TREPAREL

Text Analytics in the EU Fusepool project

II-SDV '13 - Nice

Treparel
Delftechpark 26
2628 XH Delft
The Netherlands
www.treparel.com

Anton Heijs
CTO
anton@treparel.com

April 15, 2013

Agenda

- Introduction of the EU Fusepool project, Treparel and the consortium partners
- Background on objectives of the Fusepool project
- User adaptive system
- Data pooling and linking
- Large scale machine learning and text analytics
- Summary

TREPAREL

The EU Fusepool Consortium

Fusepool

Bern University of Applied Sciences

European Network of Living Labs

searchbox

SEVENTH FRAMEWORK PROGRAMME

xerox

Geo

TREPAREL

Treparel (Delft, The Netherlands) is a global software provider in Big Data Text Analytics and Visualization.

Global companies, government agencies, software vendors or data publishers are using Treparel KMX text analysis software to gain faster, reliable, precise insights in large complex unstructured data sets (like application notes, blogs, email and patents) allowing them to make better informed decisions.

As part of the Fusepool consortium Treparel integrates the KMX Text Analytics technology for advanced classification, clustering and visualization of large complex document collections.

TREPAREL

The Big Picture

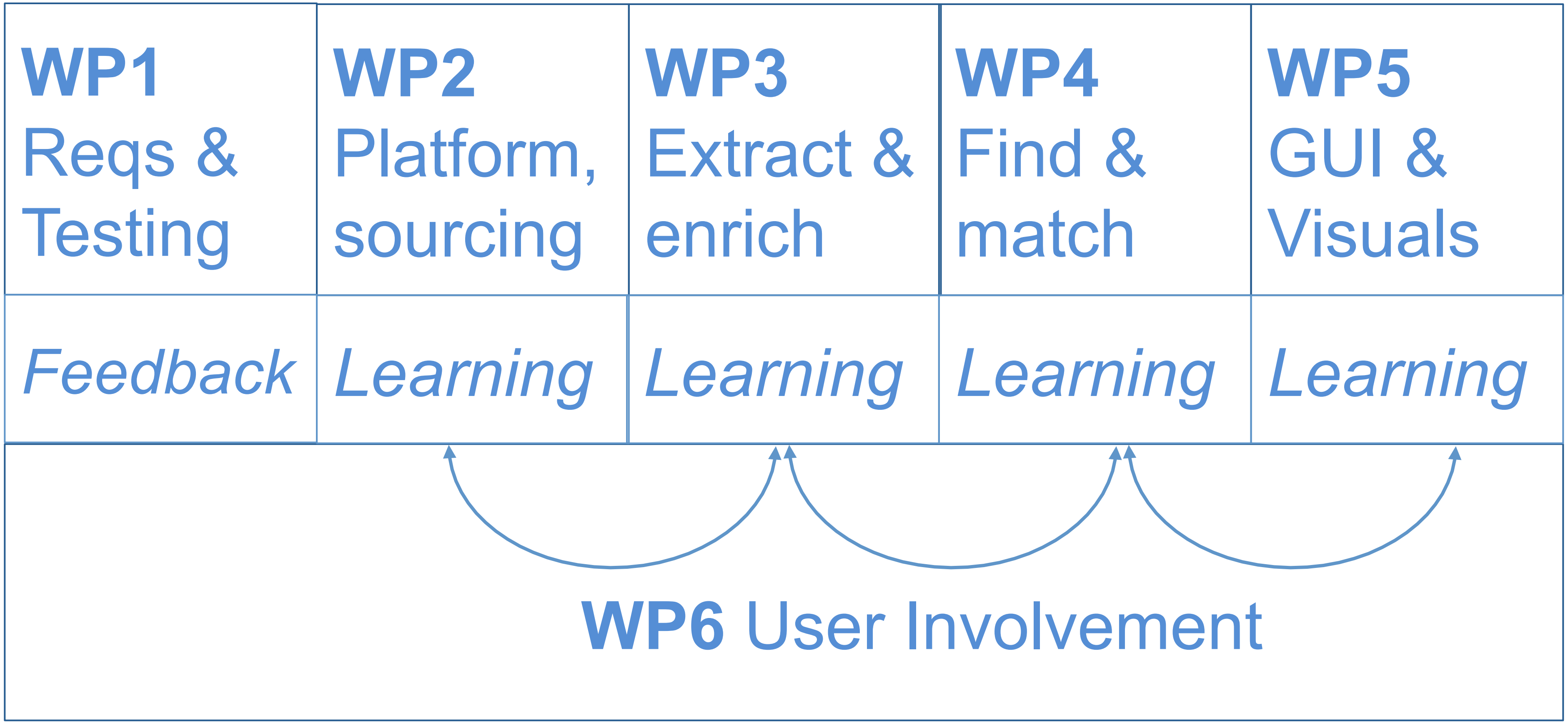
Fusepool

Each partner provides critical **building blocks:**

- ✓ **BUAS**: Dynamic user interfaces
- ✓ **ENOLL**: End users with specific needs
- ✓ **GEOX**: NER, UI design
- ✓ **SEARCHBOX**: text search and similarity semantic matching
- ✓ **TREPAREL**: text/patent classification engines
- ✓ **XEROX**: user-adaptive learning know-how

➡ **BUT BREAK-THROUGH BY INTEGRATION**

Work packages



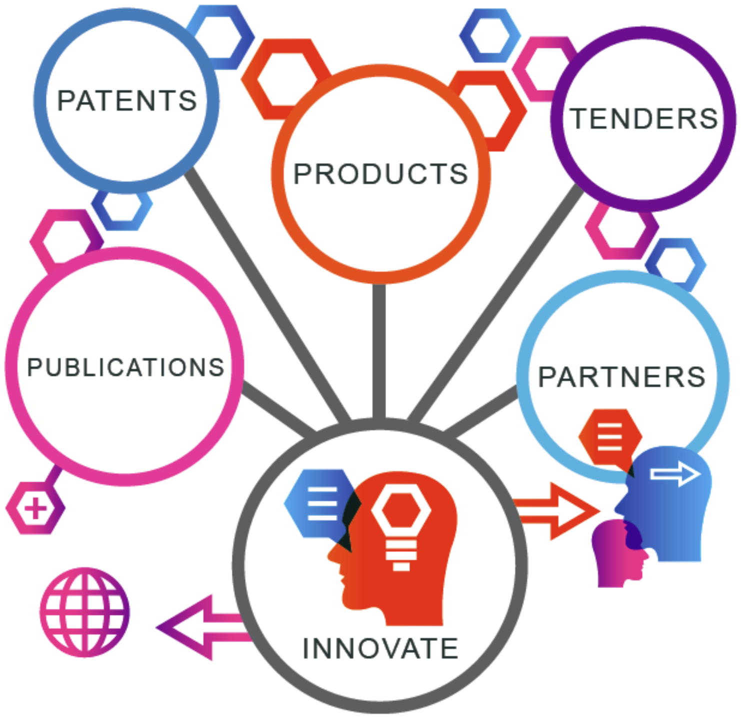
WP1 Reqs & Testing	WP2 Platform, sourcing	WP3 Extract & enrich	WP4 Find & match	WP5 GUI & Visuals
Feedback	*Learning*	*Learning*	*Learning*	*Learning*

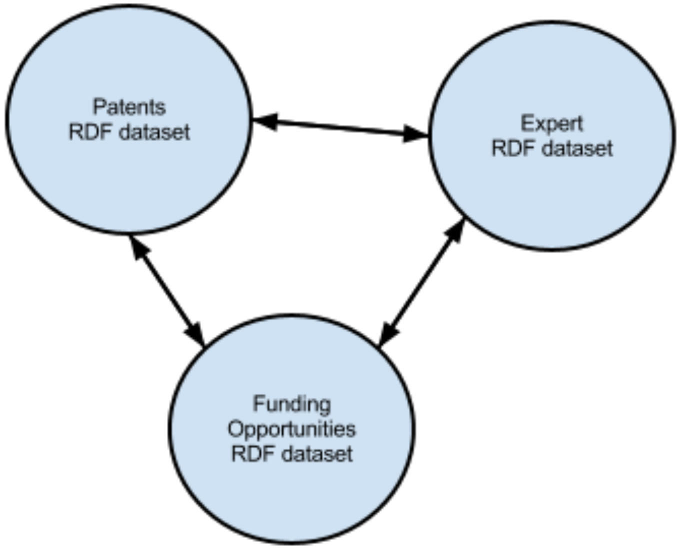
WP6 User Involvement

The EC Fusepool project is a 2 year EU project and started in July 2012

Fusing and pooling information for product development

Vision:	User-adaptive system
Living Lab:	Rapid app development
Data processing:	Sourcing & interlinking
Machine learning:	Matching & optimizing





Integrated Use Cases

Background

- SMEs have a need for technology intelligence for detecting and responding to opportunities and threats
  - This a partly driven by growth and complexity of patents and lawsuits
- Consumer intelligence to detect opinions and needs of consumers for product development
- Open innovation requiring cooperation (links between data, e.g. finding business partners)
- Focus: Machine Learning algorithms to improve matching

User-adaptive system

- **Focus**: monitor and learn specific needs and preferences of a user to align features, functionalities, and graphical interfaces
- **Adaptive**: machine learning from crowd-sourcing (rather than rule-based)
- **User-aligned prioritization**: more usable and customized interfaces, suggestions based on activity & user feedback

User-adaptive matching

- **Main goal**: automated user-adaptive matching of users to:
  - Patent analysis
  - Finding funding opportunities
  - Partner matching
- **Key asset**: information provided by the user
- **User Data Credo**: accuracy improves with quantity and quality of user data while variety (breadth) increases with number of users
- **Living lab**: Co-creation between creators and consumers of the Fusepool platform

Data sourcing

- **Sources**: internal & external content from web harvesting and structured data sources
  - using content databases and linked open data
- **Scope**: initial data corpus includes all explicitly in- and excluded sources
- **Gained value from Information**: recommendations based on machine learning from feedback

Data handling

1. **Text analysis and feature extraction**: ML & NLP methods for categorizing, named entity extraction, etc.
2. **Shared metadata models**: mapping text features to existing/custom ontologies and generation of semantic triplets
   - → High-level abstraction & persistence for reuse
   - → Lightweight storage: mostly metadata only, text indexing and abstraction uses schema-free key-value (enabling actionable facets)

Data interlinking

- **Contextualize**: terms are interlinked with same and similar terms across sources:
  - Enrich the extracted content with existing information available in the Internet
  - Interlink as much information as possible to increase the value of knowledge extraction
  - Use available public sector resources in Semantic Web and LOD format
- **Metadata**: when a user uploads texts to be matched with other content, only the metadata descriptors are transmitted
- **Data privacy**: data fusion from diverse sources without endangering user privacy

Searching & finding

- **Key search-oriented features**:
  - Search through all content in the data pool
  - Faceted search (categories, metadata, entities)
  - Integration of Linked Open Data (LOD) results
  - Cross-lingual indexing and cross-referencing
  - "Did you mean?"-functionality in case of typos and auto-completion of search queries
- **User-adaptive**: indexing and integration based on users needs (e.g. user profiling)

Adaptation & refinement

- **Adaptive search**: results are aligned to user preferences based on analysis of user implicit and explicit feedback
- **Multi-task ranking**: good trade-off between user-independent search (high coverage but lower precision) and a very customized approach
- **Query intent discovery**: analysis of the query structure and interlinking of queries

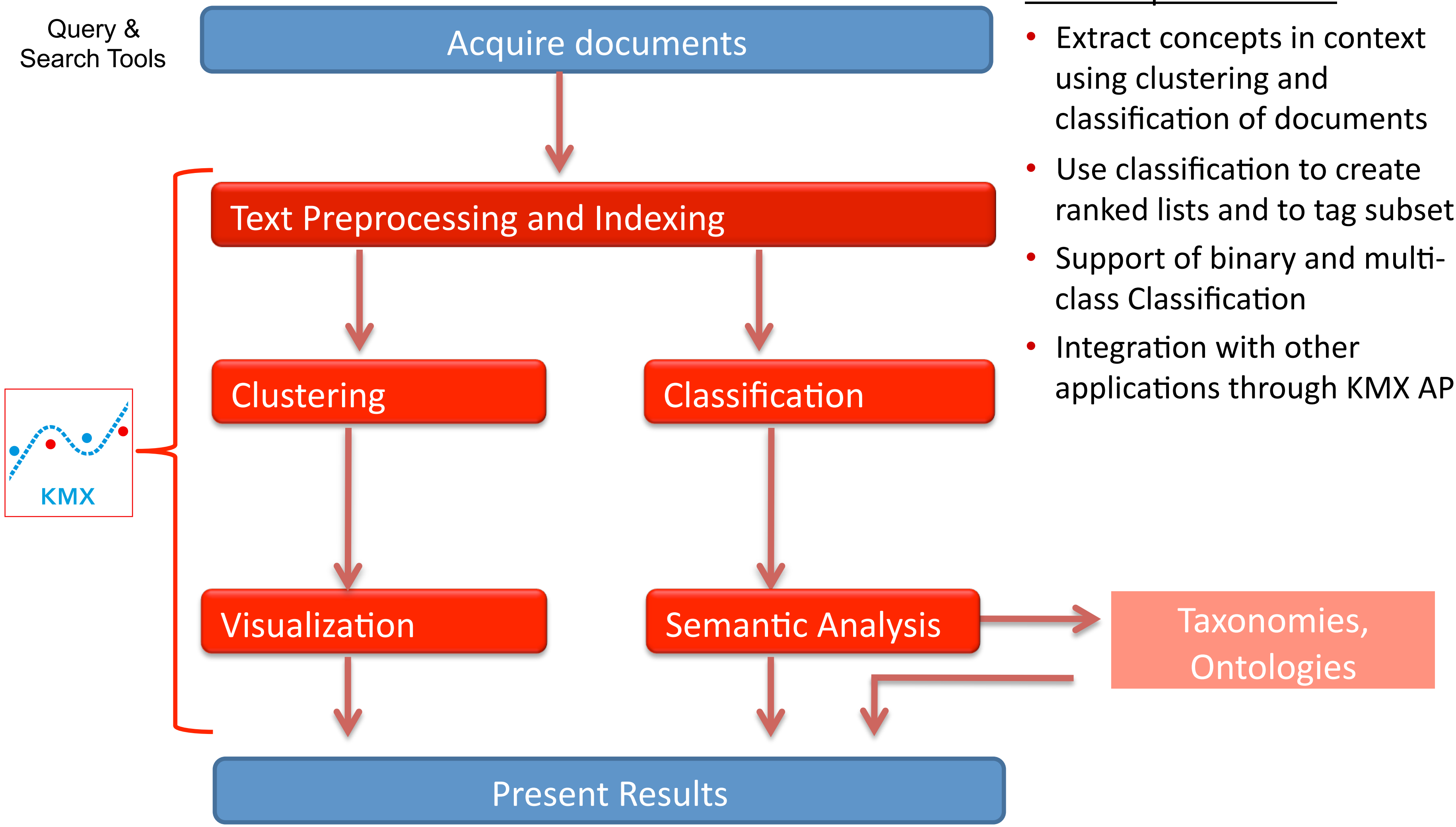
Correlating & matching

- **Search guided navigation**: semantic matching extracts contextual relationships to list related content
  - suggestions organized by categories
  - exposing facets within related content
- **Distributed rule and event model**: defines states, actions, and consequences (e.g. notifications, visualizations) for reasoning based on light-weight ontologies

Crowd sourcing & supervised automation

- **Relational learning**: related instances are used to reason about the focal instance
  - Rationality of content (links to other content, people, etc.) provide rich information
  - Similarities/dissimilarities to other content is established purely on relational properties

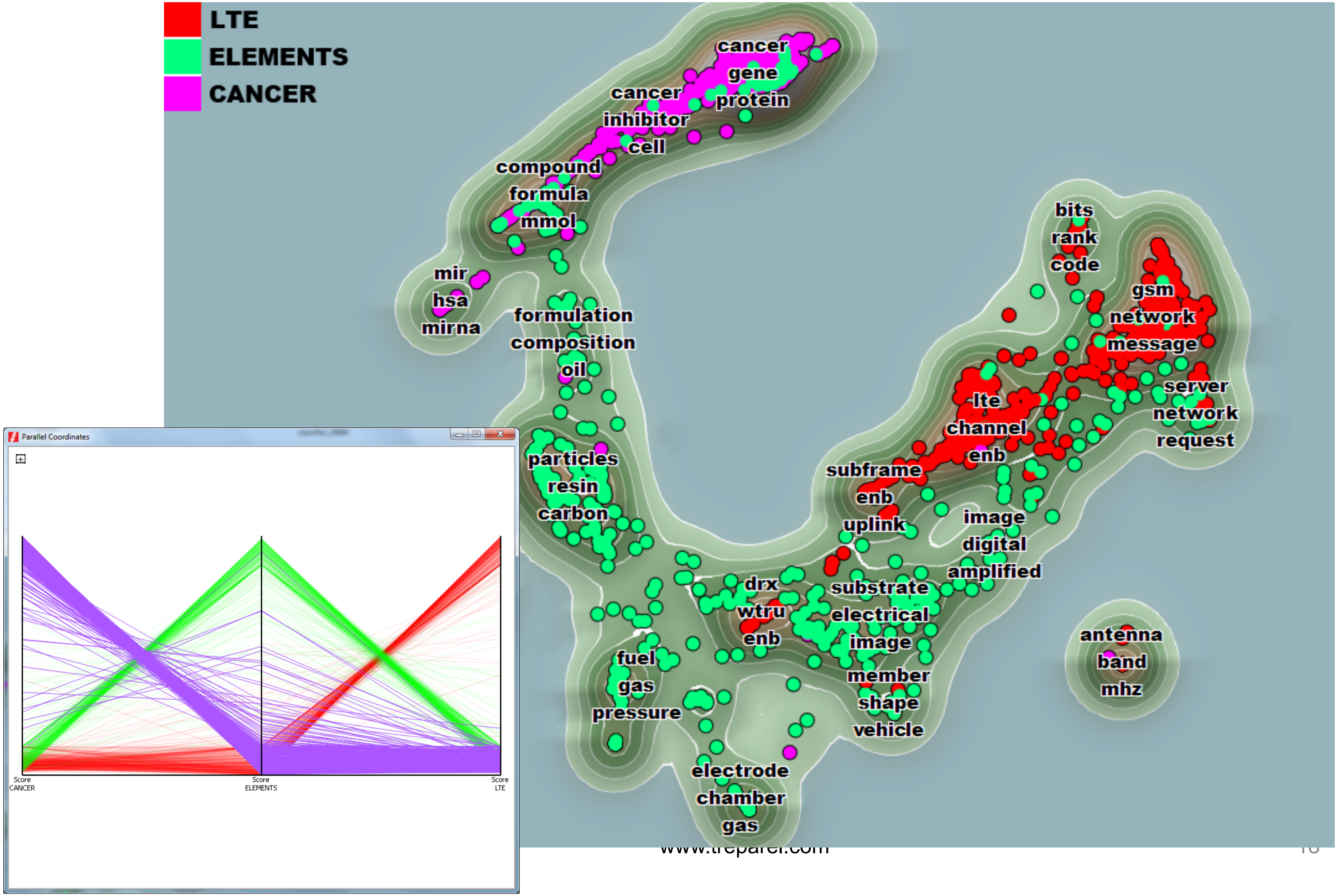
Document level analysis using the KMX technology



KMX unique functions:

- Extract concepts in context using clustering and classification of documents
- Use classification to create ranked lists and to tag subsets
- Support of binary and multi-class Classification
- Integration with other applications through KMX API

NER in the landscaping



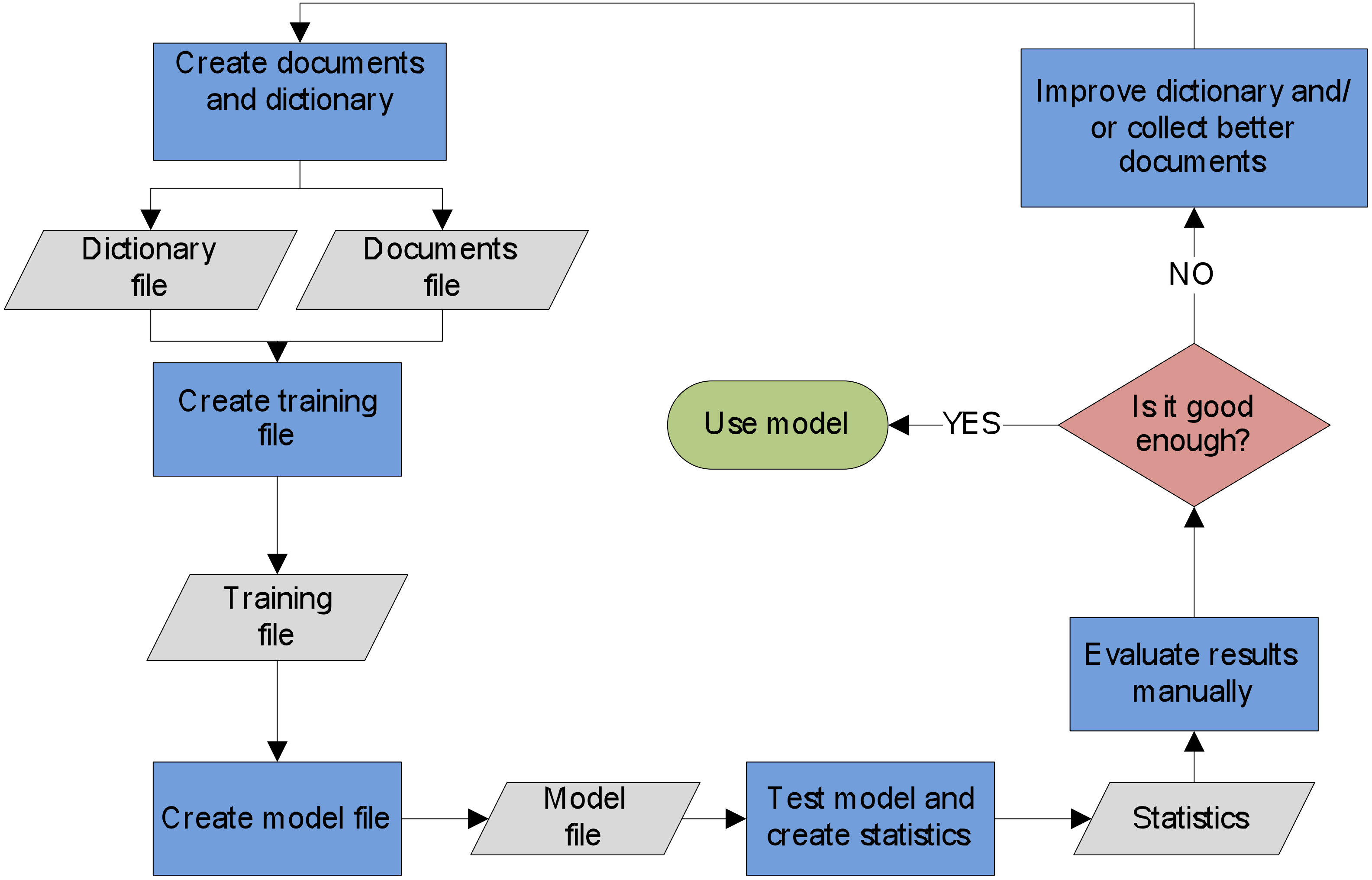
Sentence level analysis using Named Entity Recognition

- The aim of NER is to identify entities in unstructured text documents.
  - To locate : mark-up the entities
  - To classify : into predefined categories/domains
- Aim of usage
  - To recognise trends (trained and new high frequency)
  - To find all „trained” entities
  - To „discover” new entites
- NER approaches
  - Statistics based (supervised machine learning)
  - Rule based (regular expressions)

TREPAREL

Building a NER model

Fusepool

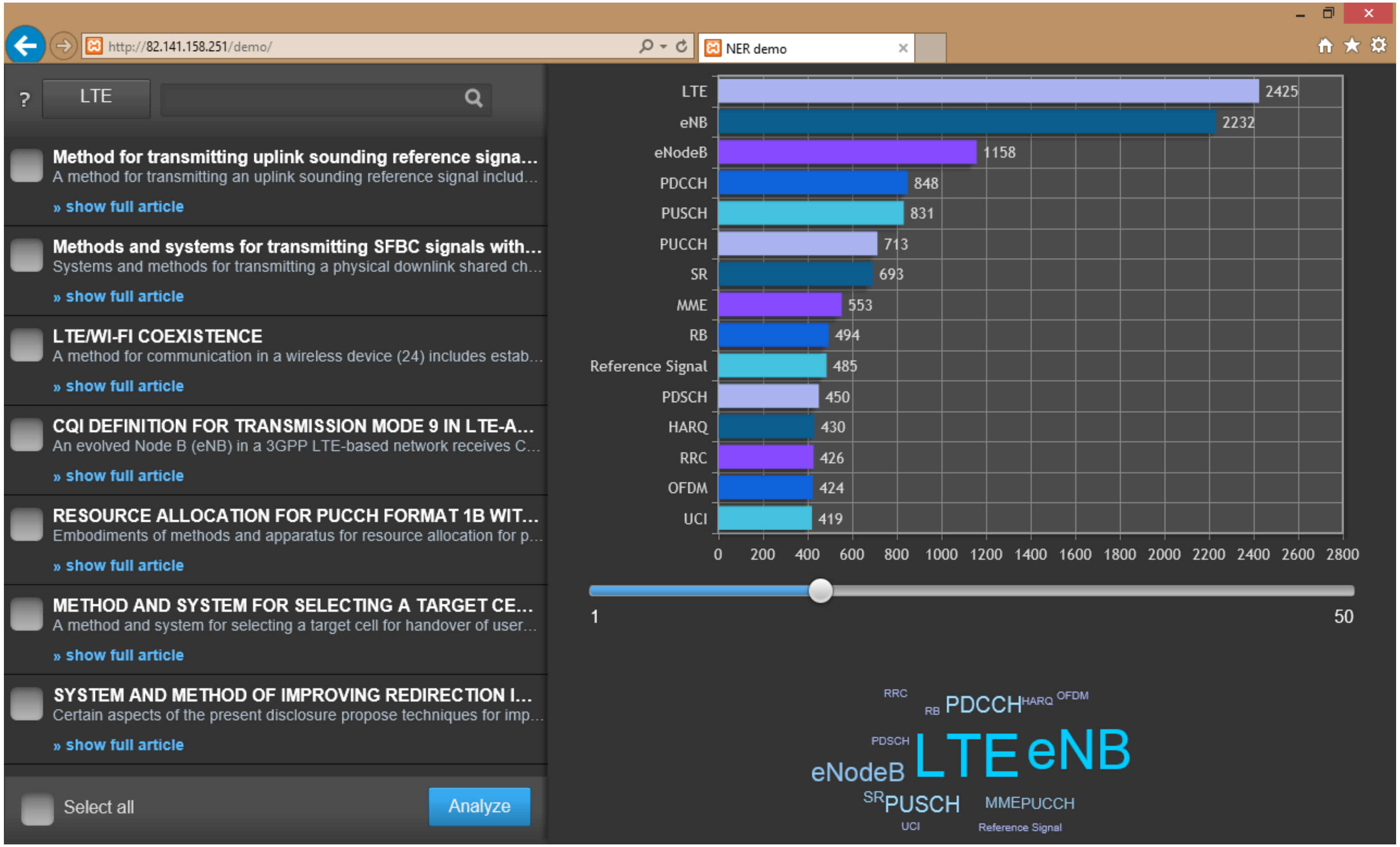

Create documents and dictionary
Dictionary file
Documents file
Create training file
Training file
Create model file
Model file
Test model and create statistics
Statistics
Evaluate results manually
Is it good enough?
YES
Use model
NO
Improve dictionary and/ or collect better documents

3 NER model examples

- ✓ Training text: 500 patents from EPO
- ✓ Training: model building by Stanford NER

- **LTE (long term evolution)**
  - F1: 88%
  - New entities: „GSM BSS”, „LTE TDD”
  - False pos: „Location Area LA”
- **Elements**
  - F1: 98%
  - False pos: „argon/hydrogen”
- **Cancer**
  - F1: 87%
  - New entities: „myeloma cell”, „tumor tissue”
  - False pos: „cell mortality”, „the test compound”

Using NER in a GUI



http://82.141.158.251/demo/

NER demo

? LTE

Method for transmitting uplink sounding reference signa…
A method for transmitting an uplink sounding reference signal includ…
» show full article

Methods and systems for transmitting SFBC signals with…
Systems and methods for transmitting a physical downlink shared ch…
» show full article

LTE/WI-FI COEXISTENCE
A method for communication in a wireless device (24) includes estab…
» show full article

CQI DEFINITION FOR TRANSMISSION MODE 9 IN LTE-A…
An evolved Node B (eNB) in a 3GPP LTE-based network receives C…
» show full article

RESOURCE ALLOCATION FOR PUCCH FORMAT 1B WIT…
Embodiments of methods and apparatus for resource allocation for p…
» show full article

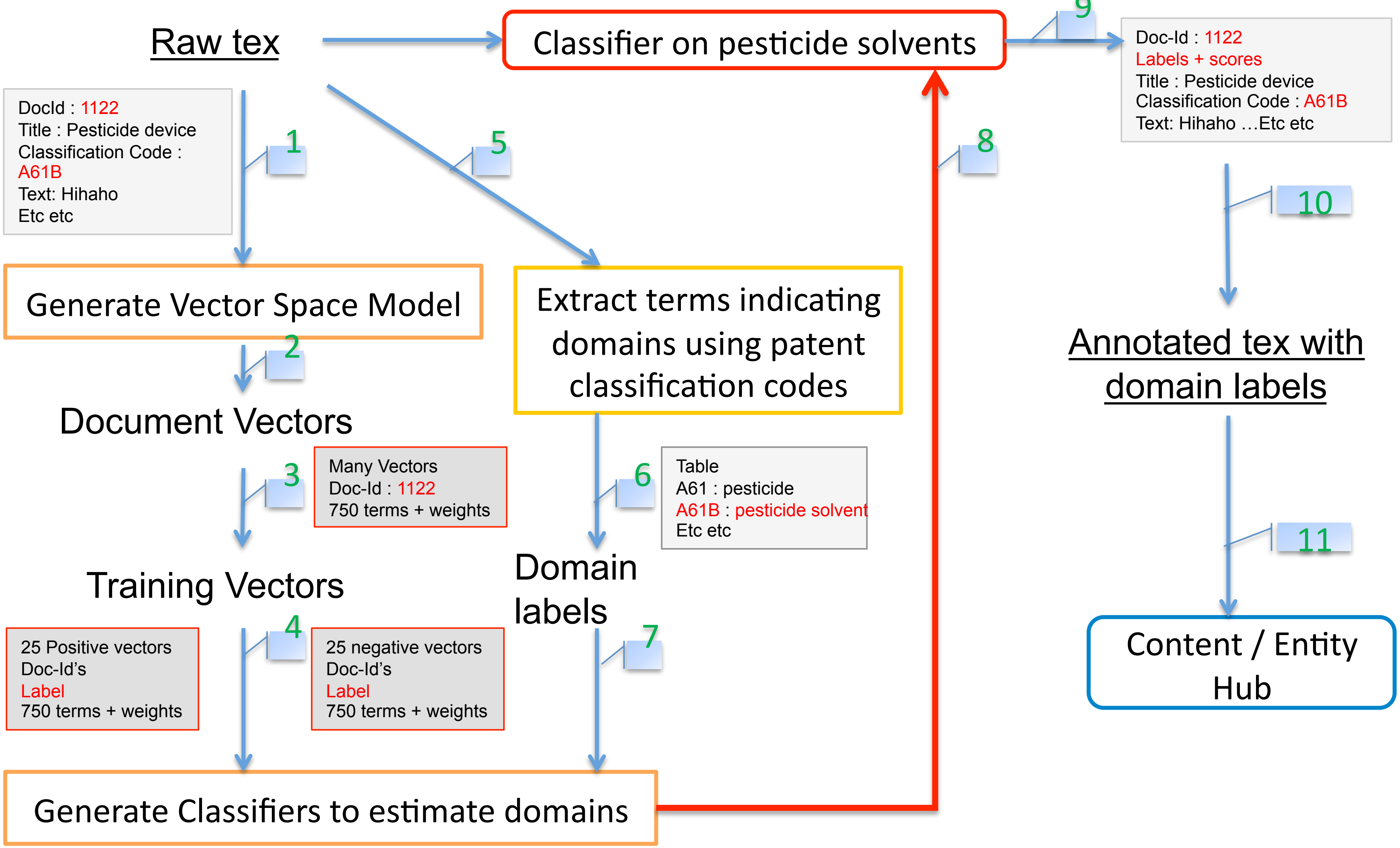
METHOD AND SYSTEM FOR SELECTING A TARGET CE…
A method and system for selecting a target cell for handover of user…
» show full article

SYSTEM AND METHOD OF IMPROVING REDIRECTION I…
Certain aspects of the present disclosure propose techniques for imp…
» show full article

□ Select all

Analyze

Combining text analysis approaches



Raw tex

Classifier on pesticide solvents

DocId : 1122
Title : Pesticide device
Classification Code :
A61B
Text: Hihaho
Etc etc

Generate Vector Space Model

Document Vectors

Many Vectors
Doc-Id : 1122
750 terms + weights

Training Vectors

25 Positive vectors
Doc-Id's
Label
750 terms + weights

25 negative vectors
Doc-Id's
Label
750 terms + weights

Extract terms indicating domains using patent classification codes

Table
A61 : pesticide
A61B : pesticide solvent
Etc etc

Domain labels

Generate Classifiers to estimate domains

Doc-Id : 1122
Labels + scores
Title : Pesticide device
Classification Code : A61B
Text: Hihaho …Etc etc

Annotated tex with domain labels

Content / Entity Hub

Software architecture

- **Data sourcing**: retrieves data from data sources via ***data integrator***
- **Storage**: raw (e.g. text) or processed data stored in database or triple store
- **Using ML** to enable learning from the crowd
- **Push & Pull**:
  - interface to consumers (web and mobile apps)
  - supporting quality and access control from portal
- **Portal**: business logic, storage of registered data sources, access control using web frontend

Open Call for Users

- **Applications received** from Finland, Germany, Spain, Greece, Bulgaria, Hungary, Italy, UK, Belgium, China, Switzerland, France, Denmark, Portugal, Ireland and The Netherlands
- **Business areas covered**:
  - Bio-medical,
  - Pharma and biotech,
  - ICT/Telecommunications,
  - Digital media,
  - Renewable energies,
  - Education,
  - Innovation/consultancy services for SMEs.
- **Mix of profiles** from SMEs, Research, SME intermediaries (incubators, Science parks, Living Labs, etc), developers and more
- **Multiple areas** of research, development and innovation areas are covered.

EC perspective

- **Data as a service**: scale economies of scale in management of data
- **Data pooling**: processes need timely aggregation and redistribution of diverse data but building own is redundant and prohibitive for SMEs
  - provide services on top of pool with high quality data
  - provide access to services on demand
- **Success criteria**:
  - Early provision of scalable basic Fusepool services
  - SME involvement and uptake of Fusepool services
  - Machine learning for data curation & user adaptation
- **Required steps**:
  - Stepwise integration of existing & new technologies
  - Early and ongoing feedback from end users

Summary

The EC FP 7 program Fusepool is all about:

- Building a platform with web enabled services for
  - Data pooling
  - Large scale text analysis
  - Large scale machine learning of user input
- Enable SME's with analytics for improving their innovation and competitive strengths using
  - SME's involvement and feedback to the Fusepool services
  - Machine learning for data curation & user adaptation

Welcome to visit Fusepool at:
www.fusepool.net