

Overcoming Language Barriers in Patent Research

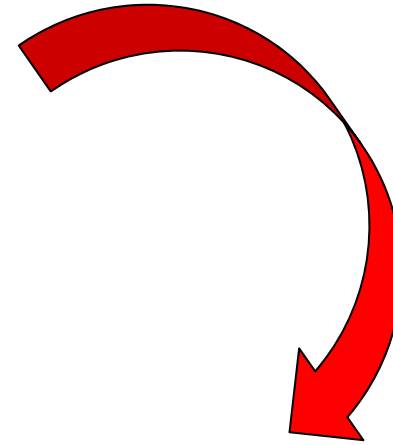
Peter J Vanderheyden
Vice President & Managing Director,
Global IP Solutions



- Who is LexisNexis
- The Language Barrier Problem
- The LexisNexis Perspective
- Underlying Assumptions
- The LexisNexis Solution – In 3 Parts
 - ❖ Full Text
 - ❖ Machine Translation
 - ❖ Semantic Search

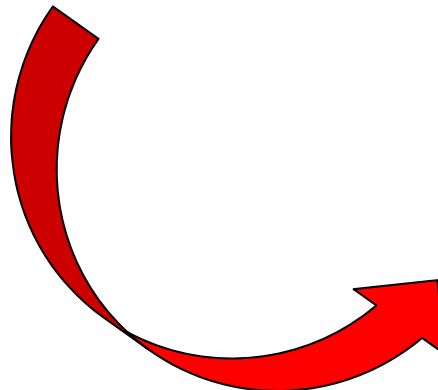
Founded: **1973**
World Headquarters: **New York City**
Parent Company: **Reed Elsevier**
Global Reach: **Customers in > 100 countries**
Employees: **> 18,000 globally**
Offices Worldwide: **110**
Revenue: **£2.5+ B 2009 (per RE Ann. Rpt.)**

Stock Symbols: **NYSE: ENL; NYSE: RUK**

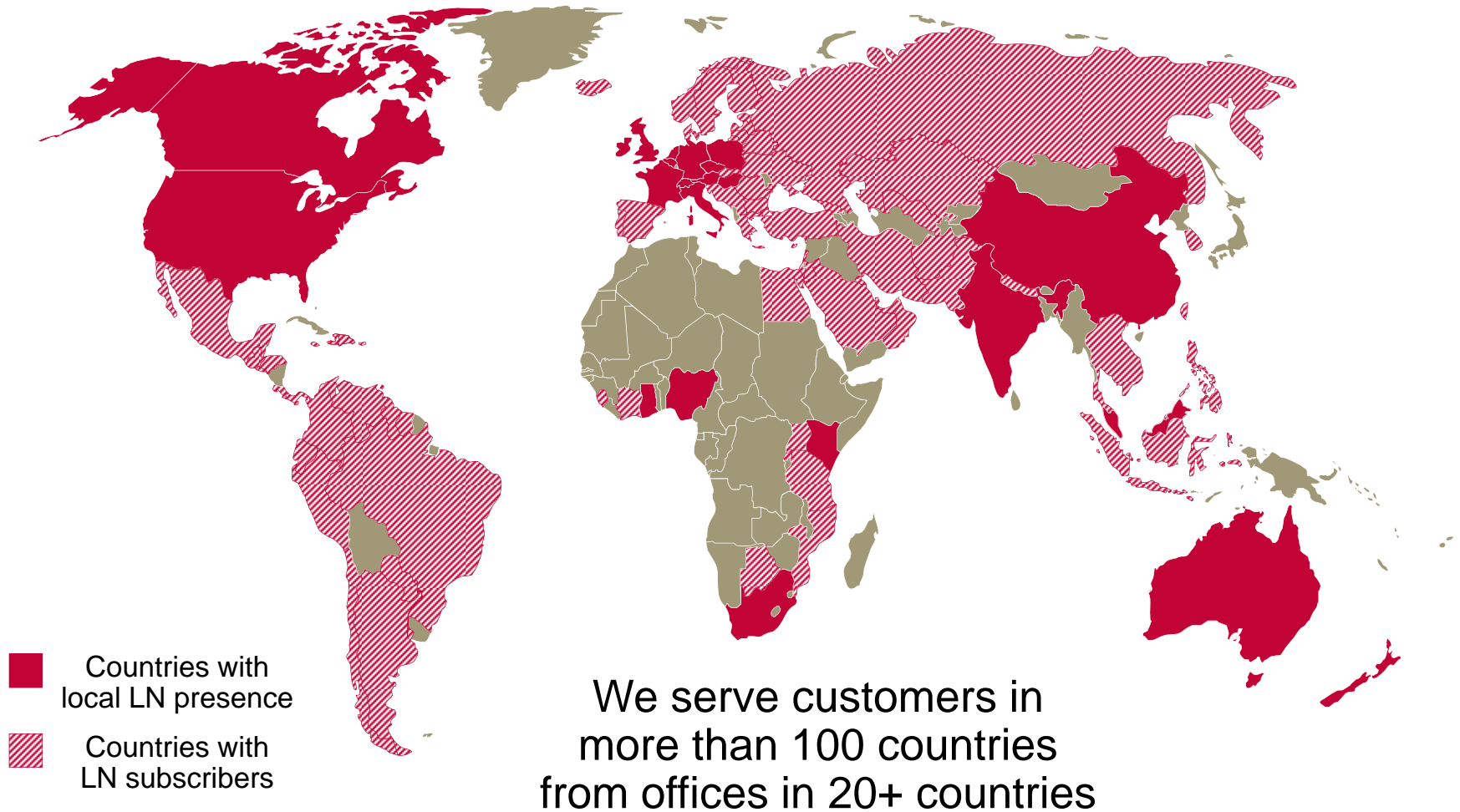


LexisNexis® is a leading global provider of **content-enabled workflow solutions** designed specifically for professionals in:

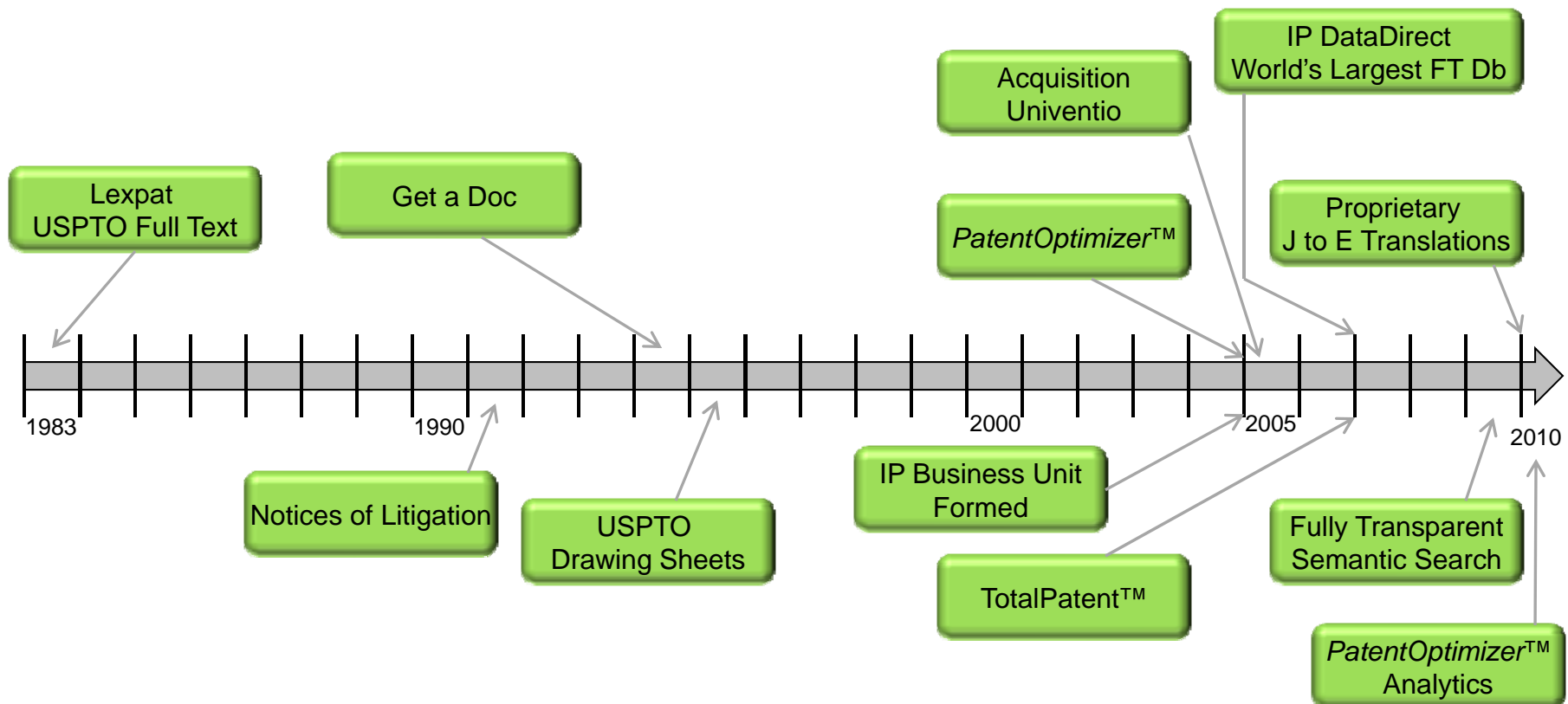
- Law Firms
- Corporate Counsel
- Librarians
- Government Agencies
- Law Enforcement
- Tax Specialists and Accountants
- Colleges and Universities
- Risk and Compliance Officers
- Insurance Companies



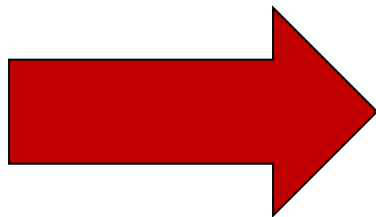
A Significant & Growing Global Presence



- Historical commitment to patent research
- Reinvigorated investment and dedication from 2005
- Industry leading solutions

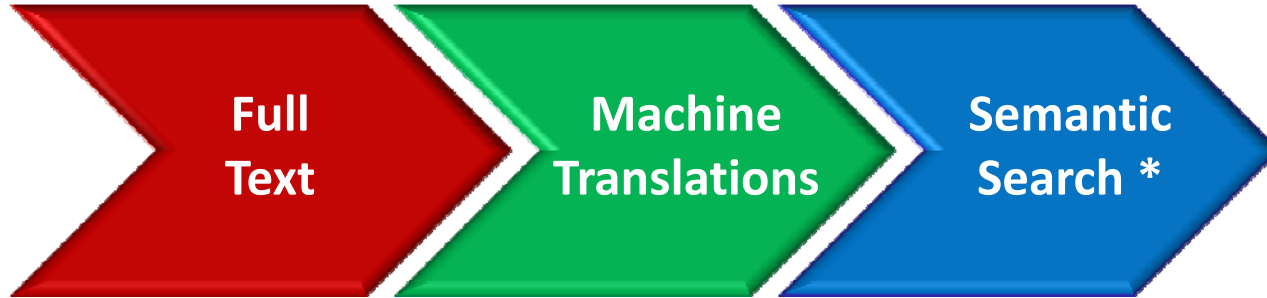


- Variety of languages for patent/tech documentation
- IT implications of these languages/characters
- Limited skill level of the average searcher relative to these many languages
- Familiarity with vocabulary, especially critical in the world of patent literature



These patents/prior art still matter regardless, and being able to find relevant patents/applications is of critical importance. In fact, as searchers begin to open up these previously less accessible collections the need to do so will accelerate for all.

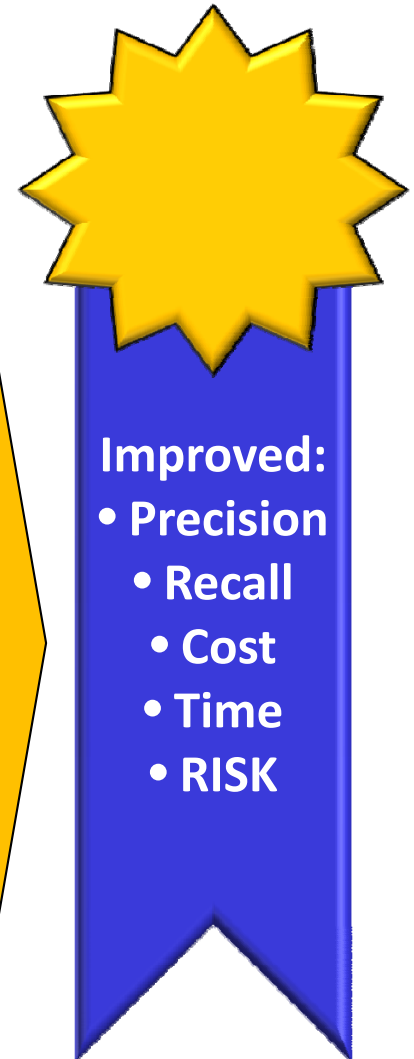
+ Professional User Experience!!



- Search where you're most likely to find the answer
- Avoid higher cost, less comprehensive A/I
- Improve chances of finding key elements
- Use current information available rapidly upon publication

- Search multiple collections simultaneously
- One query
- State of the art
- Continuous improvement
- Focus on lexical correctness

- Deep insight into knowledge base
- Extra helpful to a non-native speaker
- Enable weighted Boolean queries
- Complete transparency
- Complete control
- Ranking of results



* As an additive tool to today's expert Boolean search techniques

- There is no perfect solution/panacea
- Searching is about reducing RISK
- Normalize to a single language
- Short Term: English
 - Massive amount of learning content (for MT/Semantic)
 - High % legal/tech professionals
 - Large % of patent/NP content
 - Investment extensible to other content
- Long Term: Other Primary Languages ?
 - Economies are challenging
 - Availability of MT technologies/investment
 - Learning corpus is (more) limited

What is it ?

- *Description*
- *Claims*
- (Everything > Title/Abstracts)
- Challenges
 - # Sources (constantly improving – government/vendor)
 - OCR
 - Tagging
 - Tables
 - Figures
 - Chemical Information
 - Sequencing

Why this is important ?

- Patents are about details – words matter
- Details are in the Full Text ¹
 - “The highest density of unique entities is found in the title. In the other sections the density is nearly the same but the mean of unique entities is highest in the description. The second highest mean is found in the claim section.” (emphasis added)
 - “Only up to four percent of all entities are unique to the claim section in comparison to the description section. At least 79% of all entities are unique to the description section in relation to the claim section.” (emphasis added)
 - “...between 92% and 97% of the entities are unique to the sections claims plus description.” (emphasis added)

1 – TREC-CHEM patent database study: H. Cunningham, A. Hanbury, and S. Rüger (Eds.): IRFC 2010, LNCS 6107, pp. 161, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Full Text Access is Increasing



11 Authorities now
publish some Full Text



LexisNexis™

Vendors like LexisNexis Produce More
20 Add'l Countries including:
Canada, Italy, Netherlands, France, Portugal,
Sweden, etc.

Description Section

- Original English 18.5 M Patents
- MT English 15.5 M Patents
- Total English **34.0 M** Patents

Claims Section

- Original English 18.2 M Patents
- MT English 11.5 M Patents
- Total English **29.7 M** Patents

59M Titles and 36M Abstracts in English (native or machine translated)

What is it? Two Technologies -- Rules Based MT v Statistical MT

	RBMT	SMT
Terminology	<ul style="list-style-type: none"> • Limited number of user dictionaries • Limited number of entries per user dictionary • More dictionary entries => slower performance • Coding of morphological info • Dictionary taxonomy management • Disambiguation via semantic rules or dictionary taxonomy 	<ul style="list-style-type: none"> • Bilingual text available • Statistics performs the linguistic work as far as possible (finding of term pairs and disambiguating based on context) • Linguists get a validation function • Unlimited phrase pairs supported • More options => more computing time
Grammar	Good grammar precision: serves text “intelligibility”	Less good grammar precision, but text is still understandable.
Language coverage	<ul style="list-style-type: none"> • Limited to the language pairs offered by the system • A dictionary created for a specific language pair cannot often be “inverted” for usage in the reverse language pair 	<ul style="list-style-type: none"> • Extensible system. New language pairs can be developed. • A dictionary/phrase pair created for a specific language pair can be used for the reverse language pair as well.
Customization	Mostly out of the box solutions	Pre- and post-processing rules
Time-to-market	<ul style="list-style-type: none"> • Baseline translation immediately available if language pair is supported • If language pair is not supported, look for a different product • Customized translation requires at least 2-3 years work to get to a desired quality level 	<ul style="list-style-type: none"> • Baseline translation immediately available if language pair is supported • If language pair not yet supported, baseline engine can be available in 6 months’ time • Customized translation for existing baseline available in 2-3 months’ time (dependent on language pair)

JP = around **4.7 million documents** with full text

Due to the **novel and technical** character of patents, **terminology** is relatively **unique**.

RBMT

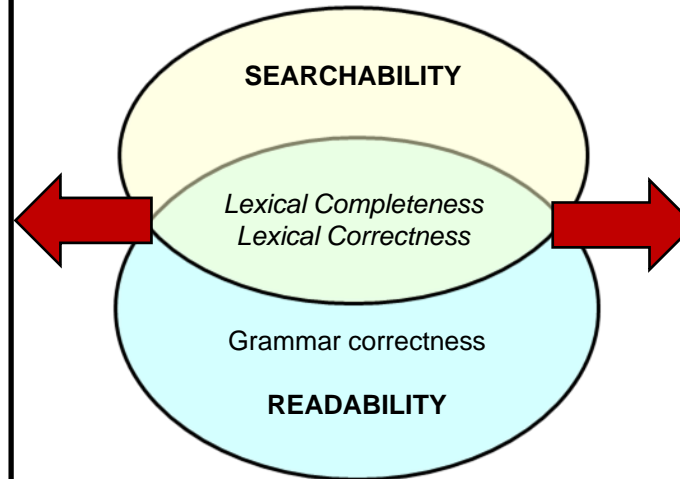
Assumption:

1 new word every 10 documents
=> around 450.000 term pairs

1 terminologist = 300 words/day
=> 1500 days needed (around 4 years)
=> 4 terminologist needed for one whole year

Costs:

0,50 Euro per solved terminology pair
+ terminology extraction/validation costs
+ import/coding costs
+ dictionary maintenance costs



SMT

Bilingual materials available

- Extract
- Sentence-align
- Train Translation Model (probabilistic mapping of equivalences between source and target words)
- Identify words not included in Translation Model
- Resolve unknown words

=> Automated alignment process
=> Human validation of aligned sentences (circa 3 months time)
=> Remaining unknown words are less than ¼ initial estimate (around 100.000) – require 3 linguists for 3 months
=> No morphological or semantic coding required

Our feasibility study points to SMT,
HOWEVER
readability and intelligibility are largely dependent on grammar (weaker point for SMT)

Hybrid solution would combine the strengths of the two approaches!

Test set: 30 (15 JP and 15 WO) sample files

Reference: English reference translation for calculation of BLEU, F-Measure and TER score.

Domain:

IPC Section (F)

IPC Section + Class (F23)

IPC Section + Class + Subclass (F23C)

Systems:

SMT 1, hybrid, ad-hoc developed patent-specific baseline engine

RBMT 1, baseline

RBMT 1, imported JPO dictionary, no coding

RBMT 2, patent-specific terminology routines

RBMT 3, patent-specific developed terminology and TM support

SMT 1, not customizable

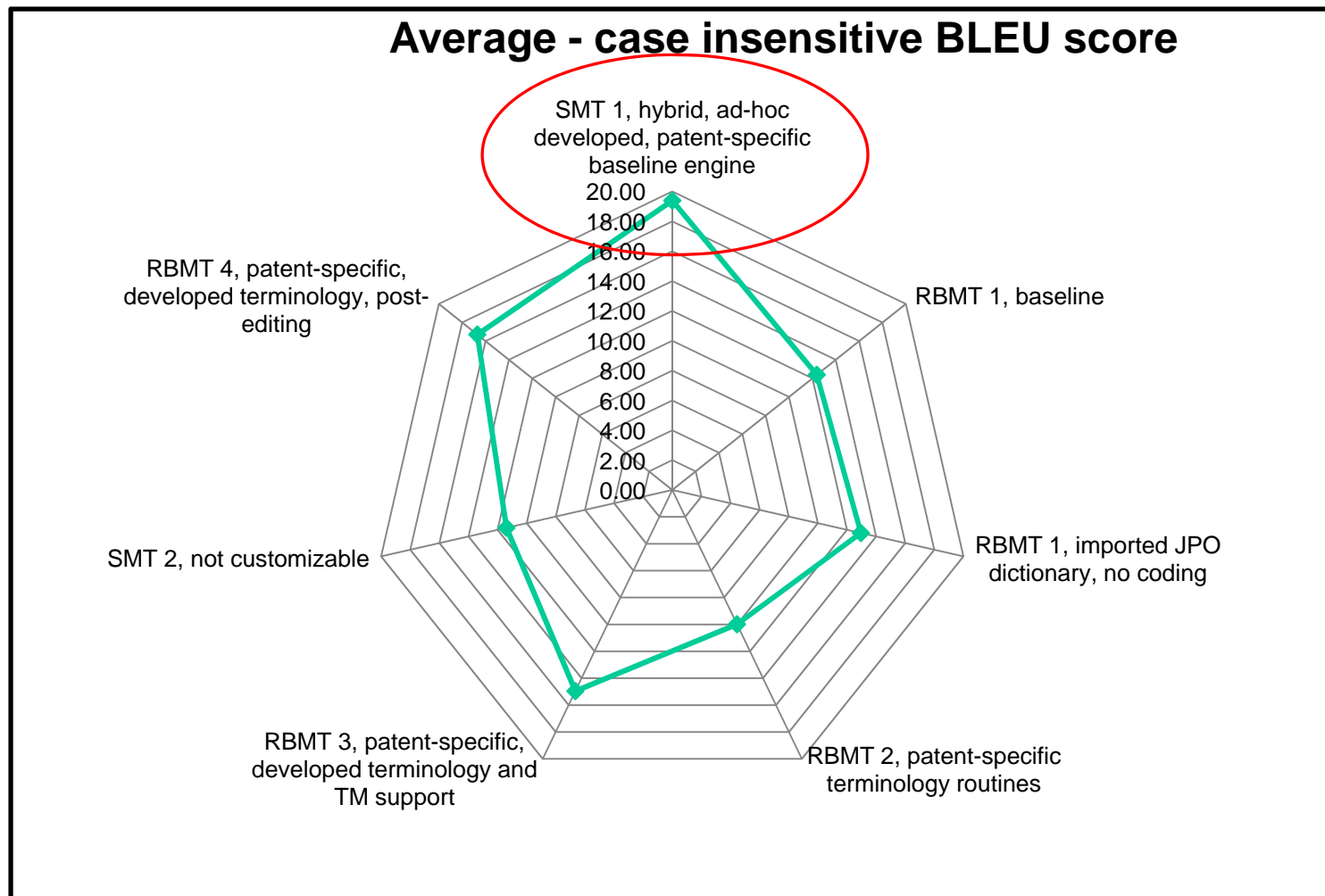
RBMT 4, patent-specific, developed terminology, post-editing

Challenges of the evaluation:

Heterogeneity of the systems

Transparency of the systems

Quality scores were calculated per subset (titles, abstracts, description, claims), per IPC category and as a general average and reflected human evaluation.



JP06173175A Original

【目的】 ポリエステル系繊維だけから構成されているにも拘らず、異色に染め分けられ、例えば玉虫効果や空効果、ジャガードの柄効果や表裏で色の異なるリバース効果などを表出するポリエステル系繊維製品、及びその無地染めによる染色方法を提供すること。【構成】 染着特性を異にする複数タイプのポリエステル系繊維から成る繊維構造物を用いて、分散染料およびまたはカチオン染料にて異色に染め分けるという手段を採用した。【効果】 系内で部分的に配向度等が異なるポリエステル系繊維を選択すれば、一本の系内で異色染めすることが可能になり、また要すれば、特殊断面型ポリエステル等の外見差異による効果との相乗効果も期待できる。

PAJ Abstract

PURPOSE: To prepare a polyester fiber product dyed in different colors in spite of the use of exclusively a polyester fiber and developing e.g. iridescent effect, grandrelle effect, jacquard pattern effect and reverse effect to give different colors on the obverse and reverse faces and to provide a dyeing method for dyeing the fiber product by plain dyeing.

CONSTITUTION: A fiber structure composed of plural types of polyester fibers having different dye abilities is dyed to different colors with a disperse dye and/or a cationic dye. When a polyester fiber having partially different orientation degrees in single fiber is used, the single fiber can be dyed to different colors. If necessary, a synergistic effect between this effect with a variety of colors and the apparent difference in a polyester fiber, etc., having a specific cross-section can be expected.

LexisNexis

Object: Polyester fiber comprises only nevertheless, sprinkly on two different color, for example eight grandrelle iridescent effect results, Jacquard pattern effect of different colors such as reverse effect at two sides and a polyester-based fiber product, as well as a method for dyeing according to the plain printing.

Constitution: Dyeing characteristics comprise a plurality of different types of polyester-based fibers using a fiber structure, dispersion dyes and/or dyed with a cationic dye at two different color.

Results: Partially within the yarn such as polyester fiber is selected degree of orientation is different in that case, only one within the yarn can be dyed, and if necessary, such as a polyester or a special cross-sectional appearance of a synergistic effect due to difference in effect can be expected.

Google

Objective: Despite being composed of only polyester fibers, are unique to 染Me分Ke, effects and heather effects Tamamushi example, polyester fiber products to reverse the effect of different expression and color jacquard pattern and the side effects, to provide a method of dye staining and its solid. [] Configuration for the fiber structure consisting of several different types of polyester fiber dyeing properties, was adopted by means of dye in different colors to unique cationic dye and disperse dye or. [Effect] if you select a different polyester fibers and yarns in a partial degree of orientation, will be able to dye yarn in a unique one, and if required also differences in appearance such as polyester-type special section can be expected synergies with effect.

JP2003178146A
Original

【課題】携帯電話の情報配信網を利用して、結婚式等の慶事及び通夜、葬儀等の弔事に関して、その場所や日時等の様々な情報をユーザが必要に応じて簡単に得ることができるようにする。【解決手段】慶弔情報提供システム(S)は、携帯電話の情報配信網(1)と、慶弔情報データベース(2)と、慶弔情報提供網(3)と、電報センター(4)とで構成されている。慶弔情報提供システム(S)は、慶弔情報提供網(3)から慶弔情報を提供するようにし、慶弔情報を慶弔情報データベース(2)へ蓄積する。携帯電話の情報配信網(1)のユーザは携帯電話の情報配信網(1)を介して慶弔情報データベース(2)から慶弔情報を取り出すことができ、携帯電話から祝電又は弔電を電報センター(4)に申し込むことができるようになっている。

PAJ Abstract

PROBLEM TO BE SOLVED: To enable information, which concerns the **place, time**, or the like of **happy** occasions such as a wedding and of **sorrowful** occasions such as a wake or funeral, to be acquired easily and as required by using a **cellular phone** data distribution network.

SOLUTION: A **celebration/condolence information providing system** S comprises a **cellular phone data distribution network** 1, a celebration/condolence information database 2, a **celebration/condolence information providing network** 3, and a telegraph center 4. The system S is designed to provide celebration/condolence information via the network 3 and to accumulate information in the database 2. A user of the network 1 can retrieve information from the database 2 and request the center 4 to dispatch a celebration/condolence telegram.

LexisNexis

PROBLEM TO BE SOLVED: An information distribution network is utilized as a **portable telephone**, a wedding and of **sorrowful**, such as related to the funeral condolences, various information such as the **date area** that a user can simply obtain as necessary.

SOLUTION: Condolatory information providing system (S) has, in the information **distribution network (1) and a portable telephone**, condolatory information database (2) and, the **information network condolatory** (3) and, the telegram center (4) comprises. Condolatory information providing system (S) well, condolatory information from the information providing network condolatory (3) so as to perform, condolatory information database (2) to store the information. An information distribution network (1) and the user can take out the portable telephone condolatory condolatory information database (2) via the information from the information distribution network (1) of a cellular phone, so that the telegram of condolence or service request message center (4) from the portable telephone.

Google

[Challenge] using a mobile information delivery network, such as funeral and wedding occasion, for the funeral and 弔事 can easily obtain the necessary information to the user a variety of **locations and dates**, etc. prevent. **[MEANS FOR SOLVING PROBLEMS] Information Systems for family events** (S), the distribution through **mobile phone information** (1) **information database** for family events (2), a **network of information for family events** (3) Center telegraph (4) is composed with that. Information system for family events (S), the information network for family events (3) to provide information for family events from the information database information for family events for family events (2) to accumulate. Mobile information delivery network (1) The user of a mobile information delivery network (1) through the information database for family events (2) can retrieve information from for family events, or a telegram of congratulations from the Center for cellular phone cable (4) has been able to sign up.

JP2007332877A Original

【課題】この発明は、内燃機関の制御装置に関し、燃料性状を簡単な構成で精度良く検出することを目的とする。【解決手段】燃料のセタン価の検出が可能な条件が成立しているか否かが判別される（ステップ100）。その条件の成立が認められる場合には、吸気弁52の閉じ時期を通常時より遅くする制御が行われる（ステップ102）。続いて、機関回転変動の大きさが検出される（ステップ104）。その機関回転変動の大きさに基づいて、セタン価が算出される（ステップ106）。上記の処理によれば、吸気弁52の遅閉じより、実圧縮比が低下するので、圧縮上死点近傍の温度および圧力が低下する。このため、低セタン価の燃料が機関回転変動に及ぼす影響を増幅することができる。よって、セタン価を高精度に検出することができる。【選択図】図7

PAJ Abstract

PROBLEM TO BE SOLVED: To accurately detect fuel properties through a simple structure in a control unit of an internal combustion engine. **SOLUTION:** Whether a condition where cetane value of fuel is detectable is established or not is determined (step 100). When establishment of the condition is determined, control is performed to set the closing timing of an intake valve 52 later than usual (step 102). The magnitude of engine rotation fluctuation is successively detected (step 104). A cetane value is calculated based on the detected magnitude of engine rotation fluctuation (step 106). According to the above processing, since the actual compression ratio is reduced by delaying the closing of the intake valve 52, the temperature and pressure at about the compression top dead center are reduced. Therefore, the influence of fuel of low cetane value on engine rotation fluctuation can be amplified. Consequently, the cetane value can be accurately detected. COPYRIGHT: (C)2008,JPO&INPIT

LexisNexis

PROBLEM TO BE SOLVED: To the present invention, a control device for an internal combustion engine, the fuel properties can be accurately detected with a simple constitution. **SOLUTION:** Cetane number of a fuel which is capable of detecting whether the conditions are satisfied (step 100) is determined. When the satisfaction of the conditions for, when the closing timing of the intake valve 52 is slower than the normal control is performed (step 102). Subsequently, the magnitude of engine speed variation (step 104) is detected. On the basis of the magnitude of the change of the engine rotation, (step 106) to calculate the cetane number. According to the above processing, inlet valve 52 of about reason, the actual compression ratio is decreased, the temperature and the pressure in the vicinity of the compression top dead center is lowered. Therefore, the fuel is a low cetane number influence on the engine speed fluctuation. Therefore, the cetane value can be detected with high accuracy. Selected drawing: Fig. 7

Google

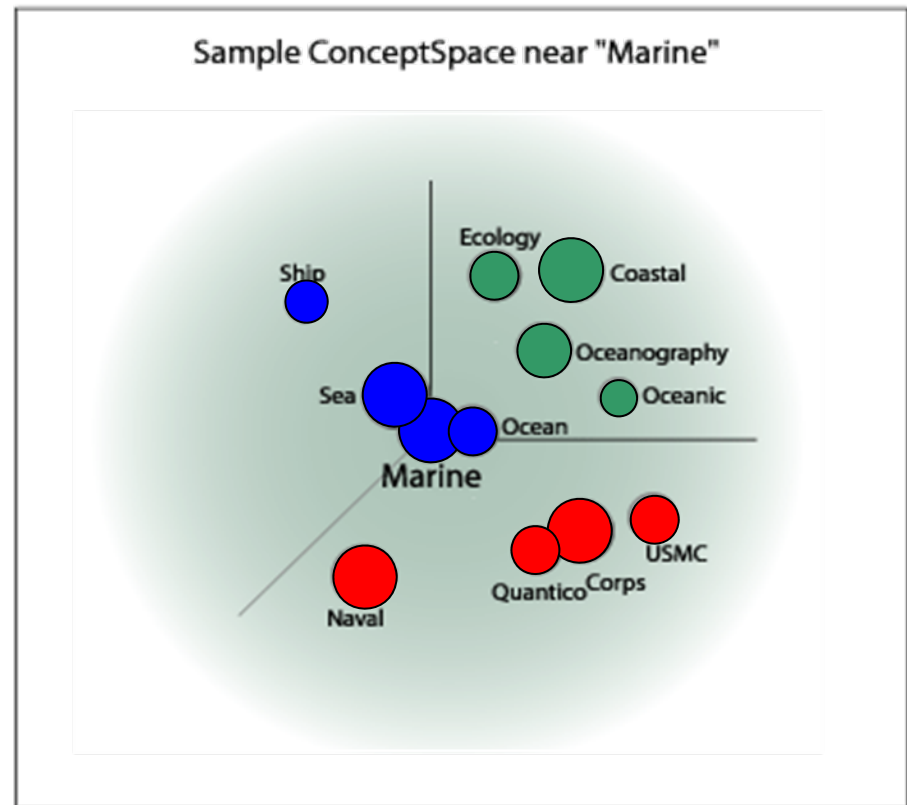
[Challenge] This invention relates internal combustion engine control unit, which aims to accurately detect the fuel property with a simple configuration. [] The remedies that determines whether conditions exist that can detect the fuel cetane number (step 100). When they found that the formation conditions, control is performed slower than normal closing time of the intake valve 52 (step 102). Then, the magnitude of change is detected engine revolution (104 steps). Based on the magnitude of the rotational body, the cetane number is calculated (step 106). According to the above process, 閉じより late intake valve 52, so the actual compression ratio decreases, lower temperature and pressure near the compression top dead center. This can amplify the effect of the rotational body low cetane fuel. So, can accurately detect the cetane number. Figure 7 - Selection Chart

What is Semantic Searching ?

Semantic Search or concept search technology provides users the ability to search with not only keywords, but concepts. Concept Search uses machine learning methods to surface meaning in documents while disambiguating semantic search queries.

- **Semantic Search Methods: Latent Semantic Analysis (or LSA) , Bayesian Inference and Support Vector Machines (SVM)**
 1. Analyzing statistical co-occurrence of terms and phrases to surface the relationship between words inside of a hyperspace.
 2. Using these relationships to search on concepts and the meaning of words, phrases or larger sets of textual input (sentences, paragraphs, etc.)
 3. Ranking by putting the corpus of documents into a model that ranks their content in relation to the search query, adjusted by the LSA, and then comparing them to one another.

- **LSA extracts every contextual relation among every word or phrase within a collection (learning document set). It then generates a vector space representation of all terms based on those relations. Within that space, proximity is a strong indicator of conceptual similarity.**
- **The result: similarities can be identified based on concepts within the material.**



Customer Challenges with Traditional Semantic Search: **THE BLACK BOX**
(Users **MUST** Surrender Transparency, Control and Scale)

TRANSPARENCY

- Semantic Search is effective and many times better than straight keyword searching, but thus far has locked users out of understanding **how** results are generated. We are forced to “trust” its algorithms, trust its function and trust its results. Virtually no system transparency.

CONTROL

- There is a surrendering of “control” with semantic search. Users cannot control how a query is constructed, much less the algorithms that create the search logic.

SCALE

- Semantic Searching requires a semantic index. Control of the search corpus via content indexing (and thus control over formatting, storage, etc.). This just does not scale. The web may never be semantically indexed.

The Result: *Traditional Semantic searching is not a practical search alternative to traditional search methodology. It may even be a liability.*

The LexisNexis Approach

LexisNexis shows the user **EXACTLY** what the Brains inferred from their initial query

LexisNexis allows the user to interact and change the inference, add/remove and change weighting of terms.

LexisNexis semantic search can be applied to patents and non-patent literature, hosted or on the web

User is back in control of the search

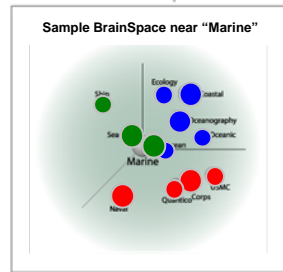
1) User builds Query

Semantic Query
 * Words
 * Sentence
 * Paragraph
 * Document,
 etc.

2) Brain is Selected



3) Intelligence/Inference: Query terms extracted and inferred terms from the brain are added.



4) Transparency: QueryCloud is generated. User is shown everything generated by the machine learning.

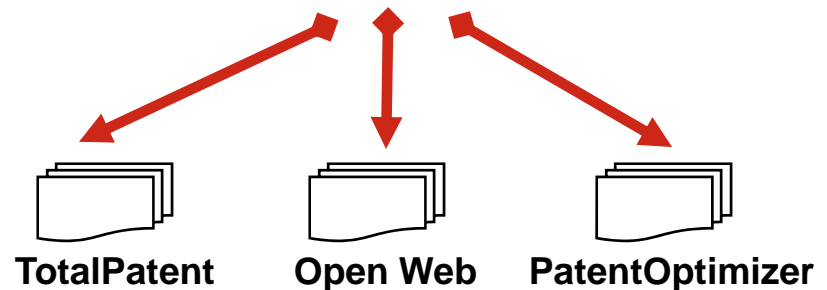
Total time elapsed = 1 second

5) Automatic Query Generation: The System uses the words extracted and generated by the brain to formulate fully optimized queries for each index being searched.

Query Creation/Optimization

- Simple Boolean
- Nested Boolean
- Ranked OR
- Term Weights

- Query 1:** lorem and (rutrum (.8) or voluptat (.74) or elit(.63) etc.
- Query 2:** lorem and ipsum and (rutrum or vlupat or elit etc.)
- Query 3:** lorum ipsum rutrum voluptat elit etc



Preview Semantic Terms and Results
How Do I...?

Change and re-analyze your search terms, click or drag a semantic term to change its status in your query, and see how your changes affect your results. [Learn More](#)

Your Search Terms

↑
↺ Regenerate Terms
↓

Terms Generated by Semantic Analysis

↶ Undo
20 of 20 terms used

+ Add

REQUIRED (Boolean "AND")

"mitral valve"

OPTIONAL (Boolean "OR")

valve
mitral
atrioventricular

regurgitation
leaflets
chordae

"papillary muscles"
tendineae
tendinae

papillary
"chordae tendineae"

"chordae tendinae"

annulus
tricuspid
repair
subvalvular

"posterior leaflet"
"tricuspid leaflets"
prolapse

EXCLUDED (Boolean "NOT")

HOLDING AREA (Not included in your search)

tendonae	posterior	"tricuspid annulus"	"anterior leaflet"
"ventricular outflow tract"	insufficiency	"leaflet coaptation"	
replacement	"heart repair"	cordae	anterior
"ventricular outflow"	"native aortic"	"tricuspid repair"	diseased
"outflow tract obstruction"	coaptation	"aortic leaflets"	
"valvular insufficiency"	function	"aortic annulus"	

Your search terms and their priority in results:

REQUIRED: ("mitral valve"[H]) **OPTIONAL:** "valve"[H], "mitral"[H], "atrioventricular"[H], "regurgitation"[M], "leaflets"[M], "chordae"[M], "papillary muscles"[M], "tendineae"[M], "tendinae"[M], "papillary"[M], "chordae tendineae"[M], "chordae tendinae"[M], "annulus"[L], "tricuspid"[L], "repair"[L], "subvalvular"[L], "posterior leaflet"[L], "tricuspid leaflets"[L], "prolapse"[L]

Most relevant **20** of **2434** results: Retrieve All Results Cancel

US20090043381A1 2009-02-12 Atrioventricular valve annulus repair systems and methods including retro-chordal anchors

English Abstract:
Methods, devices and systems are disclosed to treat atrioventricular valve regurgitation accessed through the vasculature, and by standard and minimally invasive surgical techniques. Isolated leaflet fixation and annulus treatment systems are developed. [More](#)



US7635386B1 2009-12-22 Methods and devices for performing cardiac valve repair

English Abstract:
The present invention is directed to methods and devices for repairing a cardiac valve. Generally, the methods involve a minimally invasive procedure that includes creating an access in the apex region of the heart through which one or more instruments may be inserted so as to repair a cardiac val... [More](#)



US20090177274A1 2009-07-09 DEVICE FOR REPLACING THE CHORDAE TENDINEAE OF AN ATRIOVENTRICULAR VALVE

English Abstract:
A device (10; 10'; 10"; 10'") for replacing the chordae tendineae (CT) of an atrioventricular valve comprises a cuspidal end (10a; 10a'; 10a"; 10a'41), able to be fixed along a portion (C1P) of the free edge (FM) of a cusp (C1; AC1) of the atrioventricular valve, and an apical end (10... [More](#)



US20090177266A1 2009-07-09 METHODS, SYSTEMS AND DEVICES FOR CARDIAC VALVE REPAIR

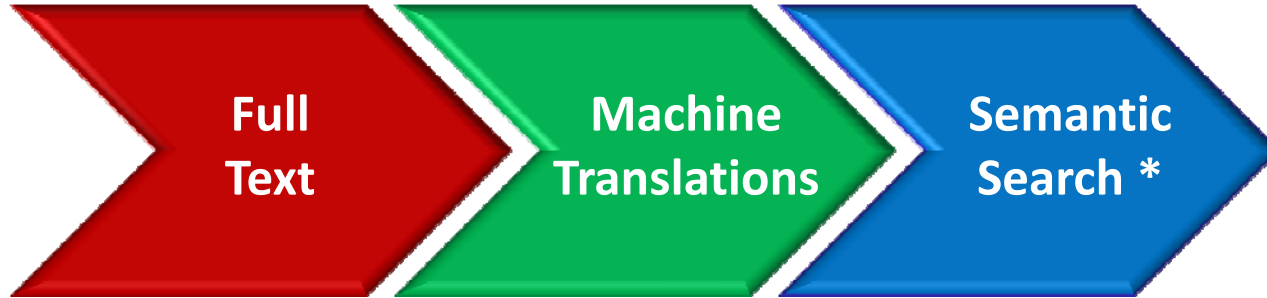
English Abstract:
Disclosed are methods, systems, and devices for the endovascular repair of cardiac valves, particularly the atrioventricular valves which inhibit back flow of blood from a heart ventricle during contraction. The procedures described herein can be performed with interventional tools, guides... [More](#)



Why this is important

- Exposes “intelligence” (complete transparency) of a massive corpus of learning content for application to the search process (full text of US patent Db + millions of Elsevier scientific journals)
- Gives searcher complete control over how the intelligence is used
- Helps the non-native speaker take advantage of deep learning and insights on related concepts that can help their search process
- It can find relevant documents that might otherwise be missed
- It can rank the documents, raising the most relevant to the top of the list and thus saving time in culling the answer set

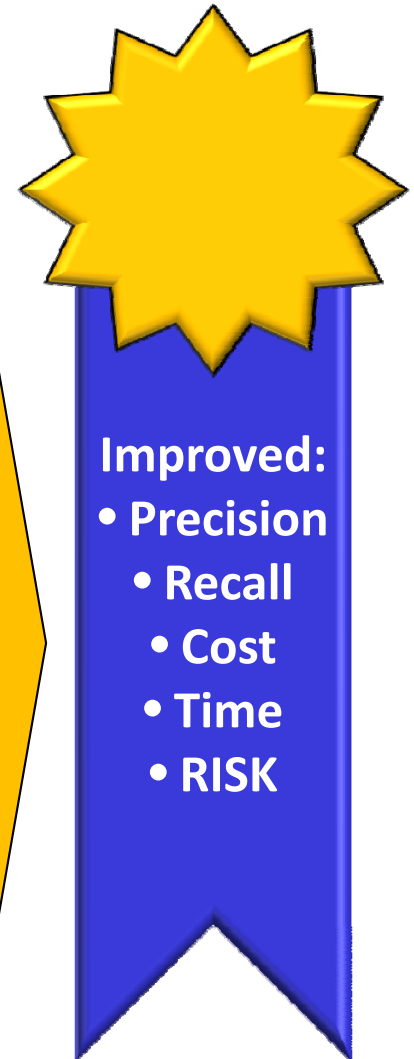
+ Professional User Experience!!



- Search where you're most likely to find the answer
- Avoid higher cost, less comprehensive A/I
- Improve chances of finding key elements
- Use current information available rapidly upon publication

- Search multiple collections simultaneously
- One query
- State of the art
- Continuous improvement
- Focus on lexical correctness

- Deep insight into knowledge base
- Extra helpful to a non-native speaker
- Enable weighted Boolean queries
- Complete transparency
- Complete control
- Ranking of results



* As an additive tool to today's expert Boolean search techniques

- Big problems require creative solutions – our work demands it of us
- Vendors have to be bold in their efforts & investments, always pushing the envelope of what's possible
- Users have to be curious, creative and adventuresome in their use of new tools. They MUST experiment – accepting that the perfect solution does not exist but that improvement is possible and should be leveraged as it is made
- When we work together (vendors, users, governments, WIPO) we move our industry forward, accelerate the development of meaningful technologies and improve the lives of all people

Thank You!