FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Rethinking Scholarly Communication with AI:

Keywords Are Not the Answer To Your Search

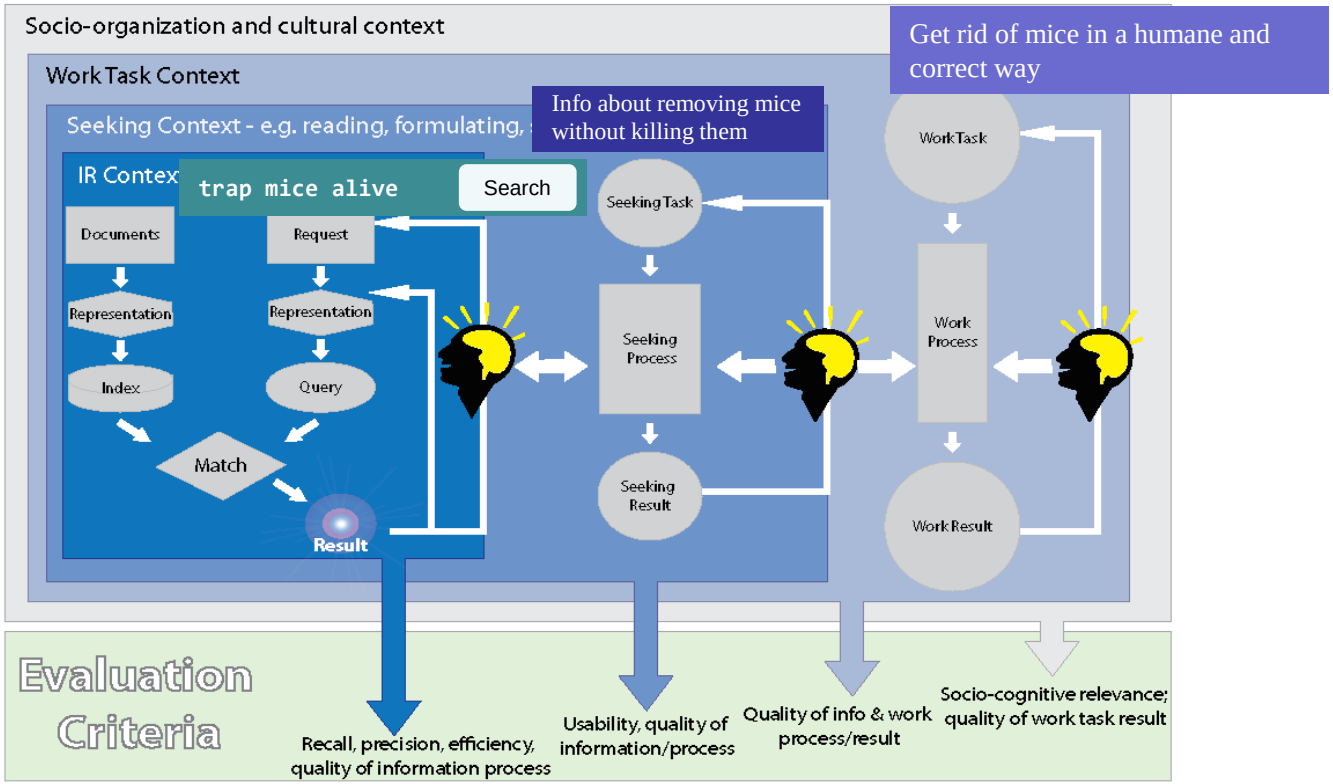
2019-06-21

Linda Andersson

OAI 11 – The CERN-UNIGE
Workshop on Innovations in Scholarly Communication

Behind the “Search” Button

The good exterminator & Co.



Modification and slide created by:
Dr. Mihai Lupu | Studio Director | Research Studio Data Science

Model first presented in:
Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer-Verlag, Berlin, Heidelberg.

Evolution of Indexing Technologies

1950 1960 1970 1980 1990 2000 2010 2020

pure counting
term frequency
position in sentence

and others

SMART
IDF
cosine similarity
and many more

Latent Semantic Analysis

WWW appears

Semantic Web appears

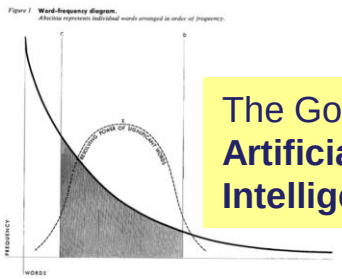
Random Indexing

Deep Learning
Speech
Vision
NLP
IR

WIKIPEDIA
The Free Encyclopedia

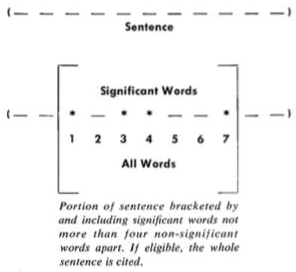
Probabilistic models for IR

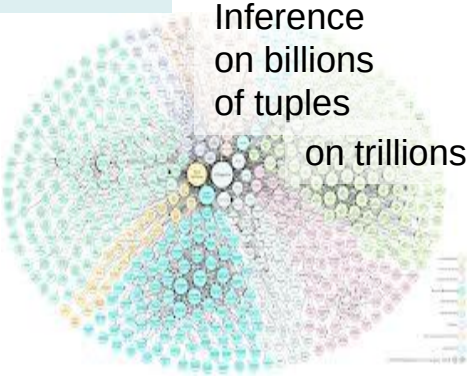
Language Models



The Golden Age of **Artificial Intelligence**

Expert Systems, Knowledge bases (e.g. Cyc)





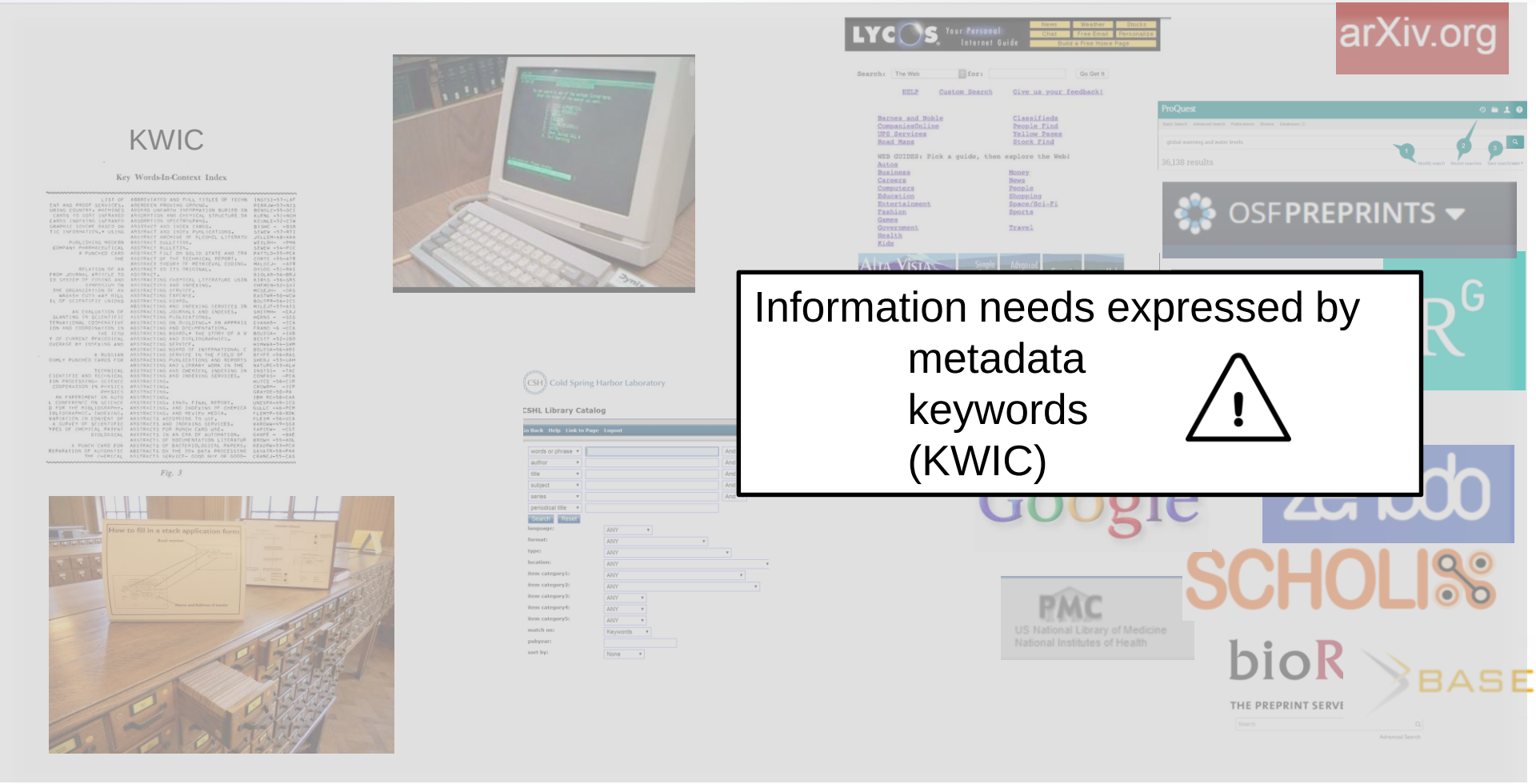
Inference on billions of tuples

on trillions

Slides courtesy to Dr. Mihai Lupu | Studio Director | Research Studio Data Science

Evolution of the Searching Functionalities

1950 1960 1970 1980 1990 2000 2010 2020

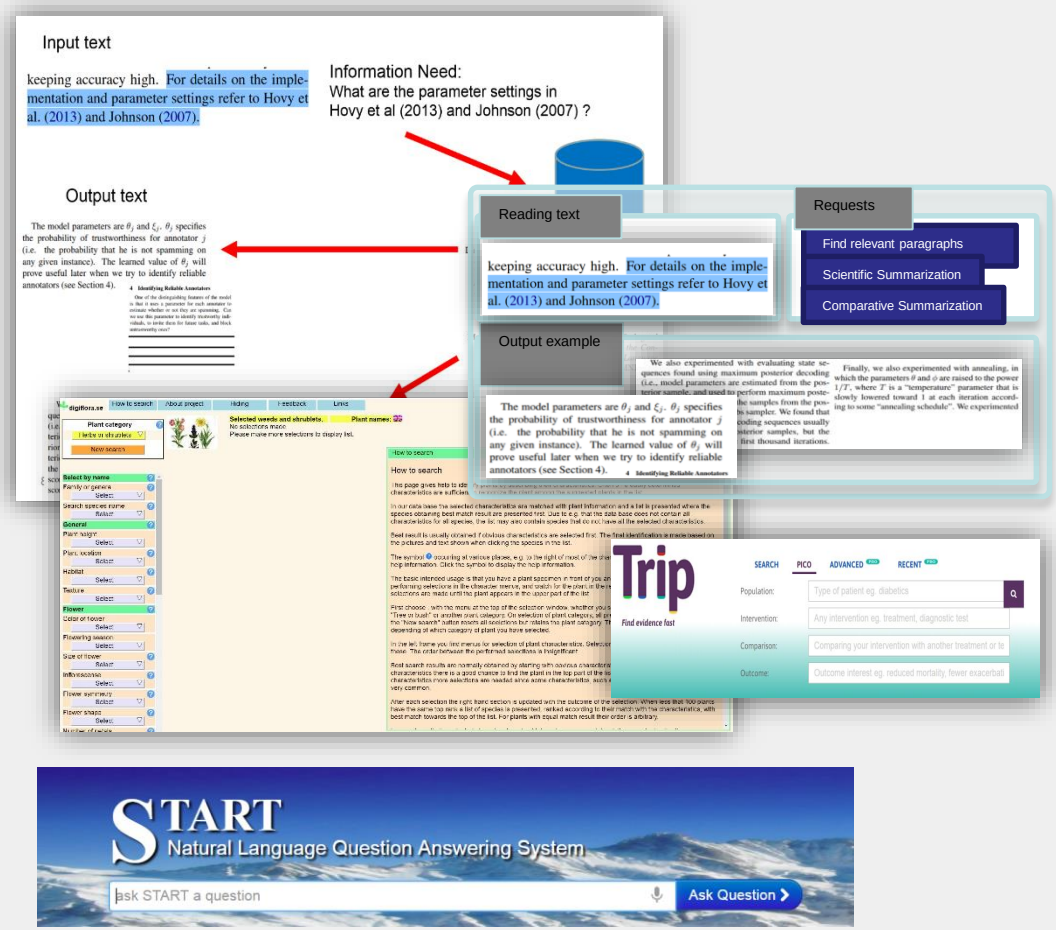


Information needs expressed by
metadata
keywords
(KWIC)

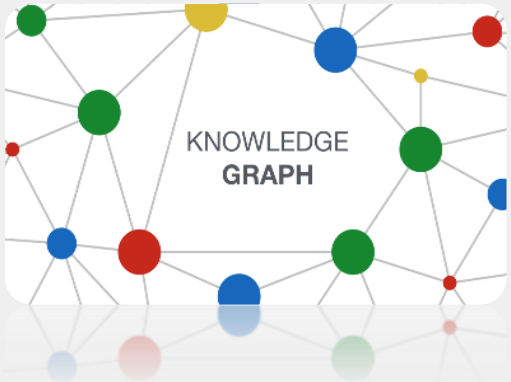
“Information Needs” Potpourri of Various Needs

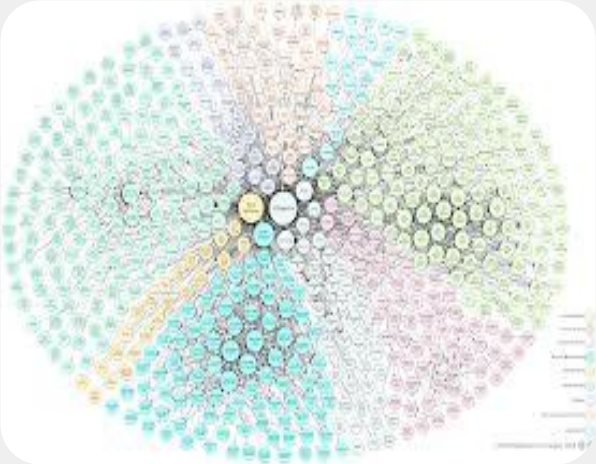
therefore Identify & Categorize

Known item search

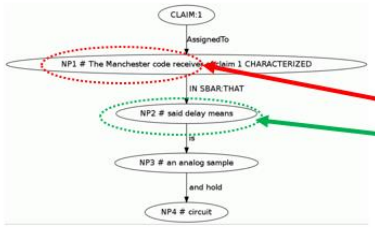


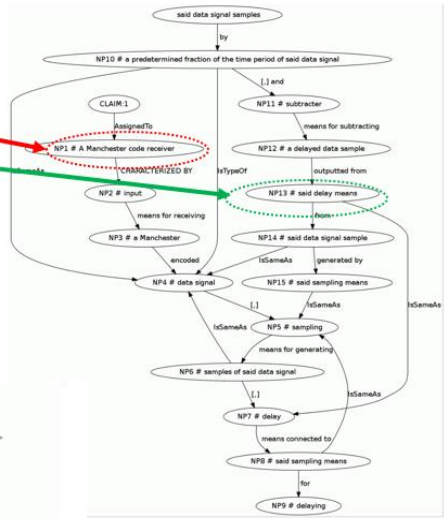
Explorative search

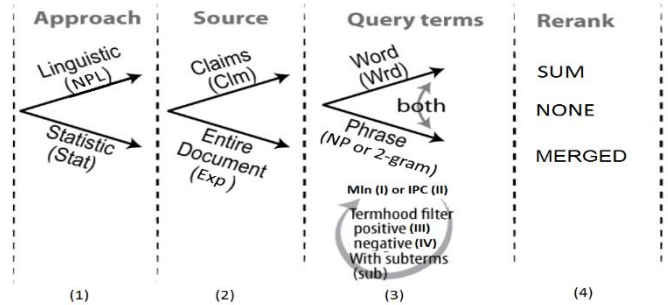


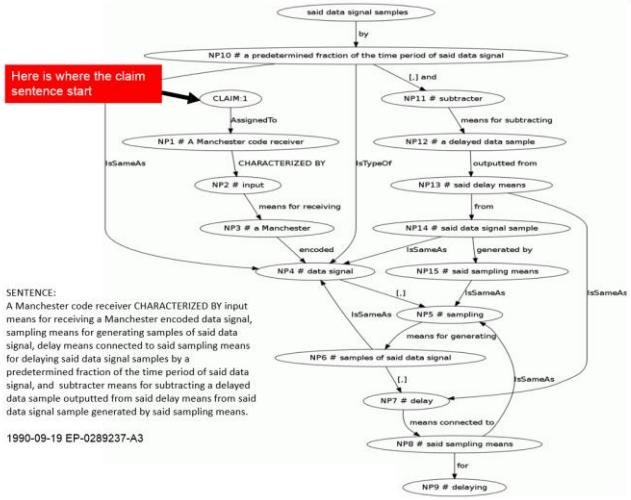


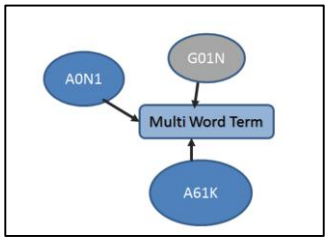
Use Case: Patent Text Mining

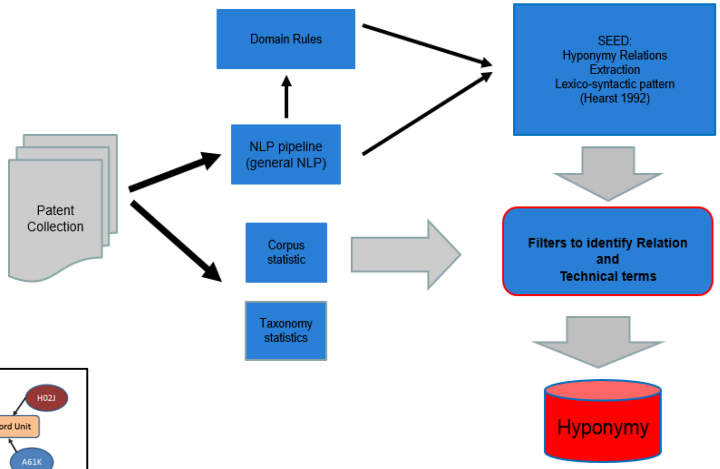


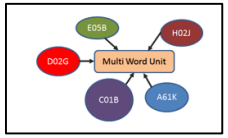












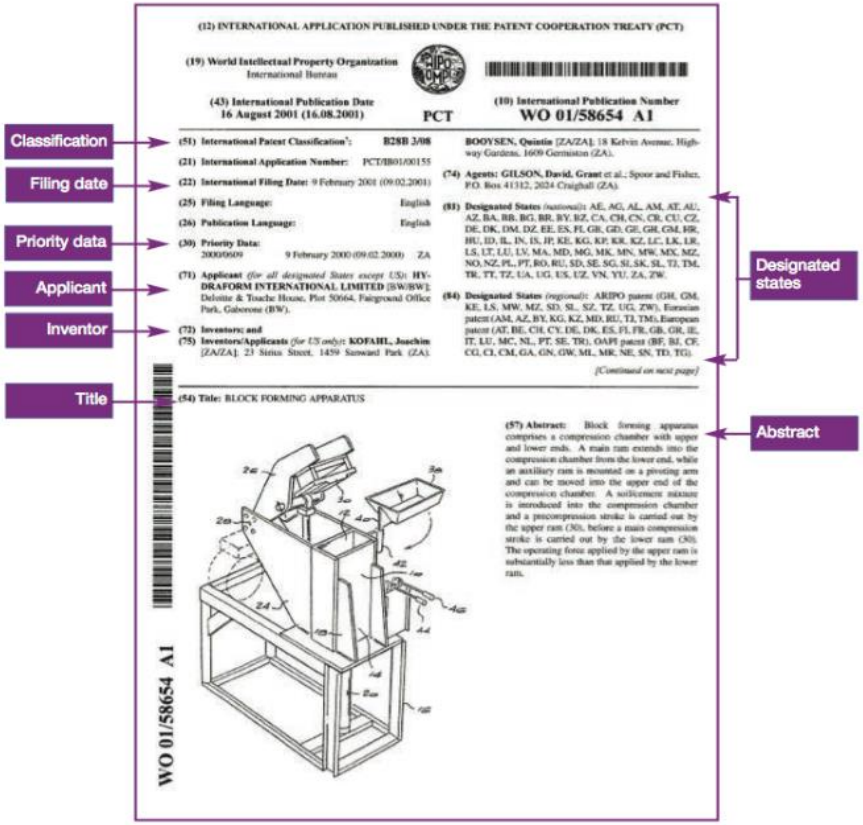
$$JoinedSimilarity = \sum_{\substack{i,j=1,n \\ i \neq j \\ i<j}}^{N} \frac{\cos\left(\vec{w_i} \quad , \quad \vec{w_j}\right)}{N}$$

SELF OPTIMIZER

FACULTY OF !INFORMATICS

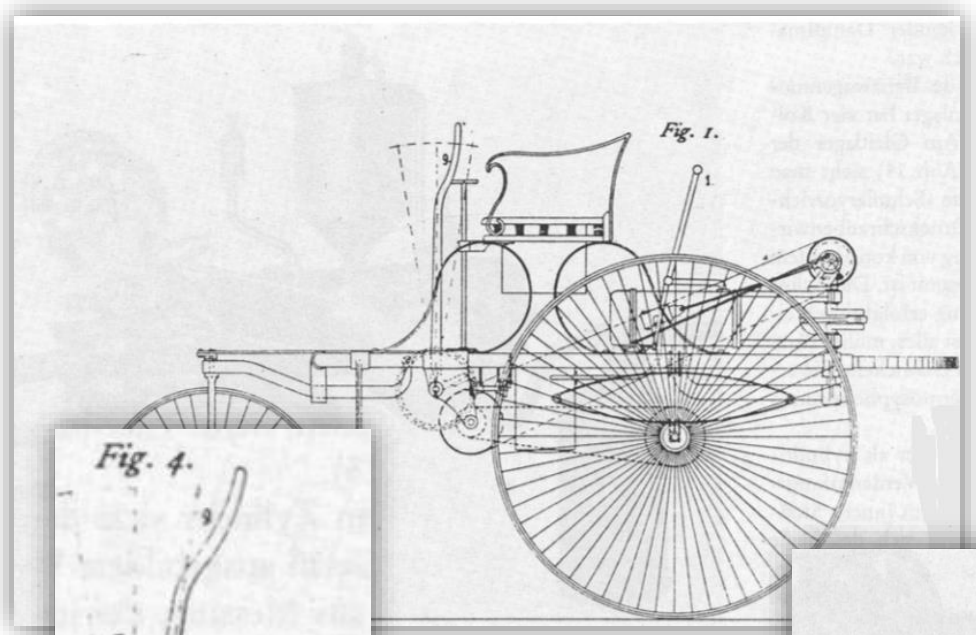
Patent Documents vs. Scientific Publications

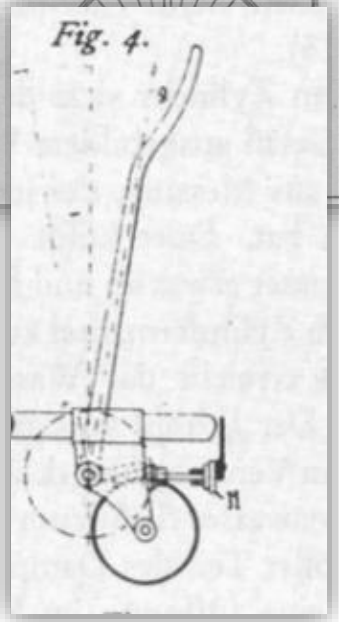
- Meta data
  - Bibliographic data
  - Classification Schema
- Text Discourse
  - Section Title, description and Claim
  - Multilingual
  - Technical terms
  - heterogeneous distribution
  - ***Patentese***
- Information Need
  - Prior art since mid 1900-centuray
  - Information extraction
    - **Paragraphs, and semantic relations**



https://library.epfl.ch/en/collections-en/patents-en/

The *Information Need* is Complex



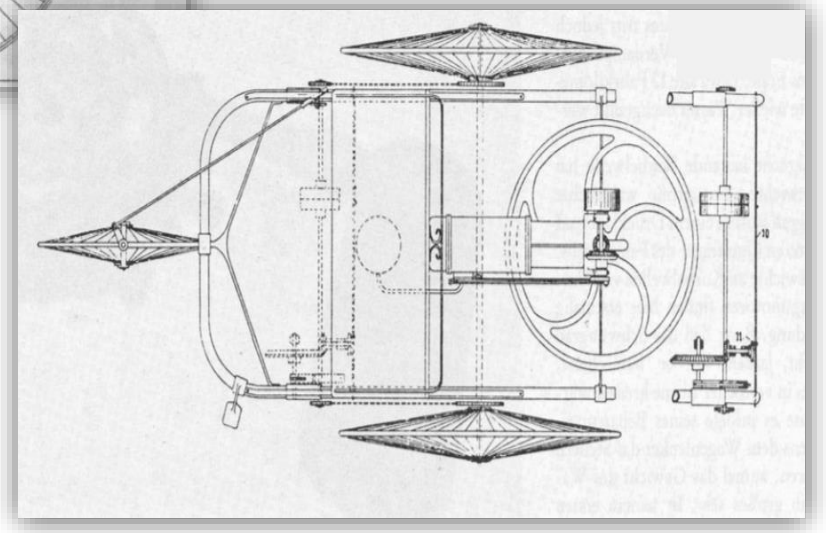


Steering mechanism

Searching for:

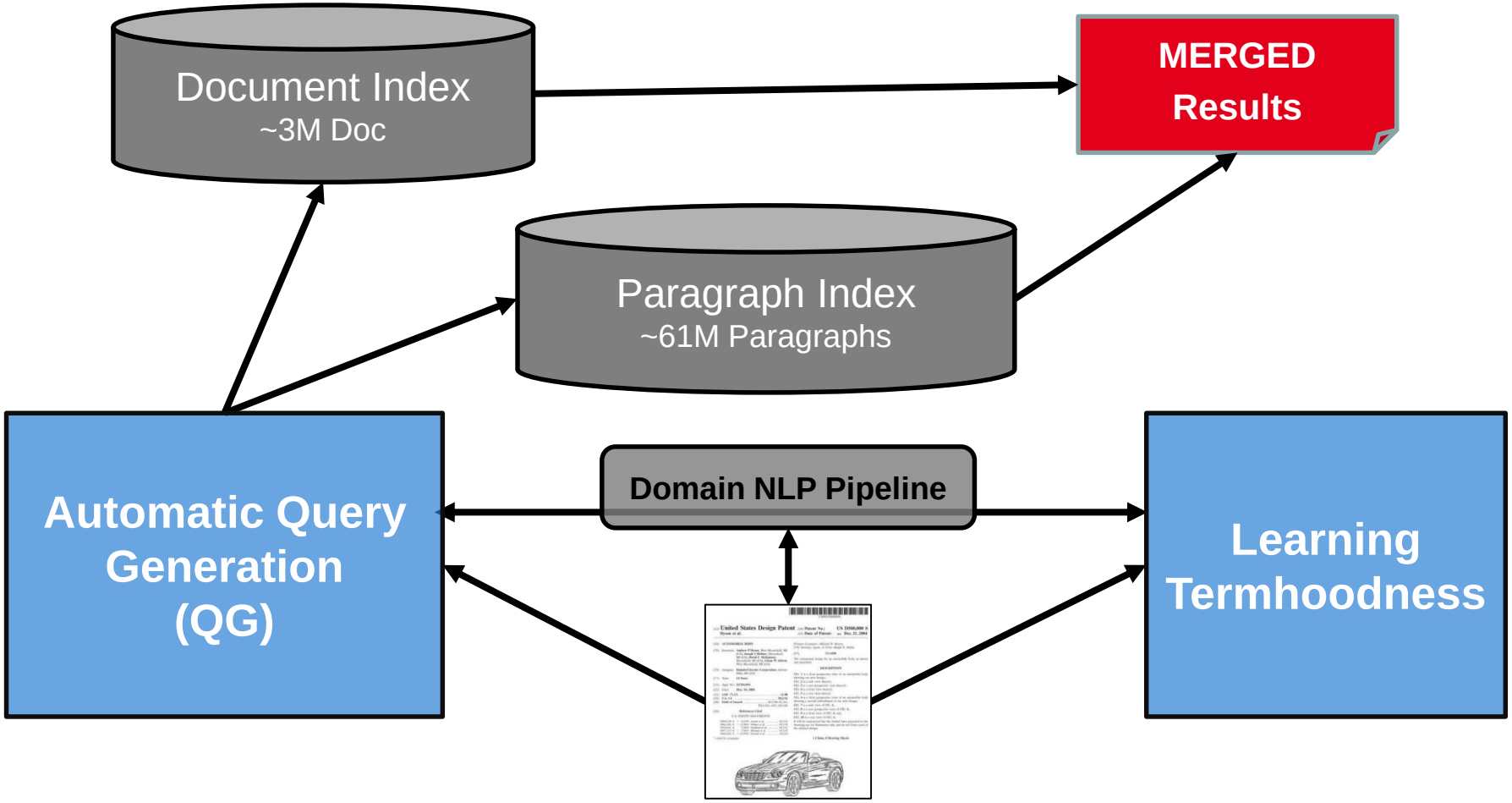
- Problem and Solution.
- Scope of the invention
- Specific technical details

Engine function



No 37435 Benz Patent – Moterwagen 1886

(Patent) Paragraph Search

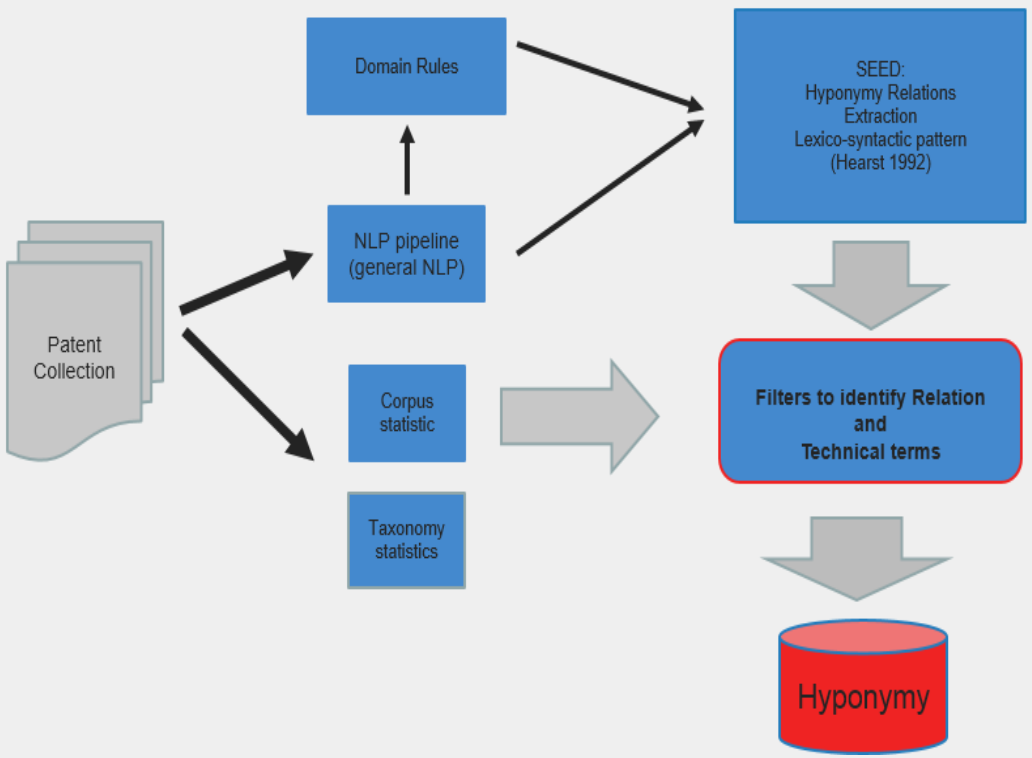


50 Topics, each topic QG 220

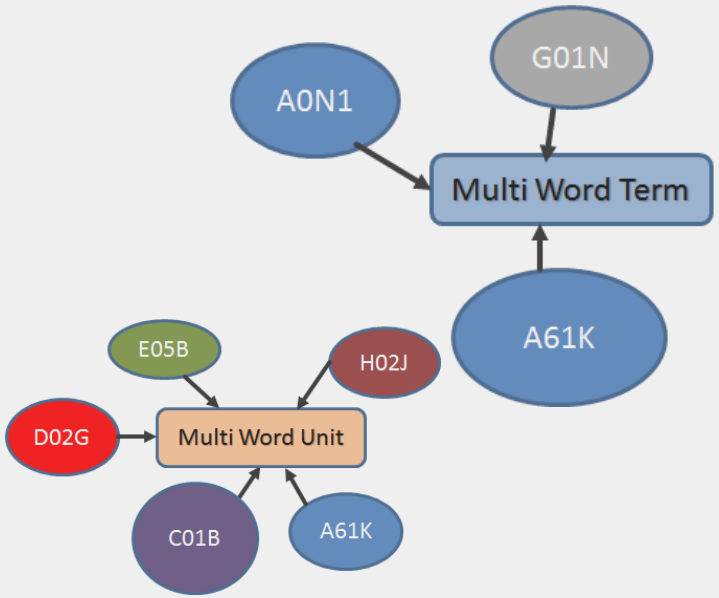
CLEF-IP 2013 Collection

Learning Termhoodness & Query Expansion

Ontology population



Terminology identification



IPC-distributional-values
Using **Machine Learning** models

Combining NLP & Distributional Semantics

Distributional Semantics identifies similarities between different words

- "Underwear" **similar to** "underpants", "undergarment", "panties", "underclothes"
- "Strength" **similar to** "strengths", "strength", "toughness", "stronger", "sfrength"

(Rekabsaz et al 2016 and Rekabsaz et al 2017)

But technical semantic relations are a mixture of single words and phrases

$$JoinedSimilarity = \sum_{\substack{i,j=1,n \\ i \neq j \\ i<j}}^{N} \frac{\cos\left(\vec{w_i}, \vec{w_j}\right)}{N}$$

- w_i, w_j represent each word vector pair cosine similarity of a MWT
- N is the number of words for a MWT

- *synthetic fibers* **synonym to** *polyester fibers*
- *thrips* **hypernym to** *bulb fly larvae*

Example of Automatic Query Formulation

Automatic query expansion terms from ontologies

brake pedal:
vehicle operating pedal,
conventional hydraulic brake system
pedal devices
position brake actuating member
brake actuating member
hydraulically-assisted rack pinion steering gear
brake operating member
conventional braking system
pair pedals

accelerator pedal
case pedal device
pedal device

US0ID500000S

(12) **United States Design Patent** (10) **Patent No.: US D500,000 S**
Dyson et al. (45) **Date of Patent: ** Dec. 21, 2004**

(54) **AUTOMOBILE BODY**

(75) Inventor: **Andrew P Dyson**, West Bloomfield, MI (US); **Joseph S Debner**, Bloomfield, MI (US); **David C McKinnon**, Bloomfield, MI (US); **Glenn W Abbott**, West Bloomfield, MI (US)

(73) Assignee: **DaimlerChrysler Corporation**, Auburn Hills, MI (US)

(**) Term: **14 Years**

(21) Appl. No.: **29/201,094**

(22) Filed: **Mar. 18, 2004**

(51) **LOC (7) Cl.** **12-08**

(52) **U.S. Cl.** **D12/92**

(58) **Field of Search** D12/90–92, 86; D21/424, 433, 206/185

(56) **References Cited**

U.S. PATENT DOCUMENTS

D408,328 S * 4/1999 Ayoub et al. D12/92
D463,436 S * 11/2002 Debner et al. D12/92
D476,601 S 7/2003 Stoddard et al. D12/92
D477,253 S * 7/2003 Minami et al. D12/92
D483,696 S * 12/2003 Howell et al. D12/92

* cited by examiner

Primary Examiner—Melody N. Brown
(74) *Attorney, Agent, or Firm*—Ralph E. Smith

(57) **CLAIM**

The ornamental design for an automobile body, as shown and described.

DESCRIPTION

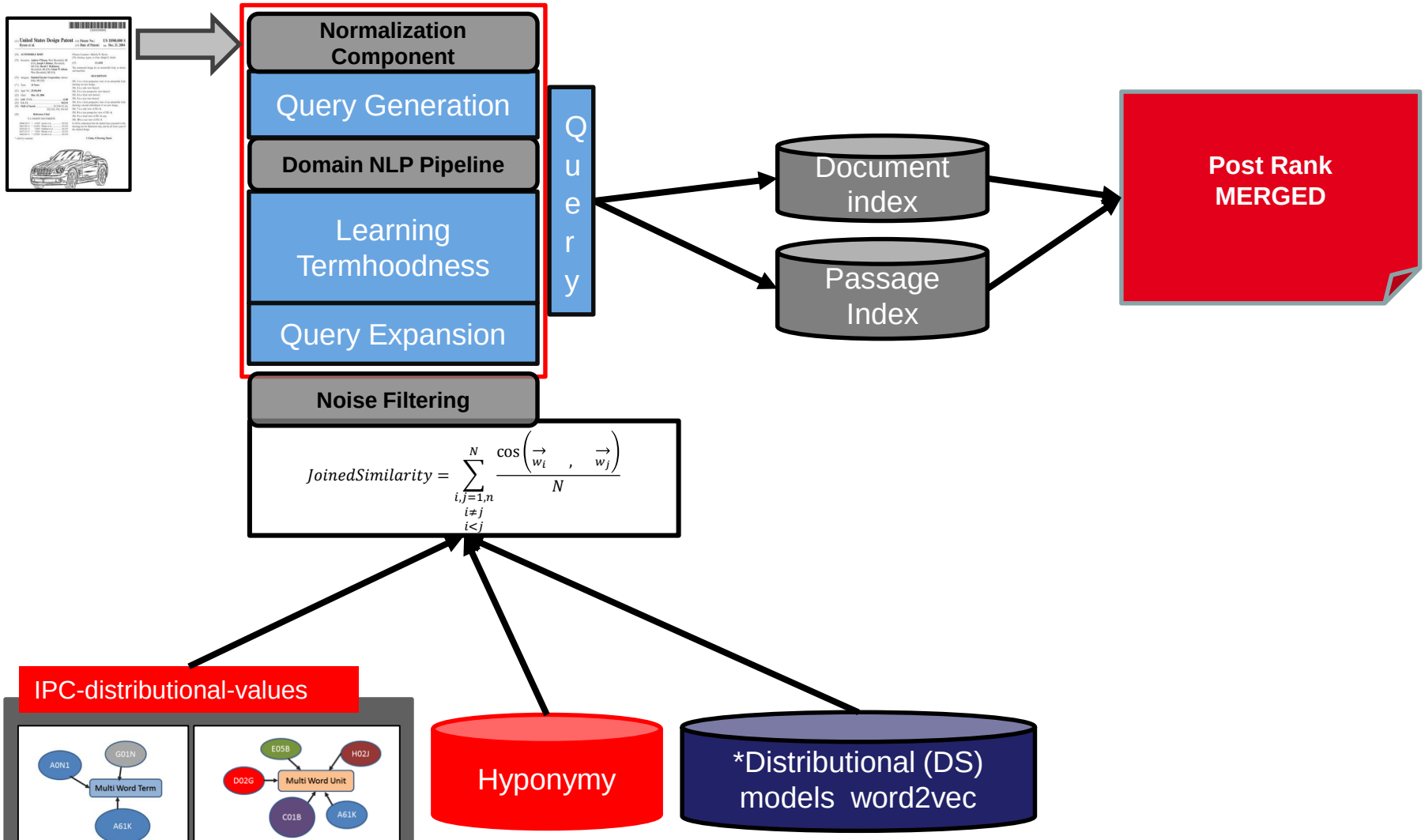
FIG. **1** is a front perspective view of an automobile body showing our new design;
FIG. **2** is a side view thereof;
FIG. **3** is a rear perspective view thereof;
FIG. **4** is a front view thereof;
FIG. **5** is a rear view thereof;
FIG. **6** is a front perspective view of an automobile body showing a second embodiment of our new design;
FIG. **7** is a side view of FIG. **6**;
FIG. **8** is a rear perspective view of FIG. **6**;
FIG. **9** is a front view of FIG. **6**; and,
FIG. **10** is a rear view of FIG. **6**.
It will be understood that the dashed lines presented in the drawings are for illustration only, and do not form a part of the claimed design.

1 Claim, 8 Drawing Sheets

Automatic Query Formulation

<QUERY>
(conured OR clutch OR connectability OR nmofs OR fclp OR dnsr OR slippage OR anda OR rotational OR acceleration OR backlash OR subordinate OR estimating OR ure OR brake OR torque OR stopped OR vehicle OR wheel OR command OR outputting OR estimate OR shock OR nsr OR driving OR pedal OR wheels OR shaft OR prohibiting OR determining OR sensor OR tws OR drive OR occurrence OR estimated OR prescribed OR stopping OR elapsed OR motor OR speed OR gdv OR instruction OR input OR output OR controller OR rotating OR accelerator OR electric OR force OR flag)
AND
("vehicle driving force control apparatus" OR "drive wheel" OR "rotational speed" OR "4wd controller" OR "clutch connection command" OR "rear wheel" OR "four-wheel drive state" OR "torque transfer path" OR "output rotational speed" OR "input rotational speed" OR "clutch control section" OR "detected parameter" OR "generation load torque" OR "torque fluctuation" OR "brake operation" OR "determination occurrence" OR "four-wheel drive vehicle" OR "input shaft" OR "4wd controller proceed" OR "wheel speed sensor" OR "output shaft" OR "response delay" OR "clutch input shaft" OR "backlash elimination" OR "drive mode switch" OR "brake pedal" OR "accelerator pedal" OR "targeted range" OR "transition time" OR "wheel speed" OR "rotational speed difference" OR "clutch connection" OR "motor torque" OR "generator load torque" OR "vehicle driving force control" OR "high rate" OR "electric motor" OR "throttle opening" OR "external disturbance" OR "vehicle driving force" OR "connected state" OR "previous equation" OR "prescribed range" OR "electric power" OR "prescribed rotational speed difference" OR "12-volt battery" OR "connection command" OR "disconnected state" OR "electric clutch" OR "four-wheel drive")
</QUERY>

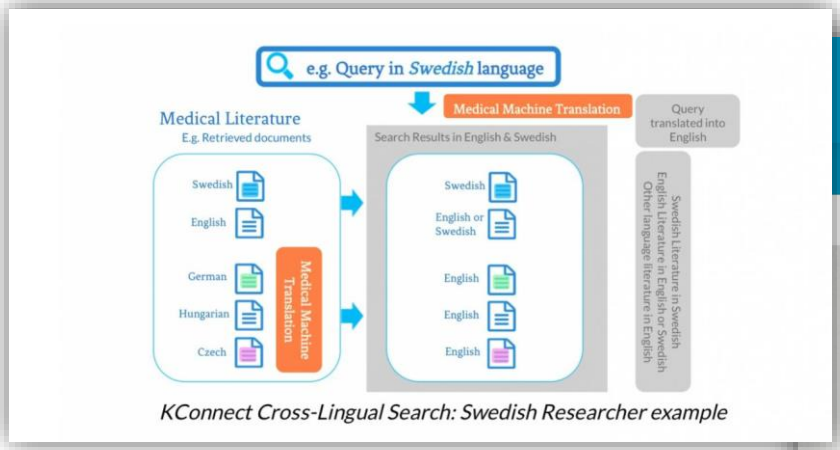
Domain Knowledge Makes AI *Smart*



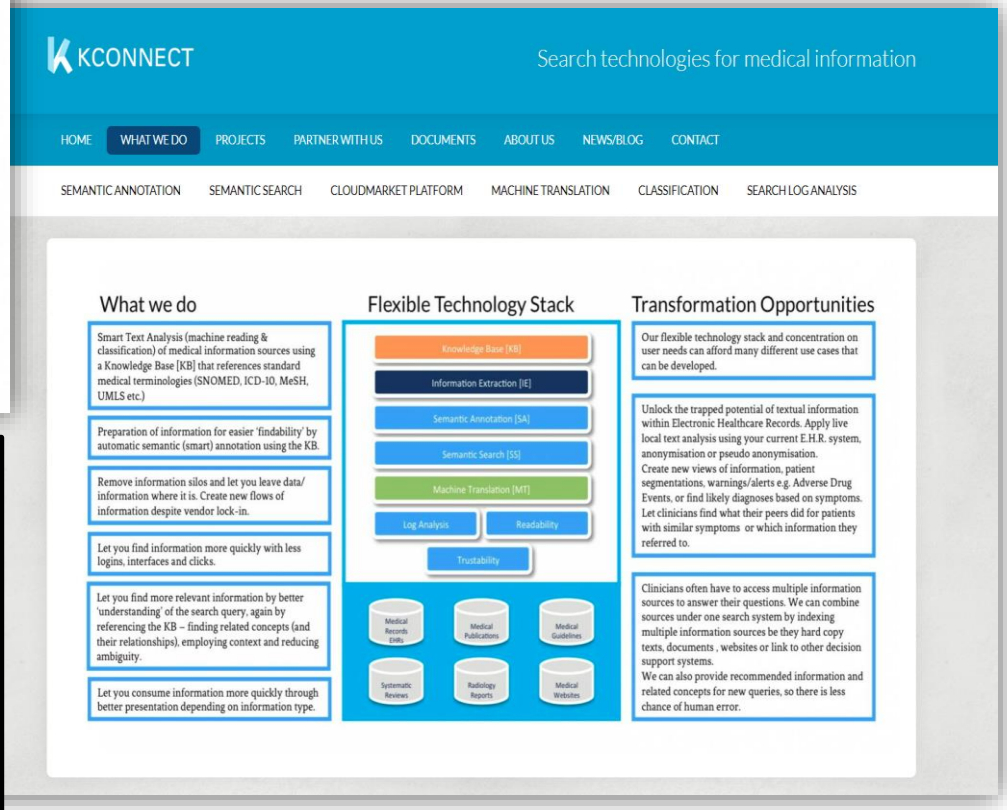
(*Rekabsaz et al 2016 and Rekabsaz et al 2017)

Use Case: Medical Information Need

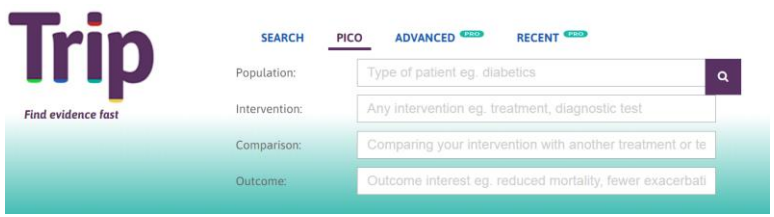
Multi-lingual retrieval



Infrastructure for medical information retrieval



Asking clinical questions using PICO entities



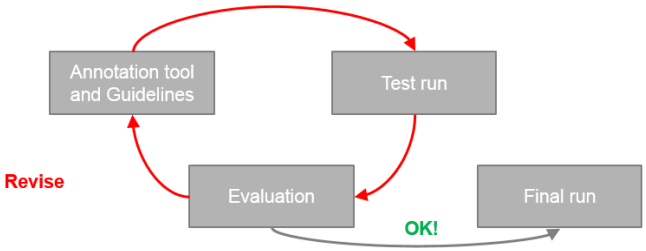
Clinical Question: PICO

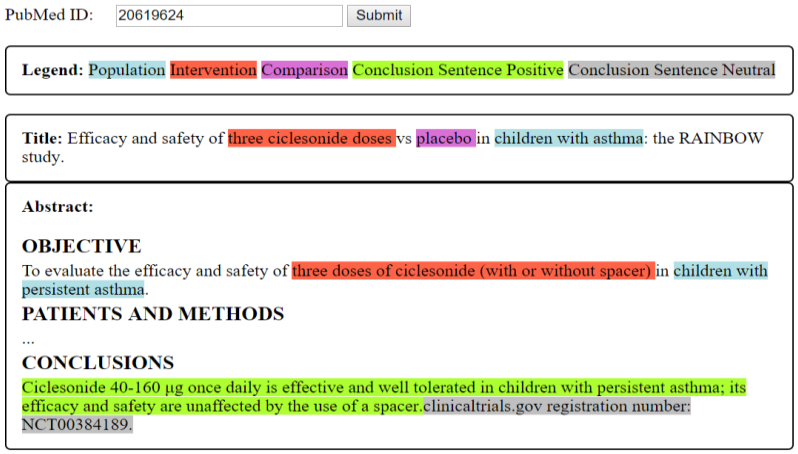
Clinical Questions	**I would like to know if Vitamin E together with aspirin is an efficient preventive treatment regimen in patients with transient ischemic attacks**		
Text Example	Vitamin E plus aspirin compared with aspirin alone in patients with transient ischemic attacks.		
Population	**Intervention**	**Comparison**	**Outcome**
patients with transient ischemic attacks	Vitamin E plus aspirin	aspirin	*vitamin E and aspirin significantly enhances the efficacy of the preventive treatment regimen*

Design Challenges

- **Complex text discourse**
  - Linguistic characteristics
  - Very complex noun phrases with post modifier clauses

- **7 Months to Beta Prototype**
  - Developing team: 1 computational linguist, 1 software engineer
- Including 4 months for annotation
  - 4 domain experts (40h each)
  - 2 Linguistic students (100h each)





Together Markus Zlabinger TU WIEN (Zlabinger et al 2018a-b)

- The Linguists identified the complex noun phrases, while the domain expert were better in identifying the population without person entity.

Learn Your Trade & Listen to Users

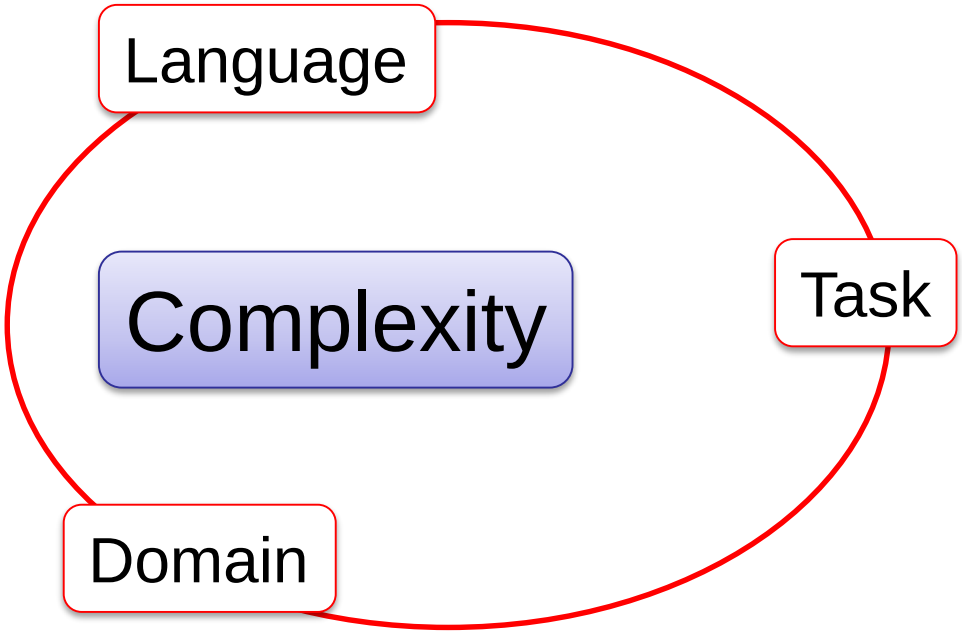
A successful text mining solution does not only focus on developing the technology or the best AI algorithm, it is much more complexed.

Keywords Are Not the Answer to Your Search

User acceptance

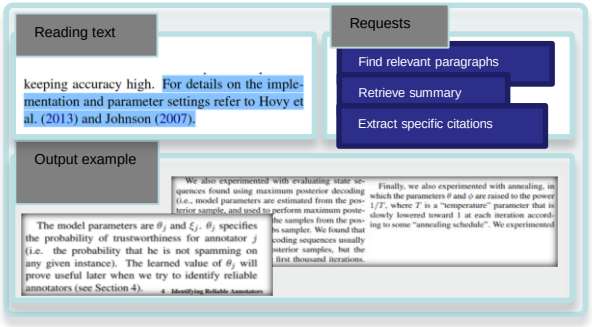
Identification

Categorisation



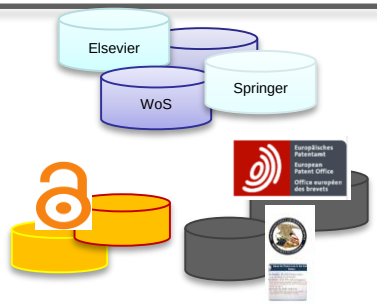
Rethinking Scholarly Communication with AI

Artificial Researcher in Science



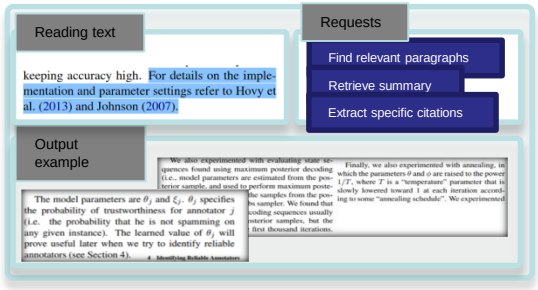
Artificial Researcher

TU WIEN Bibliothek
Your knowledge hub



vienna business agency
A service offered by the City of Vienna

Artificial Researcher in Open Access



Artificial Researcher

Bundesministerium Verkehr, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

austria wirtschafts service aws

FACULTY OF !NFORMATICS

Contact Information

PhD Linda Andersson
andersson@ifs.tuwien.ac.at
+43 (0) 699 17 821 600
http://www.ifs.tuwien.ac.at/~andersson/

Artificial Researcher

Your Scientific Knowledge Provider

linkedin.com/company/artificialresearcher
www.artificialresearcher.com

Artificial Researcher IT GmbH | Vienna - Austria

Linda Andersson - Résumé

Academic merits

- Information design 1998-2001

- General Linguistics 2001-2003
- Computational Linguistics 2006-2009

- Language Engineering 2004

- Library & Information Science 2004-2006

- Computer Science 2009-2019

Research fields

- Natural Language Understanding
- Natural Language Processing
- Information Extraction
- Information Retrieval

Awards

- 2018 *Commercial Viability award* by the Austrian Angel Investors Association.
- 2017 PhD *high potential R&D* award, i2c Award, Vienna University of Technology
- 1999 *Finalist in a Venture Cup* held at Mälardalens University College.
- *Recognized Reviewer Award* at 41st European Conference on Information Retrieval, ECIR2019 Cologne

References

Experiments in the Presentation

- **Linda Andersson, Mihai Lupu, João R. M. Palotti, Allan Hanbury, Andreas Rauber (2016) When is the time Ripe for Natural Language Processing for Patent Passage Retrieval? In Proceedings of the 25th ACM International Conference on Conference on Information and Knowledge Management (CIKM 2016)**

- **Linda Andersson, Mihai Lupu, João Palotti, Florina Piroi, Allan Hanbury, Andreas Rauber (2014) Insight to Hyponymy Lexical Relation Extraction in the Patent Genre Versus Other Text Genres In Proceedings of the First International Workshop on Patent Mining and Its Applications (IPaMin 2014) co-located with Konvens Hildesheim Germany, October 6-7 2014**

- **Linda Andersson, Mihai Lupu, Allan Hanbury (2013) Domain Adaptation of General Natural Language Processing Tools for a Patent Claim Visualization System In Proceedings of Multidisciplinary Information Retrieval, Eds. Mihai Lupu, Evangelos Kanoulas, Fernando Loizides, Lecture Notes in Computer Science, Springer Berlin Heidelberg, (70-82)**

- **Linda Andersson, Allan Hanbury, Andreas Rauber (2017) The Portability of three type of Text Mining Techniques into the patent text genre. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, Second edition, Current Challenges in Patent Information Retrieval**

- **Linda Andersson, Navid Rekabsaz, Allan Hanbury (2017) Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information The first WiNLP Workshop co-located with with the Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver**

Other papers on patent text mining

- Linda Andersson, Helena Rastas, Andreas Rauber (2014) Post OCR Correction of Swedish Patent Text - The Difference between Reading Tongue 'Lästunga' and Security Tab 'Låstunga' In Proceedings of Multidisciplinary Information Retrieval,Eds. David Lamas, Paul Buitelaar,Lecture Notes in Computer Science, Springer Berlin Heidelberg, (1-9)

- Linda Andersson, Parvaz Mahdabi, Allan Hanbury, Andreas Rauber (2013) Exploring patent passage retrieval using nouns phrases In Proceeding of the 35th European conference on Advances in Information Retrieval (ECIR'13), Eds. Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, and Stefan Rüger, Springer-Verlag, Berlin, Heidelberg, (676-679)

- Linda Andersson (2010) A Vector Space analysis of Swedish patent claims with different linguistic indices In Proceeding of the 3rd International Workshop on Patent Information Retrieval (PaIR), 2010

References

- Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik, Jon Brassey (2018) Medical Entity Corpus with PICO Elements and Sentiment Analysis Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan

- Markus Zlabinger, Linda Andersson, Jon Brassey, Allan Hanbury (2018) Extracting the Population, Intervention, Comparison and Sentiment from Randomized Controlled TrialsBuilding Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth - Proceedings of {MIE}, Medical Informatics Europe, Göteborg, Sweden, April

- Hearst. M. A. (1992) Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th Conference on Computational Linguistics - Volume 2, COLING ’92, pages 539–545, Stroudsburg, PA, USA, Association for Computational Linguistics.

- Rekabsaz, N. Lupu M, Baklanov A., Hanbury A., Dur A, and Anderson L. (2017). Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In Proc. of ACL. Association for Computational Linguistics, Stroudsburg, PA, USA.

- Rekabsaz, N., Lupu, M., Hanbury, A., and Zuccon, G. (2016). Generalizing translation models in the probabilistic relevance framework. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, pages 711–720.