FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

SEMANTiCS
Karlsruhe 2019

ADVANCING SCIENCE

# From specific problems to a generic solution: A scalable framework for analyzing Big Data of Patent Information

Ahmad Alrifai, September 09-12, 2019

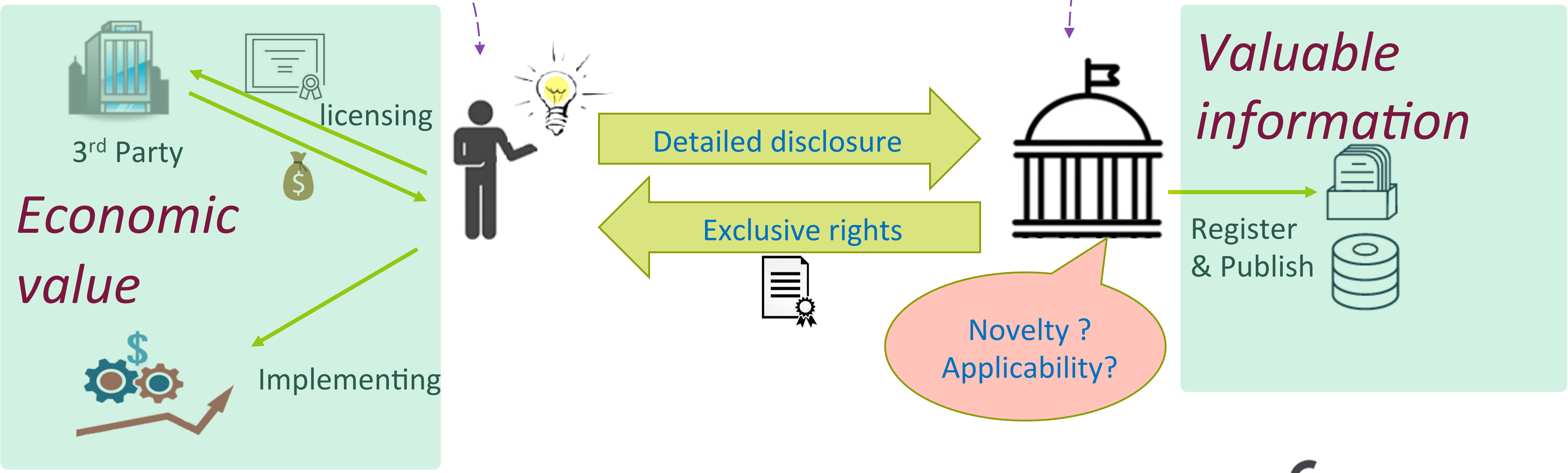
# Outline

- Patent information – An overview
- Patent search and analysis
- Text and Data mining @FIZ
- Generic scalable framework
- Ongoing and future work
- Conclusion

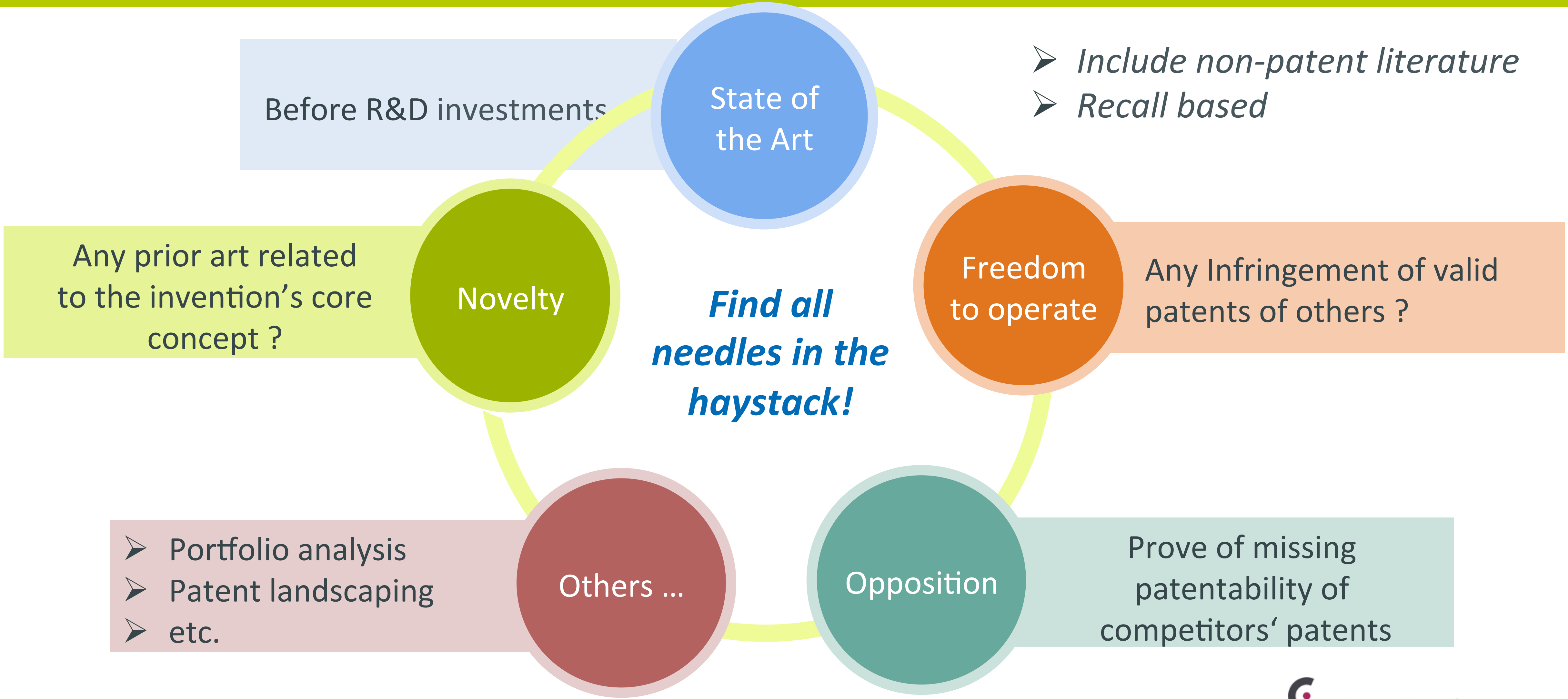
# Patent – in the full sense of the word!

*“A patent (/ˈpætənt/ or /ˈpeɪtənt/) is a set of exclusive rights granted by a sovereign state to an inventor or assignee for a limited period of time in exchange for detailed public disclosure of an invention….”*

*[source: wikipedia.org]*



# Patent search and analysis – use cases



Before R&D investments

- *Include non-patent literature*
- *Recall based*

Any prior art related to the invention’s core concept ?

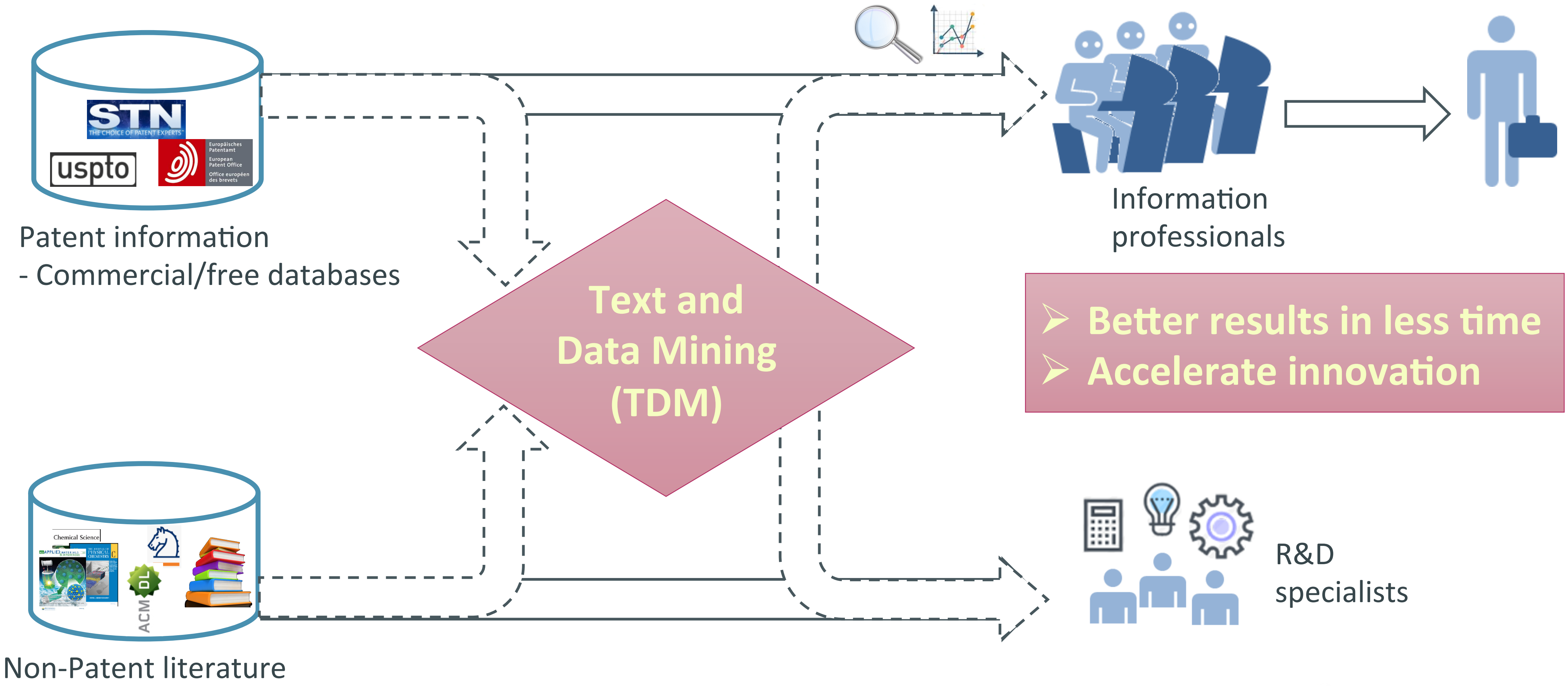
***Find all needles in the haystack!***

Any Infringement of valid patents of others ?

- Portfolio analysis
- Patent landscaping
- etc.

Prove of missing patentability of competitors‘ patents

# Patent and scientific information



Patent information
- Commercial/free databases

Non-Patent literature

Text and Data Mining (TDM)

Information professionals

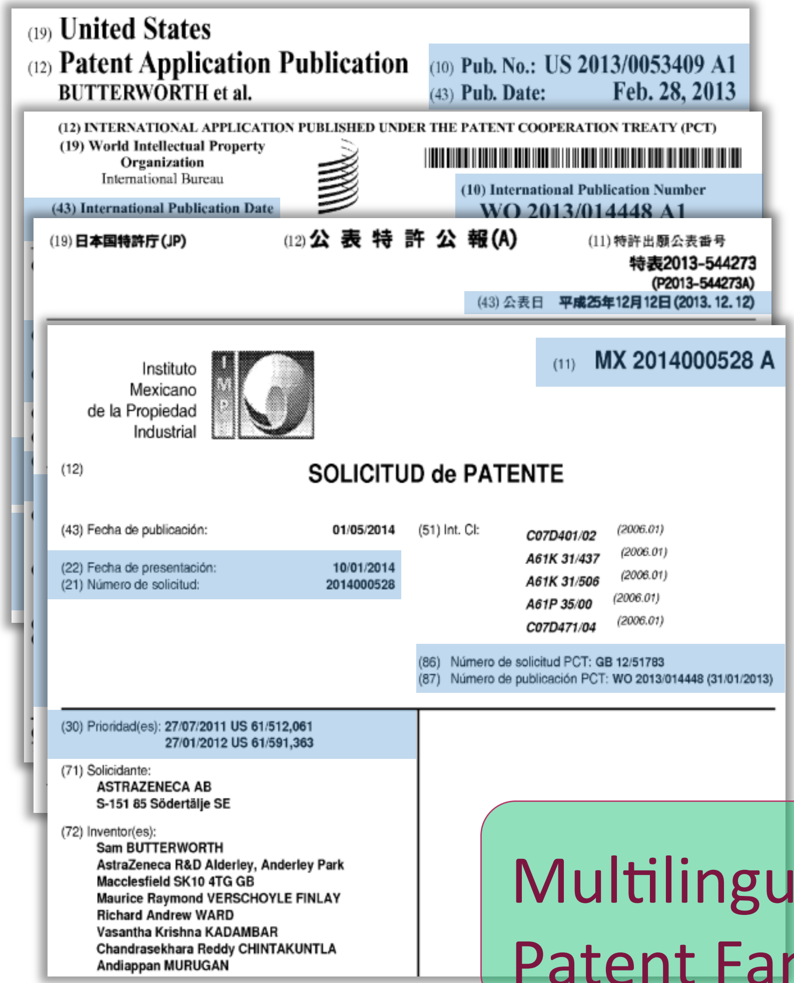
R&D specialists

➢ Better results in less time
➢ Accelerate innovation

# More than 100 Million documents

A typical chemistry patent

- *Meta-data: dates, names, classification,…*
- *Title*
- *Abstract*
- *Detailed description*
- *Claims: legal scope of the protection*



(19) United States
(12) Patent Application Publication
BUTTERWORTH et al.

(10) Pub. No.: US 2013/0053409 A1
(43) Pub. Date: Feb. 28, 2013

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)
(19) World Intellectual Property Organization
International Bureau
(43) International Publication Date
(10) International Publication Number
WO 2013/014448 A1

(19) 日本国特許庁(JP)
(12) 公表特許公報(A)
(11) 特許出願公表番号
特表2013-544273
(P2013-544273A)
(43) 公表日 平成25年12月12日(2013. 12. 12)

Instituto
Mexicano
de la Propiedad
Industrial

(11) MX 2014000528 A

(12) SOLICITUD de PATENTE

(43) Fecha de publicación: 01/05/2014
(22) Fecha de presentación: 10/01/2014
(21) Número de solicitud: 2014000528

(51) Int. Cl:
C07D401/02 (2006.01)
A61K 31/437 (2006.01)
A61K 31/506 (2006.01)
A61P 35/00 (2006.01)
C07D471/04 (2006.01)

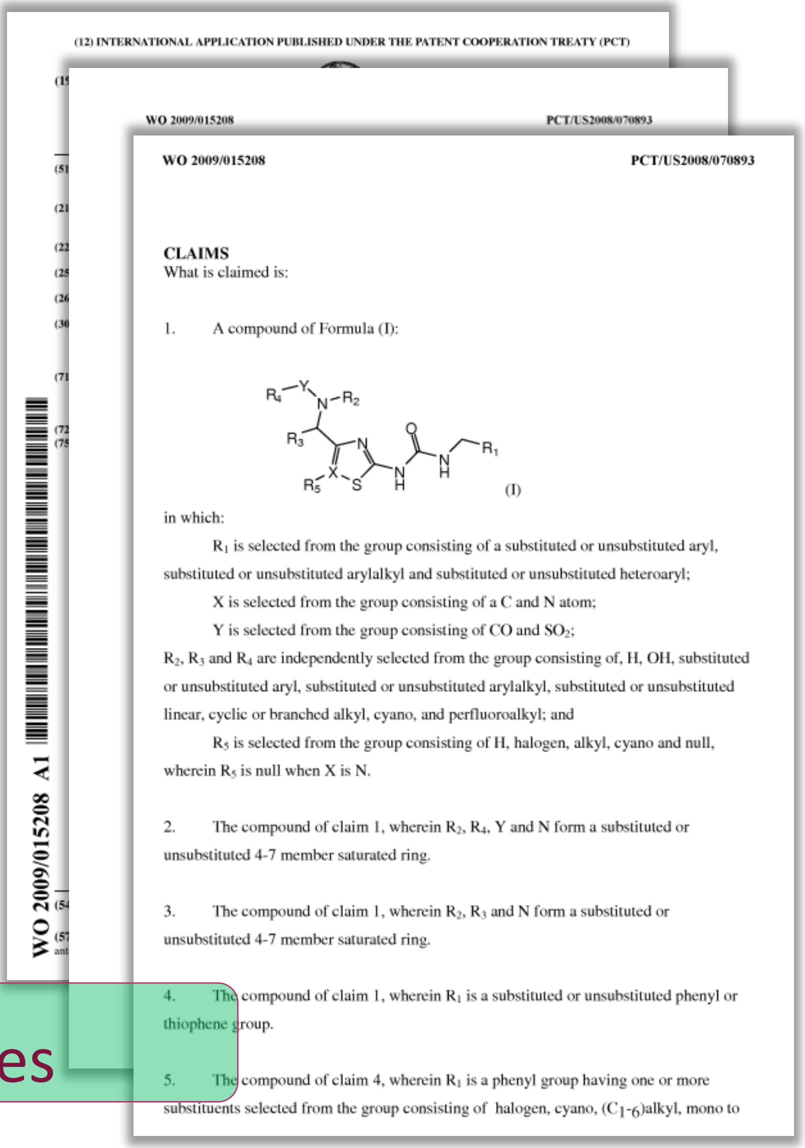
(86) Número de solicitud PCT: GB 12/51783
(87) Número de publicación PCT: WO 2013/014448 (31/01/2013)

(30) Prioridad(es): 27/07/2011 US 61/512,061
27/01/2012 US 61/591,363

(71) Solicidante:
ASTRAZENECA AB
S-151 85 Södertälje SE

(72) Inventor(es):
Sam BUTTERWORTH
AstraZeneca R&D Alderley, Anderley Park
Macclesfield SK10 4TG GB
Maurice Raymond VERSCHOYLE FINLAY
Richard Andrew WARD
Vasantha Krishna KADAMBAR
Chandrasekhara Reddy CHINTAKUNTLA
Andiappan MURUGAN

Multilingual Patent Families



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

WO 2009/015208 A1

WO 2009/015208 PCT/US2008/070893

WO 2009/015208 PCT/US2008/070893

**CLAIMS**
What is claimed is:

1. A compound of Formula (I):

(I)

in which:

$R_1$ is selected from the group consisting of a substituted or unsubstituted aryl, substituted or unsubstituted arylalkyl and substituted or unsubstituted heteroaryl;

X is selected from the group consisting of a C and N atom;

Y is selected from the group consisting of CO and $SO_2$;

$R_2$, $R_3$ and $R_4$ are independently selected from the group consisting of, H, OH, substituted or unsubstituted aryl, substituted or unsubstituted arylalkyl, substituted or unsubstituted linear, cyclic or branched alkyl, cyano, and perfluoroalkyl; and

$R_5$ is selected from the group consisting of H, halogen, alkyl, cyano and null, wherein $R_5$ is null when X is N.

2. The compound of claim 1, wherein $R_2$, $R_4$, Y and N form a substituted or unsubstituted 4-7 member saturated ring.

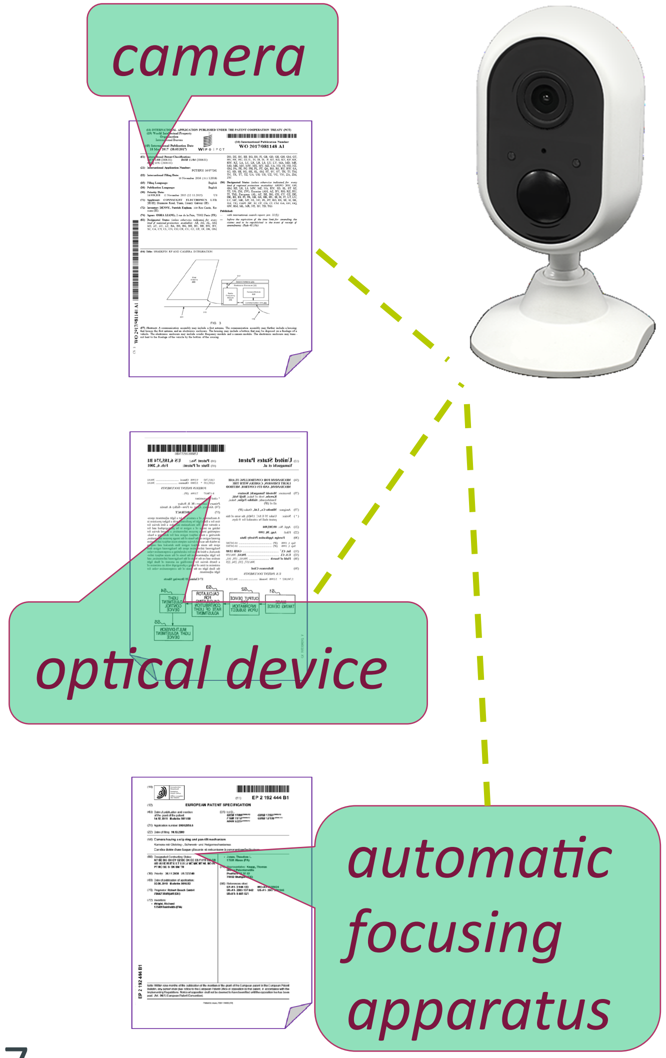
3. The compound of claim 1, wherein $R_2$, $R_3$ and N form a substituted or unsubstituted 4-7 member saturated ring.

4. The compound of claim 1, wherein $R_1$ is a substituted or unsubstituted phenyl or thiophene group.

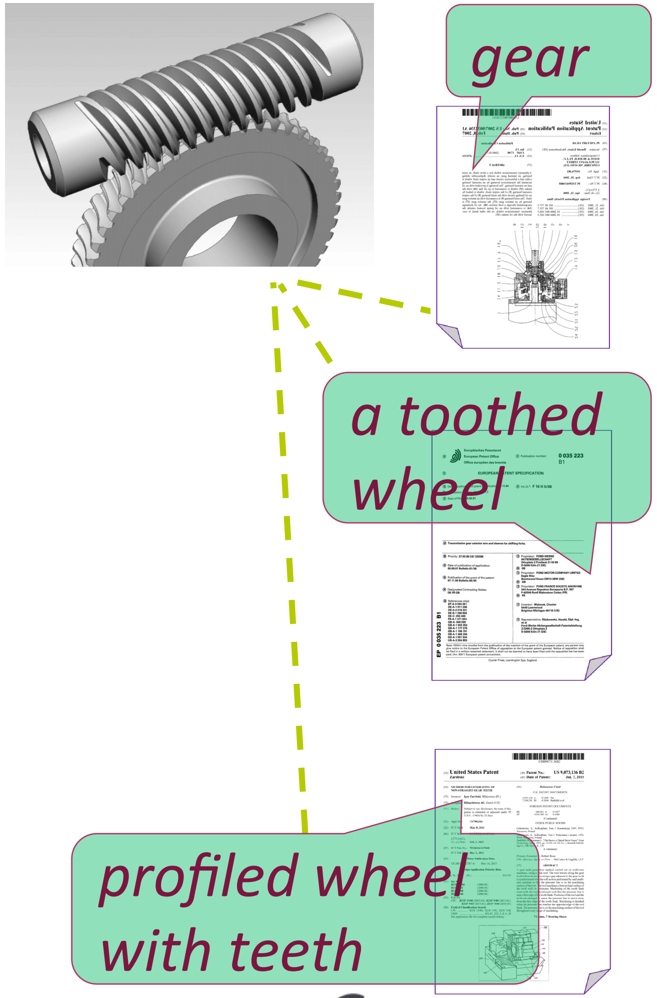
5. The compound of claim 4, wherein $R_1$ is a phenyl group having one or more substituents selected from the group consisting of halogen, cyano, $(C_{1-6})$alkyl, mono to

250 pages

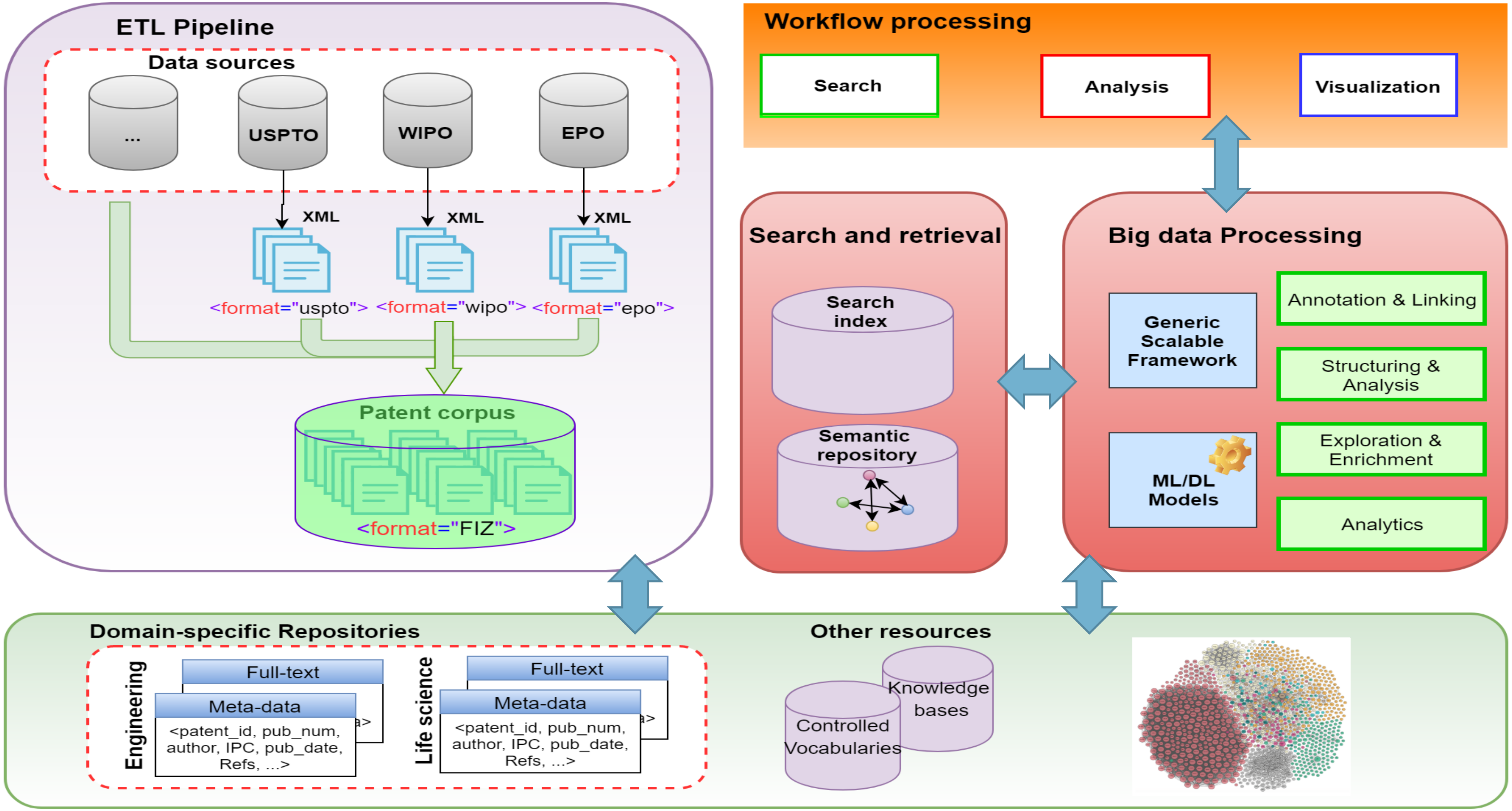
# Challenges of patent text and data mining



- Synonyms and paraphrasing
- New terminologies
- Long and complex Noun phrases
- Drawings, compounds, formulae, …
- ⋮



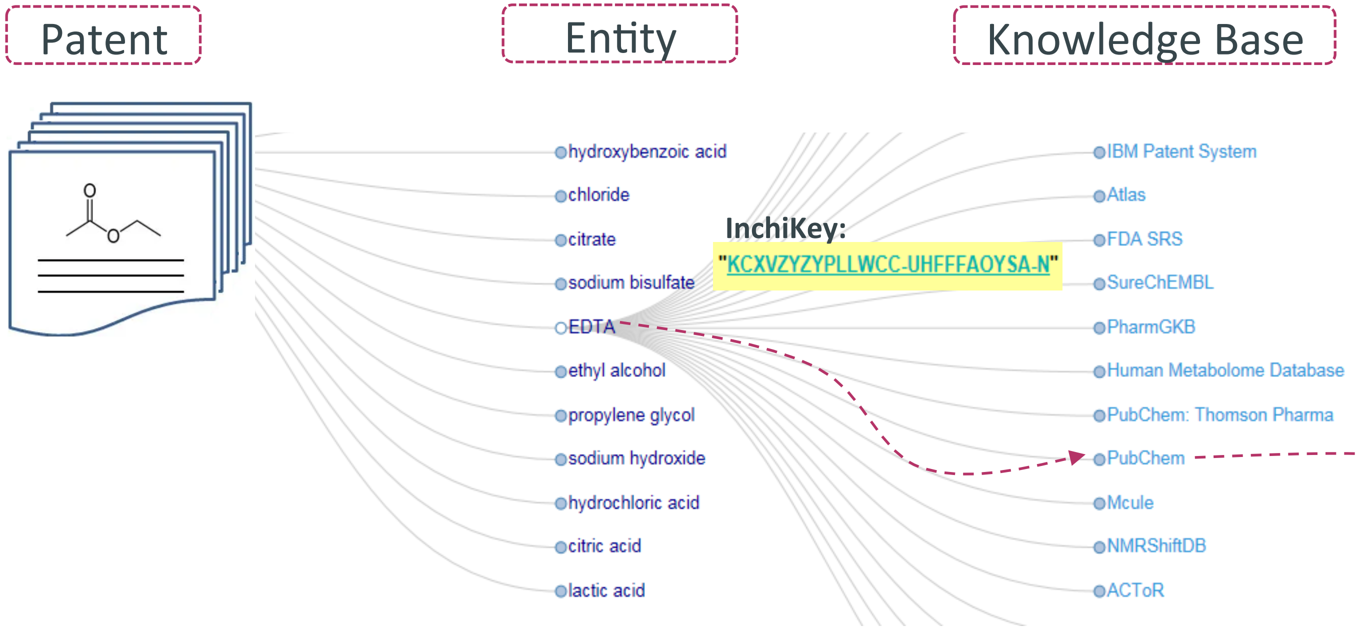
# System architecture for a powerful TDM



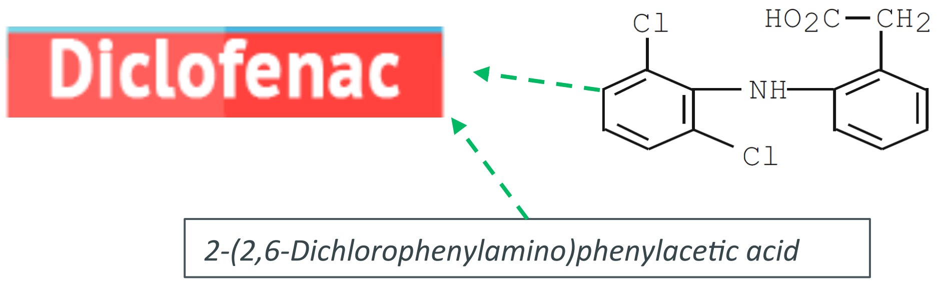
# Annotation and linking

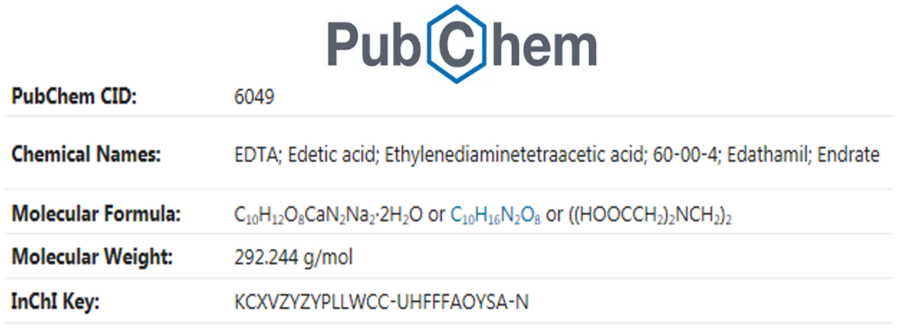
- ➢ Patent chemical search is essential for chemical and pharmaceutical domains

Annotation, linking and indexing



- ➢ Trade names, chemical names and structures





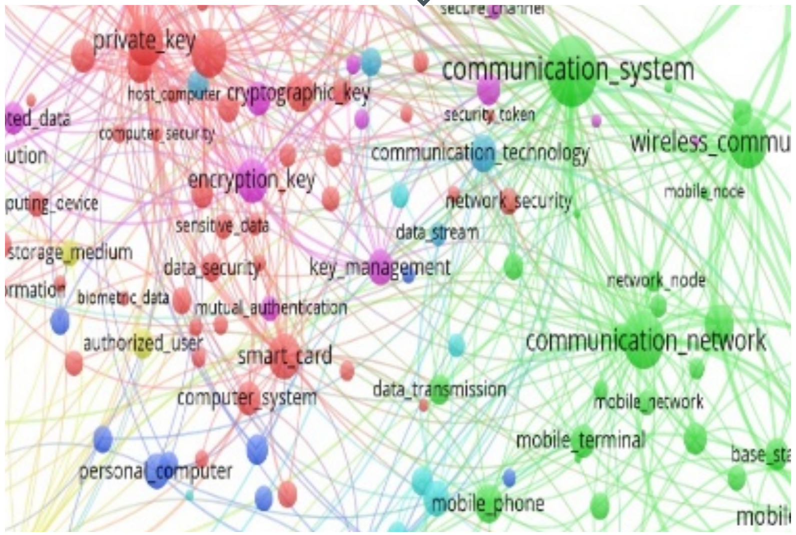
PubChem

| | |
|---|---|
| **PubChem CID:** | 6049 |
| **Chemical Names:** | EDTA; Edetic acid; Ethylenediaminetetraacetic acid; 60-00-4; Edathamil; Endrate |
| **Molecular Formula:** | $C_{10}H_{12}O_8CaN_2Na_2\cdot 2H_2O$ or $C_{10}H_{16}N_2O_8$ or $((HOOCCH_2)_2NCH_2)_2$ |
| **Molecular Weight:** | 292.244 g/mol |
| **InChI Key:** | KCXVZYZYPLLWCC-UHFFFAOYSA-N |

# Structuring and analysis

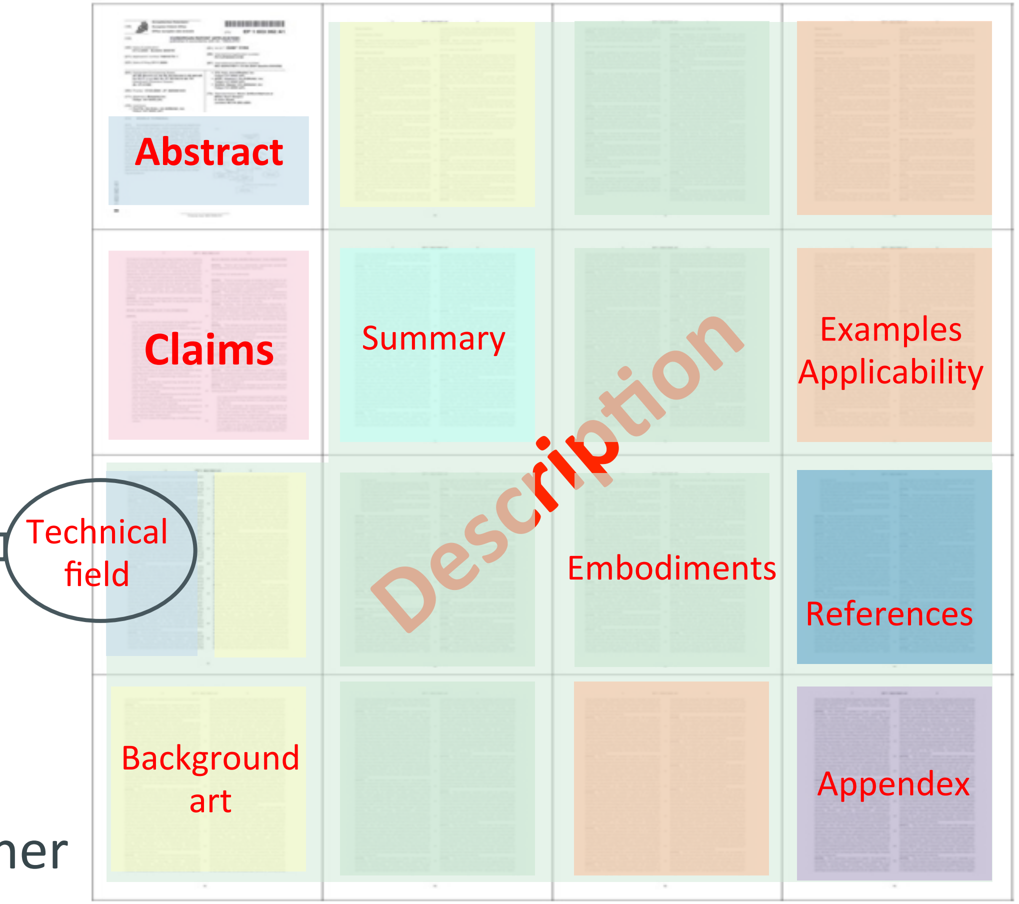
*“The invention belongs to the field of communication network, in particular relates to a domain name of the biometrics-based authentication system and method. ”*

Large-scale technical trends analysis



- Targeted search
- Input to other tasks

Segment the description part into predefined sections

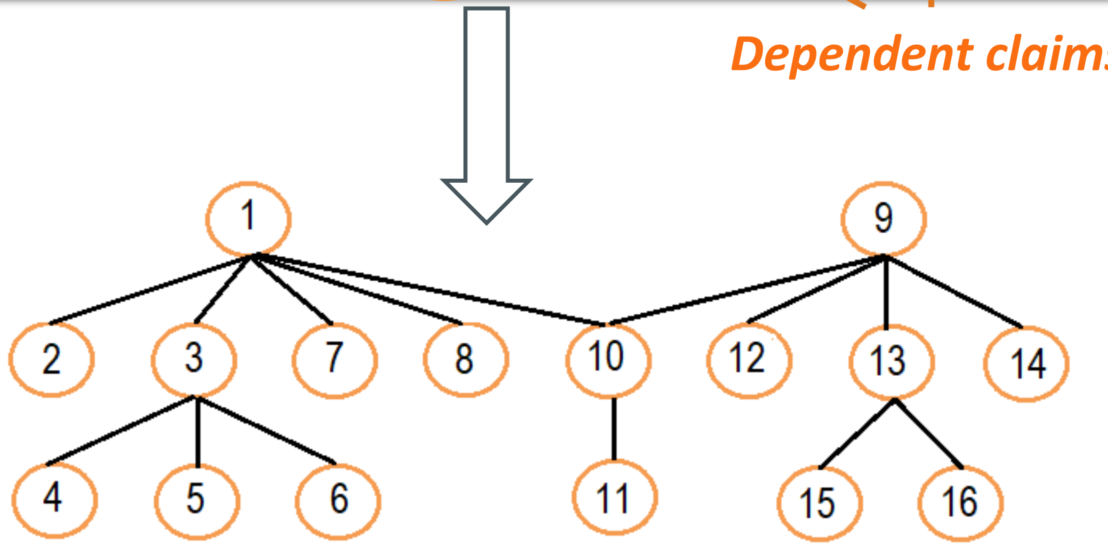


# Structuring and enrichment

## Claims Hierarchy Exploration

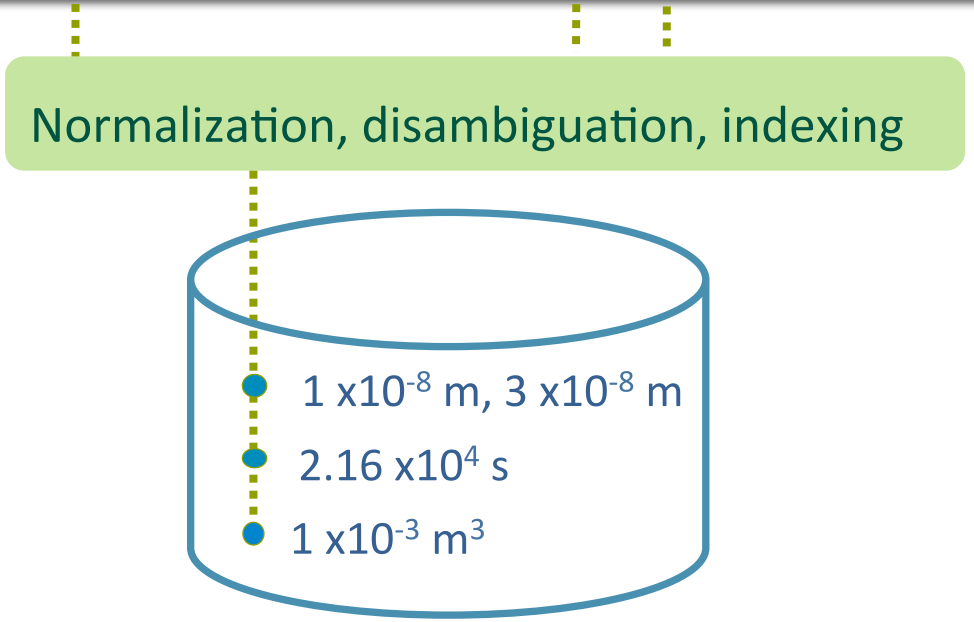
*“1. A composition comprising [...] alkenyl succinic anhydride substituted starch, [...]*
*3. The composition of claim 1 wherein the alkenyl succinic anhydride ...*
*4. The composition of claim 3 wherein ...”*

***Dependent claims***



## Analysis of numerical properties

*“The resulting CeO2 particle size measured by x-ray diffraction were in the range of 10 to 30 nm. Fig. 1 shows typical nano particles in a sample milled for 6 hours In a second experiment a 1 litre attrition...”*

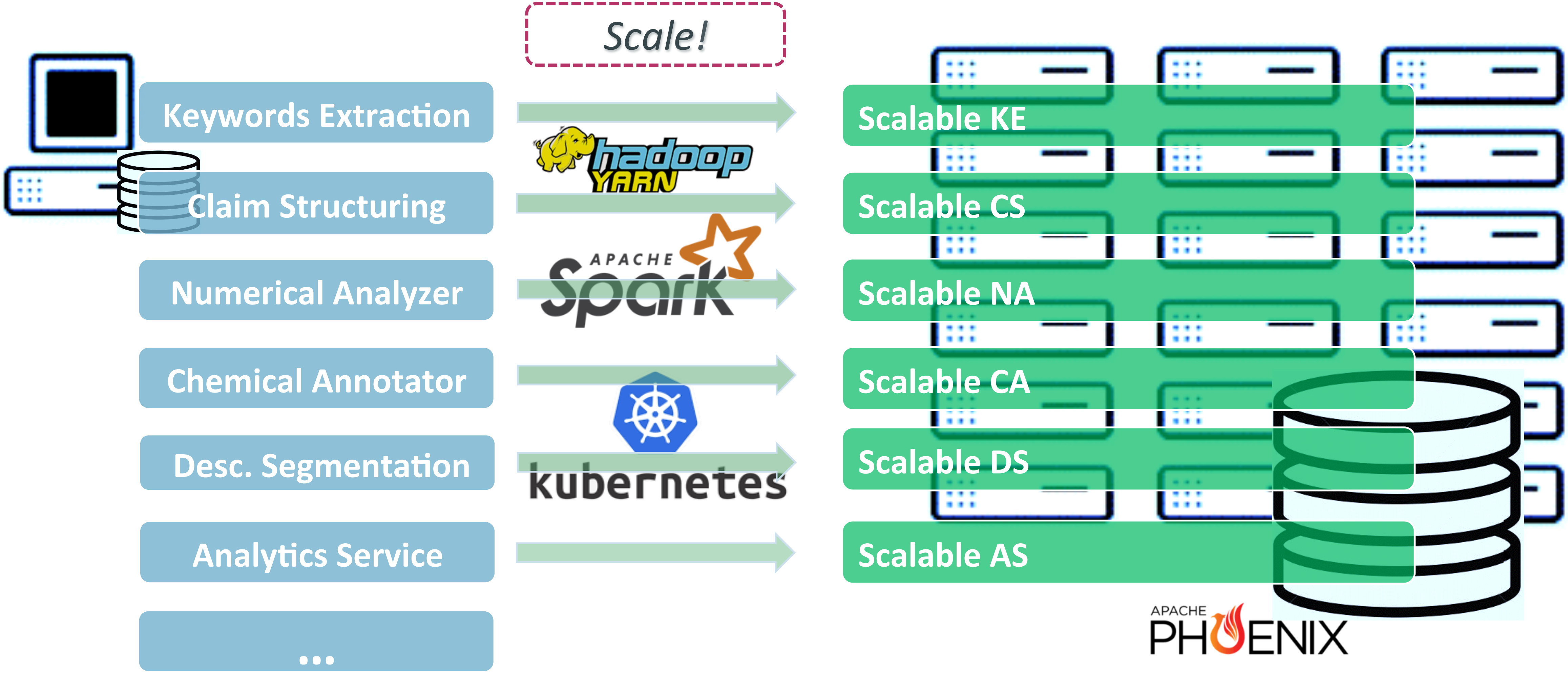


Normalization, disambiguation, indexing
$1\ x10^{-8}$ m, $3\ x10^{-8}$ m
$2.16\ x10^{4}$ s
$1\ x10^{-3}\ m^{3}$

# Need for speed

- **Performance**
  - 1s processing time per document → ~ 2 months for a database
- **Special requirements**
  - High recall is required
  - Iterative evaluations and adaptations
- **Challenges**
  - NLP tools trained on different corpora → rule-based modifications
  - Lack of in-domain training data or golden corpora
  - Typographical errors from OCR and machine translation

# It’s all about scaling



*Scale!*

Keywords Extraction

Claim Structuring

Numerical Analyzer

Chemical Annotator

Desc. Segmentation

Analytics Service

...

Scalable KE

Scalable CS

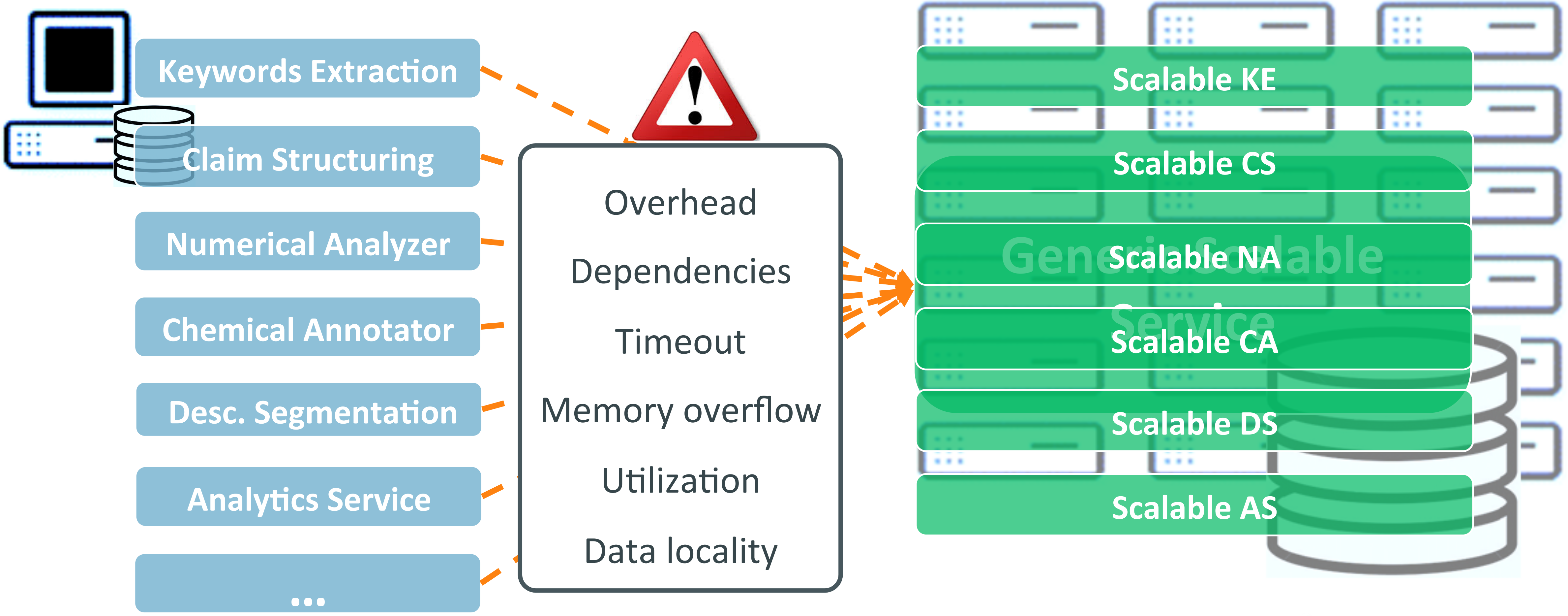
Scalable NA

Scalable CA

Scalable DS

Scalable AS

# Generic scalable framework



Keywords Extraction

Claim Structuring

Numerical Analyzer

Chemical Annotator

Desc. Segmentation

Analytics Service

...

Overhead

Dependencies

Timeout

Memory overflow

Utilization

Data locality

Generic Scalable Service

Scalable KE

Scalable CS

Scalable NA

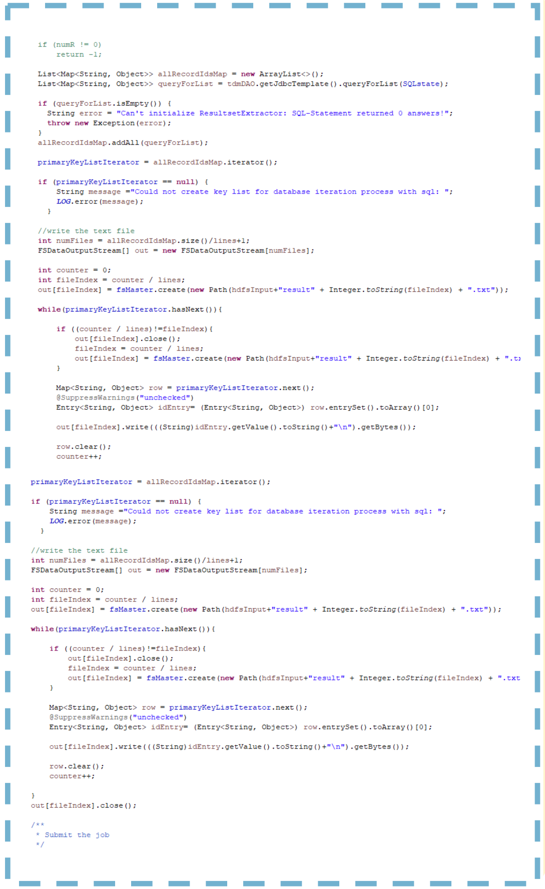
Scalable CA

Scalable DS

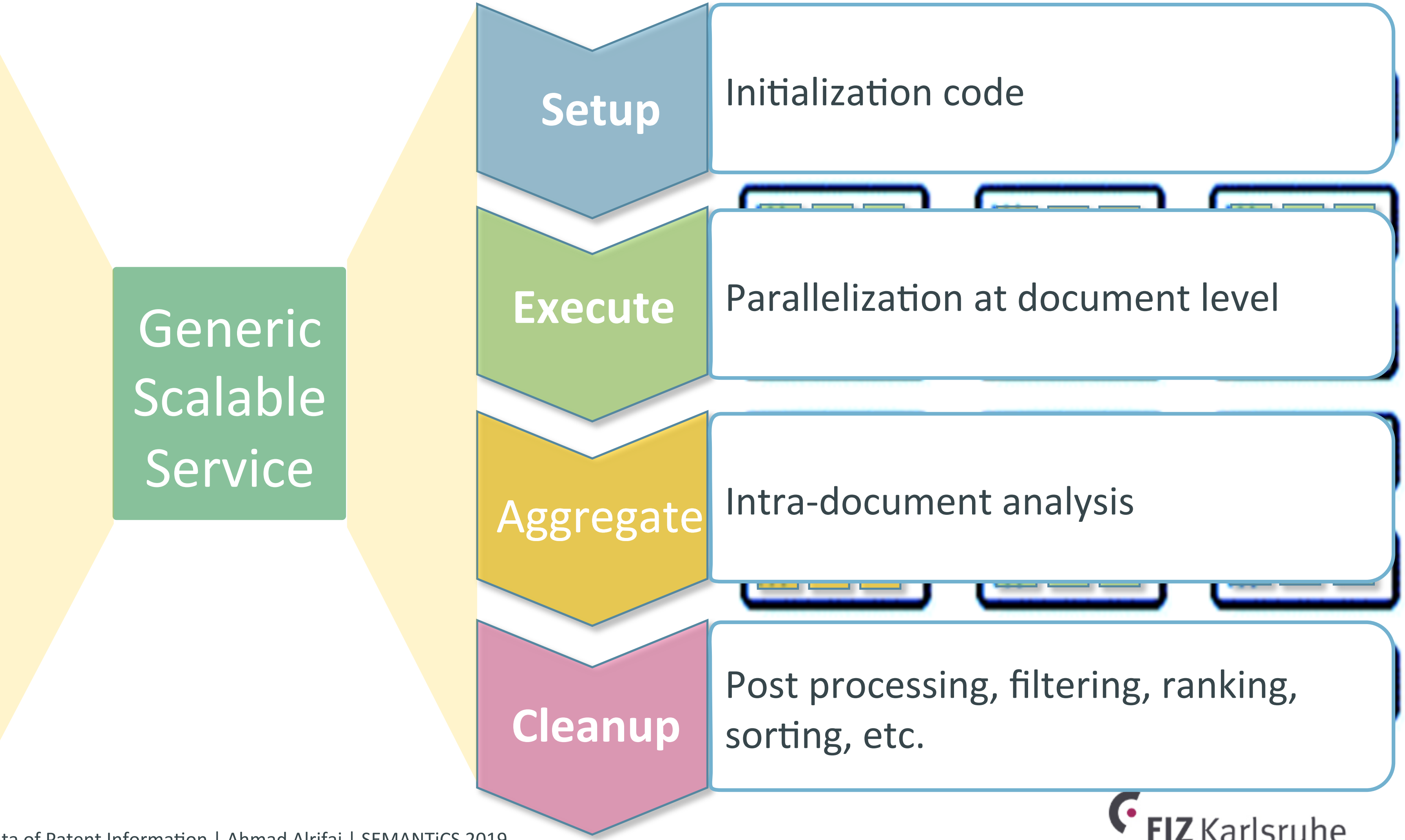
Scalable AS

# Setup, Execute, Aggregate, Cleanup

## Service.jar



```
if (numR != 0)
    return -1;

List<Map<String, Object>> allRecordIdsMap = new ArrayList<>();
List<Map<String, Object>> queryForList = tdmDAO.getJdbcTemplate().queryForList(SQLstate);

if (queryForList.isEmpty()) {
  String error = "Can't initialize ResultsetExtractor: SQL-Statement returned 0 answers!";
  throw new Exception(error);
}
allRecordIdsMap.addAll(queryForList);

primaryKeyListIterator = allRecordIdsMap.iterator();

if (primaryKeyListIterator == null) {
    String message ="Could not create key list for database iteration process with sql: ";
    LOG.error(message);
  }

//write the text file
int numFiles = allRecordIdsMap.size()/lines+1;
FSDataOutputStream[] out = new FSDataOutputStream[numFiles];

int counter = 0;
int fileIndex = counter / lines;
out[fileIndex] = fsMaster.create(new Path(hdfsInput+"result" + Integer.toString(fileIndex) + ".txt"));

while(primaryKeyListIterator.hasNext()){

    if ((counter / lines)!=fileIndex){
        out[fileIndex].close();
        fileIndex = counter / lines;
        out[fileIndex] = fsMaster.create(new Path(hdfsInput+"result" + Integer.toString(fileIndex) + ".t
    }

    Map<String, Object> row = primaryKeyListIterator.next();
    @SuppressWarnings("unchecked")
    Entry<String, Object> idEntry= (Entry<String, Object>) row.entrySet().toArray()[0];

    out[fileIndex].write(((String)idEntry.getValue().toString()+"\n").getBytes());

    row.clear();
    counter++;

primaryKeyListIterator = allRecordIdsMap.iterator();

if (primaryKeyListIterator == null) {
    String message ="Could not create key list for database iteration process with sql: ";
    LOG.error(message);
  }

//write the text file
int numFiles = allRecordIdsMap.size()/lines+1;
FSDataOutputStream[] out = new FSDataOutputStream[numFiles];

int counter = 0;
int fileIndex = counter / lines;
out[fileIndex] = fsMaster.create(new Path(hdfsInput+"result" + Integer.toString(fileIndex) + ".txt"));

while(primaryKeyListIterator.hasNext()){

    if ((counter / lines)!=fileIndex){
        out[fileIndex].close();
        fileIndex = counter / lines;
        out[fileIndex] = fsMaster.create(new Path(hdfsInput+"result" + Integer.toString(fileIndex) + ".txt
    }

    Map<String, Object> row = primaryKeyListIterator.next();
    @SuppressWarnings("unchecked")
    Entry<String, Object> idEntry= (Entry<String, Object>) row.entrySet().toArray()[0];

    out[fileIndex].write(((String)idEntry.getValue().toString()+"\n").getBytes());

    row.clear();
    counter++;

}
out[fileIndex].close();

/**
 * Submit the job
 */
```



Generic Scalable Service

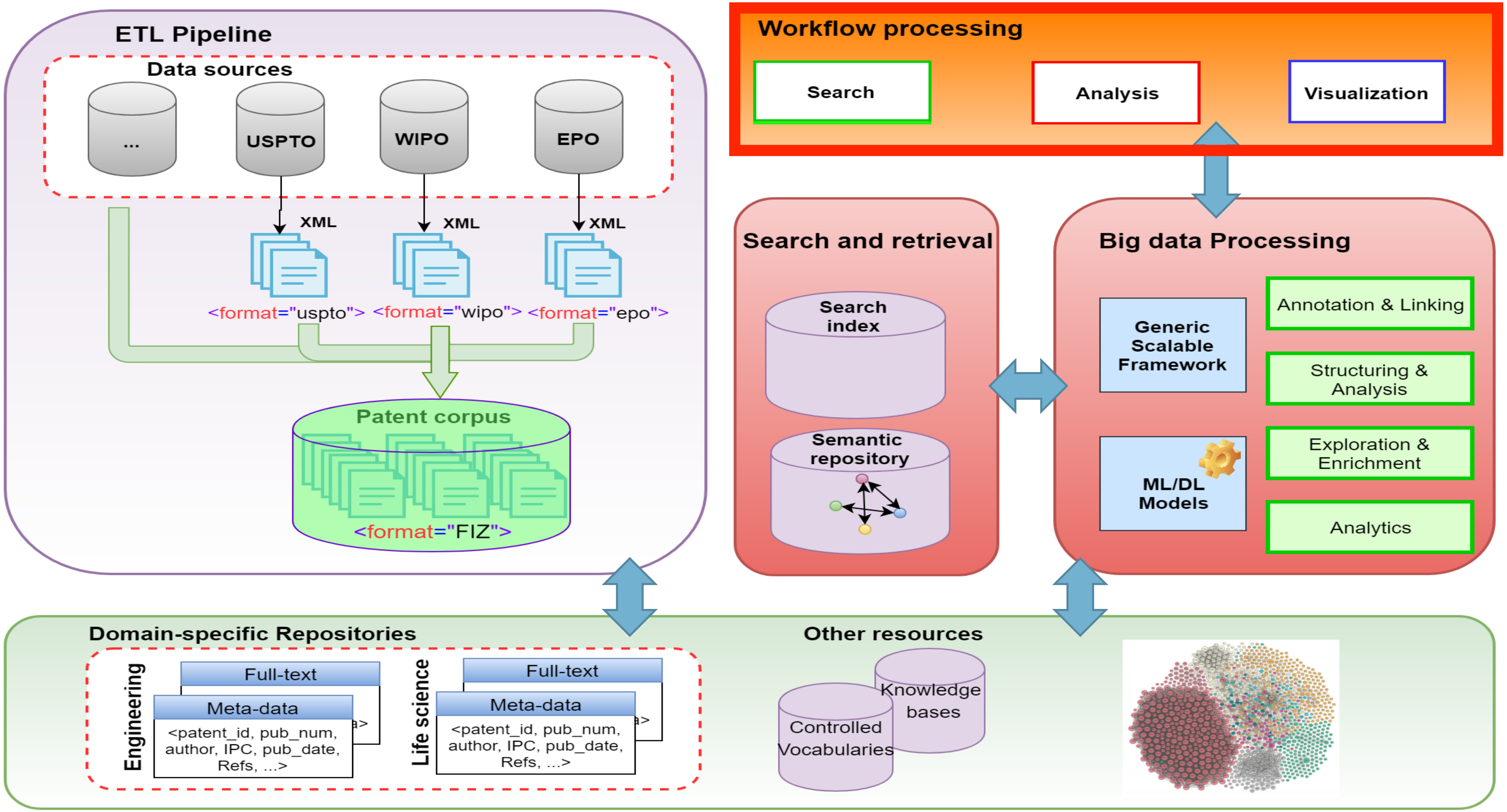
**Setup** – Initialization code

**Execute** – Parallelization at document level

**Aggregate** – Intra-document analysis

**Cleanup** – Post processing, filtering, ranking, sorting, etc.

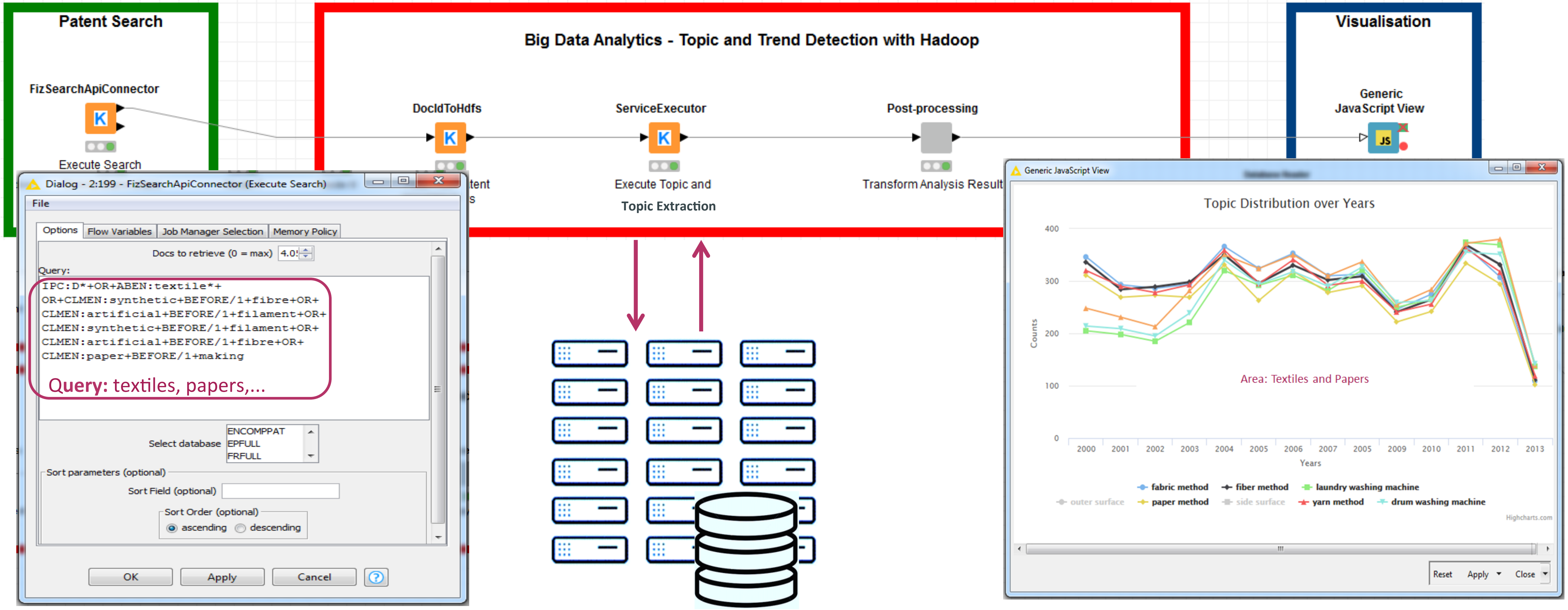
# System architecture for a powerful TDM



# Interactive workflows

Open for Innovation KNIME®

## Analysis and recognition of technology trends



Docs to retrieve (0 = max) 4.0

Query:

```
IPC:D*+OR+ABEN:textile*+
OR+CLMEN:synthetic+BEFORE/1+fibre+OR+
CLMEN:artificial+BEFORE/1+filament+OR+
CLMEN:synthetic+BEFORE/1+filament+OR+
CLMEN:artificial+BEFORE/1+fibre+OR+
CLMEN:paper+BEFORE/1+making
```

Select database: ENCOMPPAT, EPFULL, FRFULL

Sort parameters (optional)

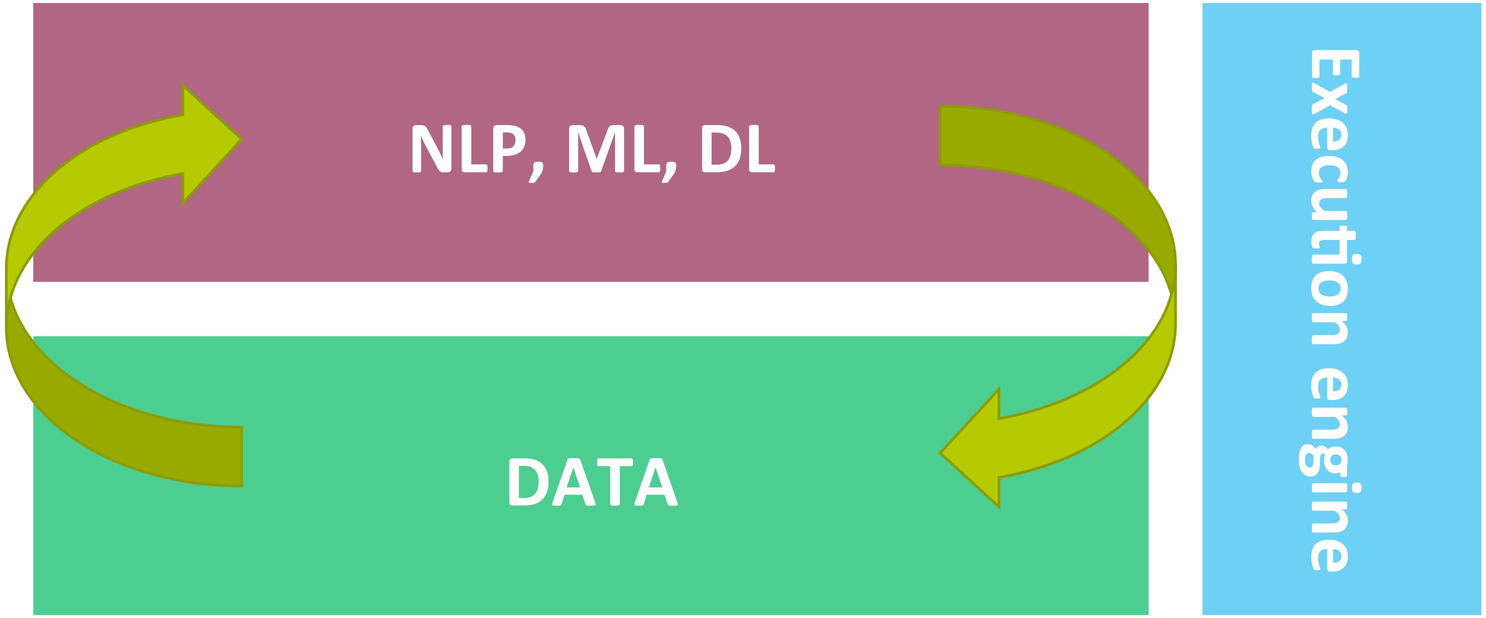
Sort Field (optional)

Sort Order (optional): ● ascending ○ descending

# Whats‘s next  - next generic Generation

Interoperability of NLP and machine learning steps in interactive workflows

- Data level: standardized formats
- Execution engine: abstractions and optimizations – Ontology for TDM

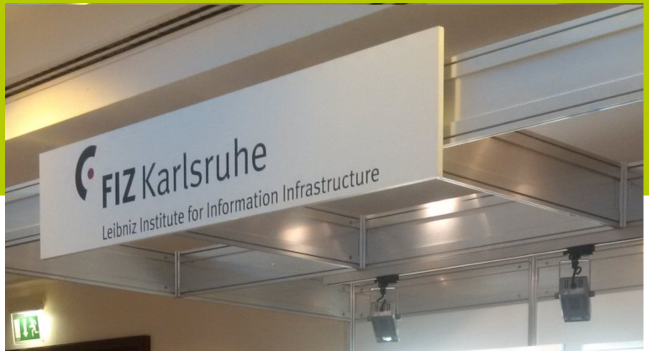


# Conclusion

- Patents form a unique and valuable knowledge source (also beyond IP domain)
- Need for the most advanced techniques to generate synergies and added values
- Patent analytics domain is catching up with adapting Machine learning and semantic technologies
- Scalable infrastructure and generic frameworks advance semantic technologies and boost the performance of TDM applications
- Considering and linking with domain-specific KBs to maximize the potentials of patent data mining
- Integrity, compatibility and interoperability of NLP and Deep learning for better comparability and reusability

# THANK YOU!
# Questions ?

## Visit us!



**Contact**

Ahmad Alrifai
Senior Researcher

+49-7247 808-545
Ahmad.Alrifai@fiz-karlsruhe.de

**Text and Data mining**
https://www.fiz-karlsruhe.de/de/forschung/text-und-data-mining-tdm

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Zertifikat seit 2016
audit berufundfamilie

Leibniz
Association