athmanhajhamou/use-of-ontologies-in-natural-language-processing-2

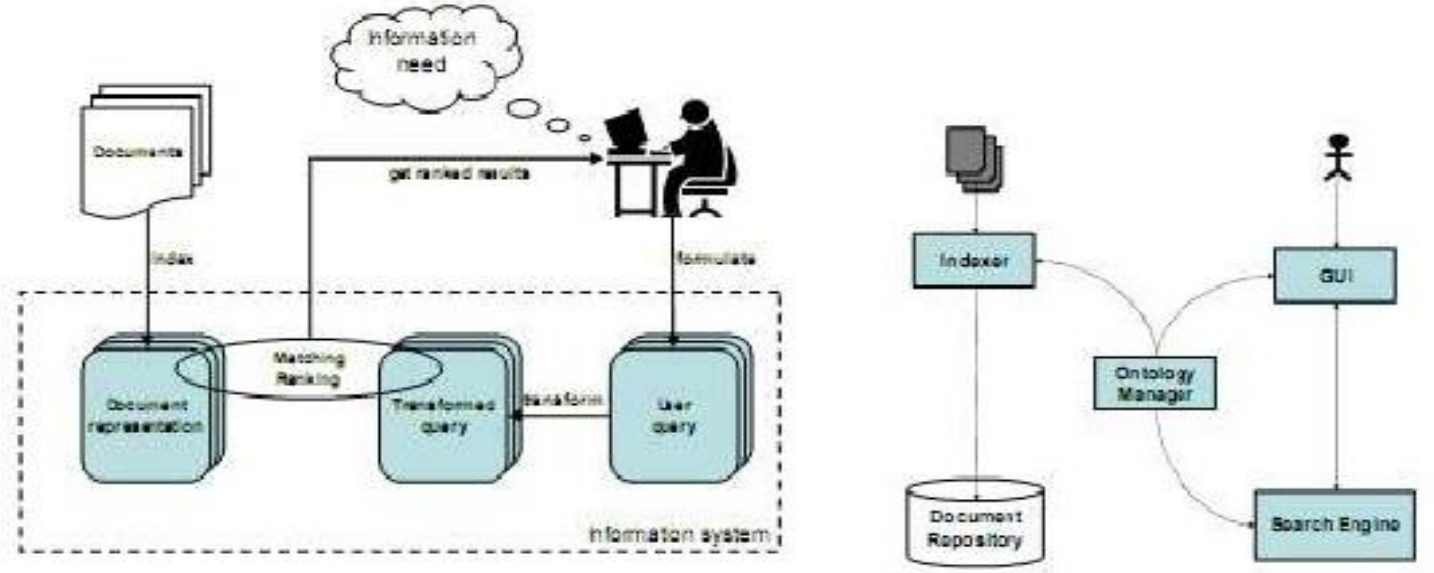

Figure: Ontology Outline in NLP

Source: https://www.researchgate.net/figure/Relation-between-natural-language-NLP-and-ontology_fig1_270471292

Ontology classifications and process

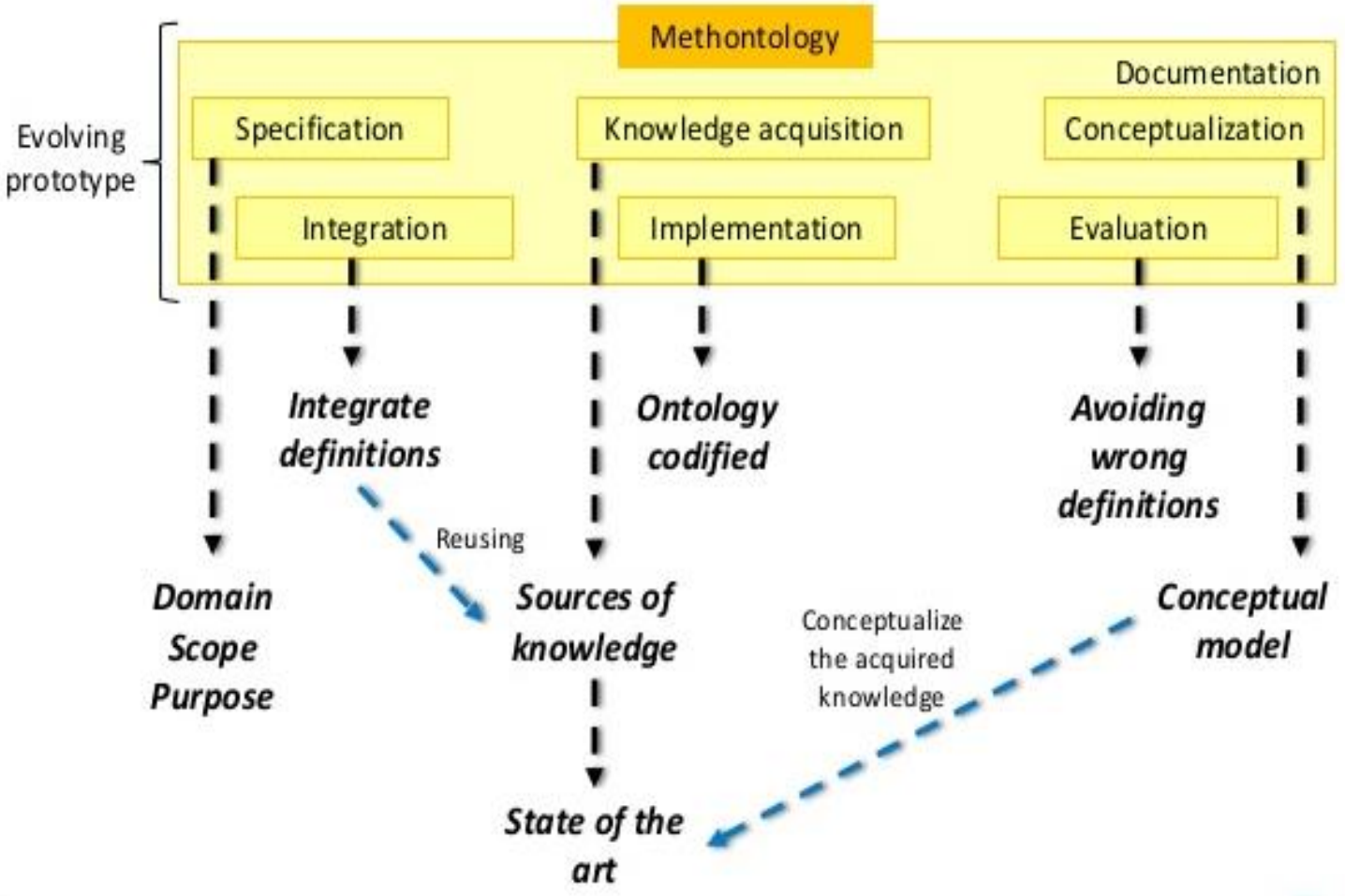

Figure: Ontology Process

Source: https://www.slideshare.net/gessiupc/rcis2014-35471364

Why ontology and its advantages

Traditional CMS	Contributions of Ontology
Match-making often ineffective because of rigid definition of contents (e.g., categories) predefined by service providers	Shared and agreed ontology provides common, flexible, and extensible definitions of multimedia contents for match-making and subsequent business processes
Difficult to specify unclear types of multimedia content out of predefined categories	Complicated use requirements can be decomposed into simple genres for elicitation of options

Figure: Advantages of Ontologies

Source: https://slideplayer.com/slide/5014295/

Ontology components

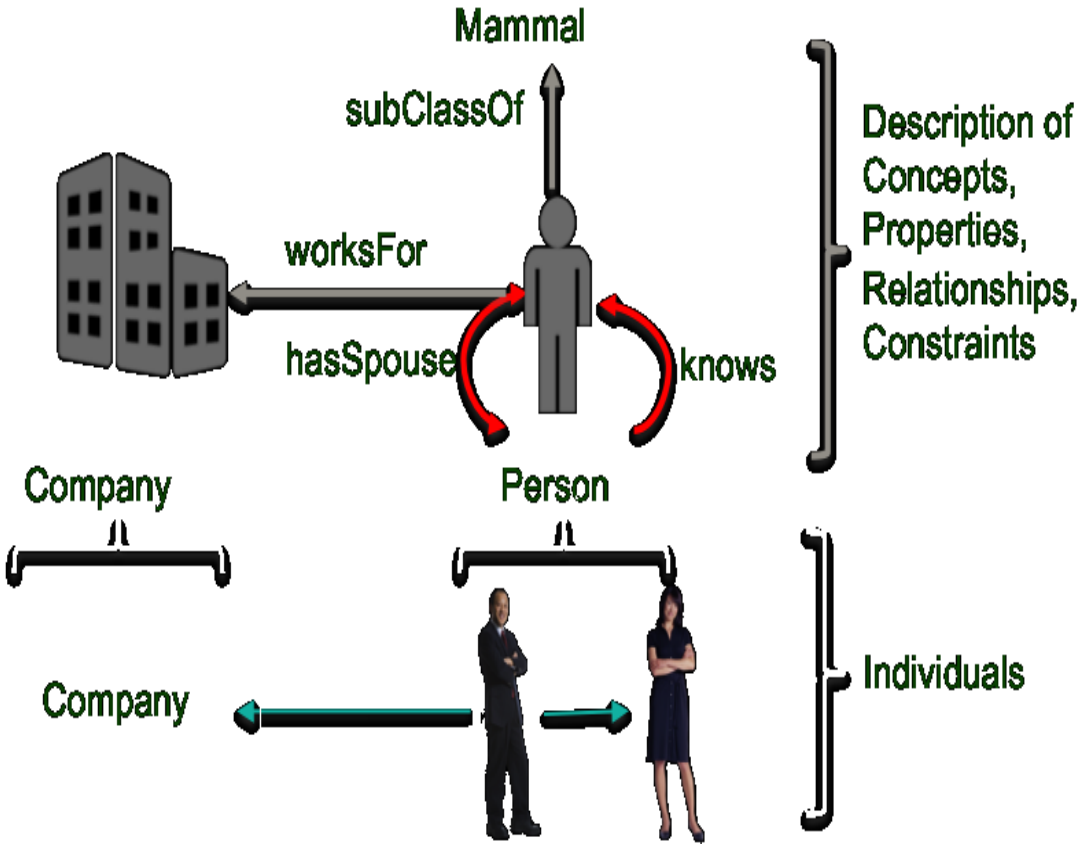

Figure: Ontology Components

Source: http://graphdb.ontotext.com/documentation/enterprise/devhub/ontologies.html

Levels of formality

- Highly informal:
  - Represented in simple natural language.
  - Wine is a winery product.

- Semi informal:
  - Represented in structured form of natural language.
  - Wine PRODUCED IN Winery.

- Semi formal: Represented in a formally defined language.

Winery $\xrightarrow{PRODUCES}$ Wine

Ontology construction approaches

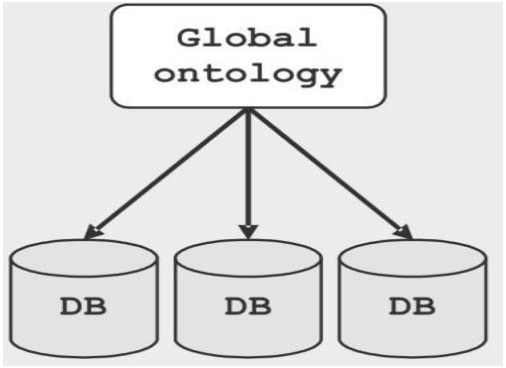

Figure: Single ontology approach

Source: https://www.researchgate.net/publication/220327569/figure/fig1/AS:411993979801600@1475238427128/Single-Ontology-Approach_Q640.jpg

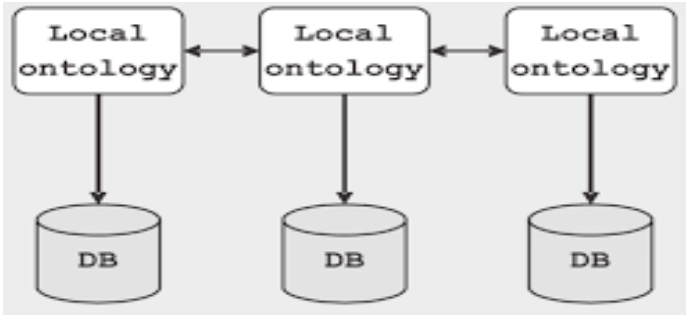

Figure: Multiple ontology approach

Source: https://www.researchgate.net/publication/220327569/figure/fig2/AS:411993979801601@1475238427796/Multiple-Ontology-Approach.png

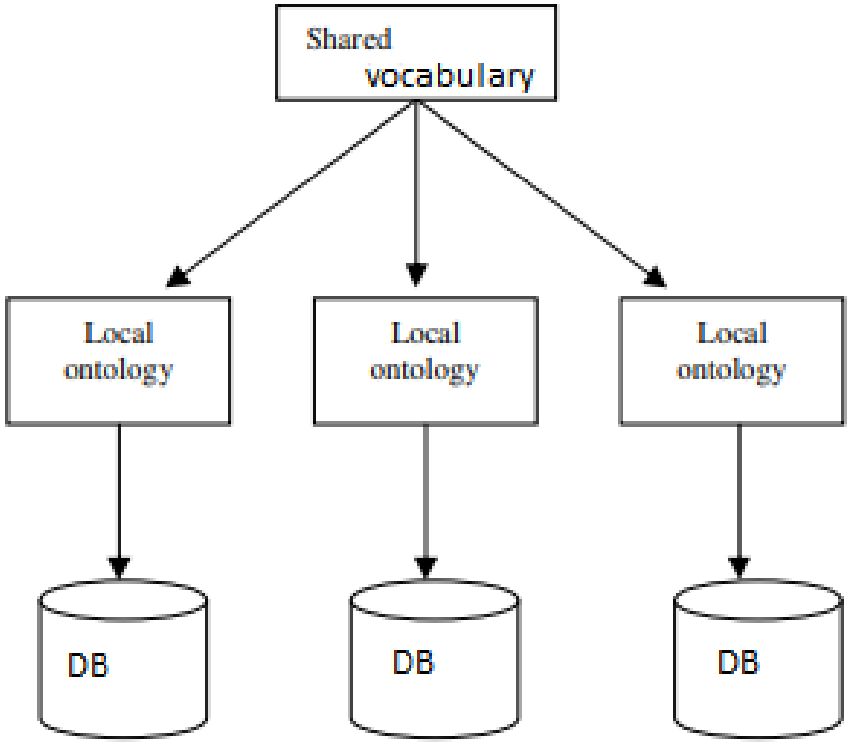

Figure: Hybrid ontology approach

Source: http://www.ijecbs.com/January2011/N5Jan2011.pdf

Ontology construction

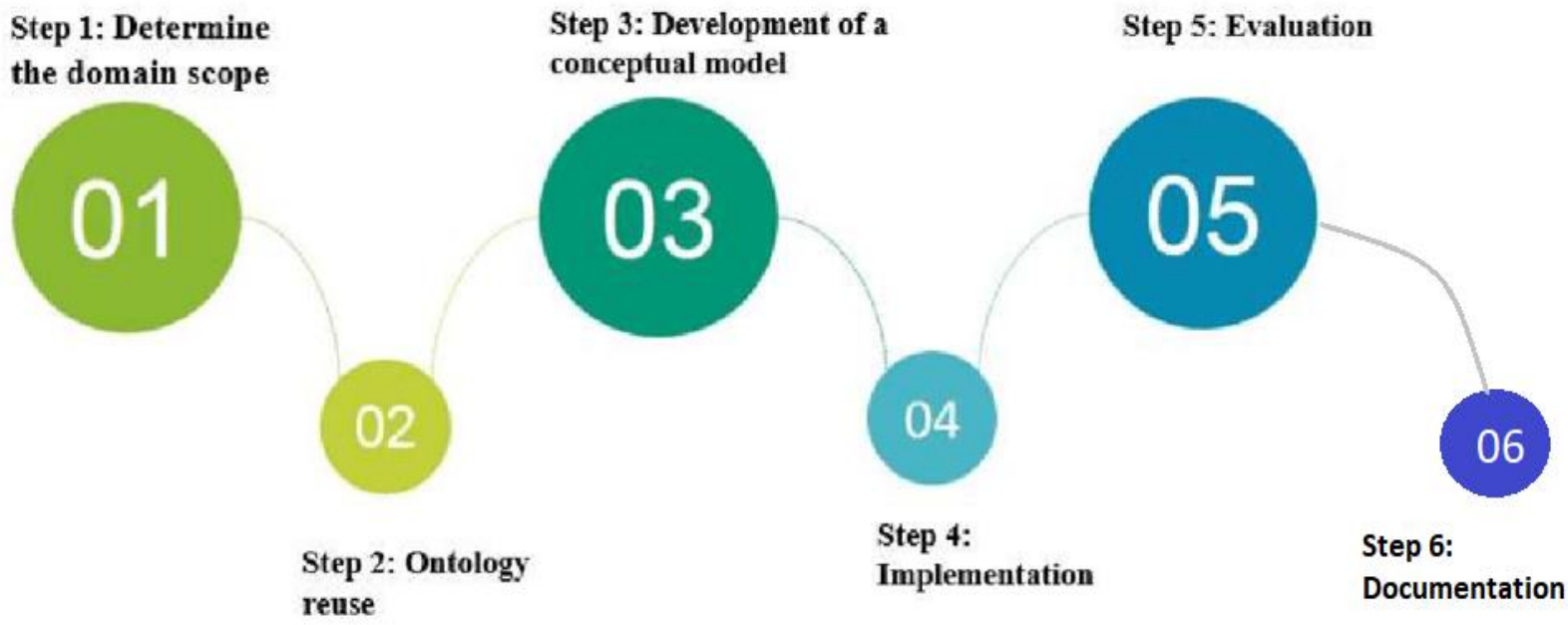

Figure: Ontology Creation Steps

Source: https://www.researchgate.net/figure/Ontology-Construction-the-basic-steps-proposed-by-Sanchez-58_fig3_329364160

Checkpoint (1 of 2)

Multiple choice questions:

1. A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
   a) Postings list
   b) Incidence Matrix
   c) Dictionary
   d) Inverted Index

2. In information retrieval, extremely common words which would appear to be of little value in helping select documents that are excluded from the index vocabulary are called:
   a) Stop Words
   b) Tokens
   c) Lemmatized Words
   d) Stemmed Terms

3. A group of related documents against which information retrieval is employed is called:
   a) Corpus
   b) Text Database
   c) Index Collection
   d) Repository

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
   a) Postings list
   b) Incidence Matrix
   c) Dictionary
   d) Inverted Index

2. In information retrieval, extremely common words which would appear to be of little value in helping select documents that are excluded from the index vocabulary are called:
   a) Stop Words
   b) Tokens
   c) Lemmatized Words
   d) Stemmed Terms

3. A group of related documents against which information retrieval is employed is called:
   a) Corpus
   b) Text Database
   c) Index Collection
   d) Repository

Checkpoint (2 of 2)

Fill in the blanks:

1. A crude heuristic process that chops off the ends of the words to reduce inflectional forms of words and reduce the size of the vocabulary is called _____.
2. _____ systems deal with queries that are limited to a domain.
3. _____ is the specification of shared conceptual terminologies.
4. _____ approach relies on Global ontology for the information resources.

True or False:

1. Stemming increases the size of the vocabulary. True/False
2. In multilingual question answering system the query and the response does not belong to a language but from multiple languages. True/False
3. Chunks created by various processes can be represented either through tags or through trees. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. A crude heuristic process that chops off the ends of the words to reduce inflectional forms of words and reduce the size of the vocabulary is called **Stemming.**
2. **Closed domain** systems deal with queries that are limited to a domain.
3. **Ontology** is the specification of shared conceptual terminologies.
4. **Single ontology** approach relies on Global ontology for the information resources.

True or False:

1. Stemming increases the size of the vocabulary. **False**
2. In multilingual question answering system the query and the response does not belong to a language but from multiple languages. **True**
3. Chunks created by various processes can be represented either through tags or through trees. **True**

Question bank

Two mark questions:

1. How is information retrieval in NLP achieved?
2. How does web-based question answering system work?
3. What is the difference between representing chunks as tags vs trees?
4. What are the advantages of ontology representation?

Four mark questions:

1. Describe the Mathematical basis model in IR.
2. Write about QA system architecture.
3. Describe the Working of information extraction.
4. What are the components of ontology?

Eight mark questions:

1. What are the design features of information retrieval and its impact.
2. Explain in detail the steps involved in creation of ontologies with examples.

Unit summary

Having completed this unit, you should be able to:

- Understand what is information retrieval and the concepts
- Learn about work with the steps in IR and perform IR
- Gain knowledge on information answering, the various types of QA, how to model a QA
- Understand the concepts of information extraction, basic ideas and operations in IE
- Learn about what is ontology construction, the types, categories and steps involved in OC