Overcoming Language Barriers in Patent Research: A LexisNexis® White Paper

Adapted from a presentation made by Mr. Peter Vanderheyden, vice president and managing director of Global Intellectual Property at LexisNexis, to a meeting of the World Intellectual Property Organization (WIPO) in September 2010

Today’s global marketplace—where all of the world’s major economies are interdependent—brings with it equally large baskets of benefits and challenges. One of these challenges is the fact that large corporations often compete with each other in the same industries, but the day-to-day communications within their businesses frequently take place in different languages.

This is especially significant when it comes to the international effort to protect intellectual property. How can U.S.–based companies effectively assert or enforce IP rights, for example, if they are unable to accurately determine which patents are on file in a foreign jurisdiction that operates with a language other than English—or, for that matter, uses an entirely different alphabet? What good does it do a legal professional in the U.S. to ascertain all of the English-language patent records in Asian countries if he or she is unable to determine the meaning of patent records from China that are entirely composed in Mandarin Chinese?

This core challenge—overcoming language barriers in patent research—is the subject of this white paper. It will explore the essence of the language barrier problem, the perspective that we bring to the issue at LexisNexis, and then offer our three-pronged solution to the problem.

Our goal with this white paper is to shine some light on the serious challenge of how we overcome language barriers in patent research and to point the way to what we believe are some exciting technological advances that make this elusive solution now more attainable than ever.
The Language Barrier Problem

There are, of course, a variety of languages used in the creation of patent records and technical documentation. At a very basic level this creates a significant IT challenge as the varying languages make it difficult to integrate the documents from these disparate countries into a single online database, particularly given the non-uniform alphabets and the use of different characters in the world’s languages.

Moreover, the skill set of the average patent researcher is not conducive to effective search and retrieval of patent records in multiple languages. The fact is that most patent researchers simply can’t be familiar with the vocabulary of other languages, which is especially critical in the world of patent literature. The problem is that patent records and other prior art are important, regardless of their place of origin and the language in which they were composed. We simply don’t have the luxury of ignoring or bypassing documents that aren’t written in English or that contain characters not included in the Western alphabet.

In order to effectively do their jobs in today’s interdependent global economy, patent researchers must have access to relevant patents, patent applications and corresponding patent documents—without running into the frequent stop sign presented by language barriers. In fact, as searchers experience growing accessibility to these previously inaccessible collections of patent records, it will accelerate the need for everyone to be working from the same library of documents worldwide.

The LexisNexis Perspective

Based on lessons learned from nearly 30 years in this business, there are five specific user benefits that we strive to achieve with every technology investment and each next-generation research offering related to intellectual property:

- **Precision**—our approach is to help researchers improve the rate with which they are able to identify the specific documents that relate most precisely to their search queries;

- **Recall**—we also place an emphasis on making sure that researchers are able to find every possible document that is relevant to their search;

- **Cost**—we focus on developing research platforms that can be used by customers in the most cost-efficient manner and yield a measurable value to the end user;

- **Time**—our goal is to put tools into the hands of researchers that are ready to help them cultivate the specific information they need in the least amount of time possible; and

- **Risk**—our belief is that patent research is ultimately about minimizing risk and that conviction shapes our entire approach. Anything that can reduce the risk of missing relevant documents is interesting to us.

It’s our belief that there is no single perfect solution or across-the-board panacea for the challenge of overcoming language barriers in patent research. This is a journey that we intend to lead. To that end, we believe that a three-pronged approach to IP research can deliver the most professional user experience and best meet the complex needs of today’s patent researcher.
A Three-Pronged Solution

Patents and patent records are ultimately about detail—this is one area where miscellaneous words contained in a document aren’t scanned through quickly; they are crucial pieces of information. Those critical details are only found in full-text searches of documents.

In the world of online research, full-text search refers to the use of a search engine that examines all the words in every stored document as it tries to match search words supplied by the user. Full-text searching techniques became common in online research back in the 1970s while bibliographic searching was the only methodology available to patent researchers for many years.

Recent research indicates that the 95 percent of the entities found in the full-text section of patent documents (claims and description) are unique to these sections, i.e., not found in the title and abstract, as illustrated in the graphic on the following page. So if patent researchers want to search where they’re most likely to find the answers they need, they’ll start with full-text. This approach avoids some of the higher costs and time delays that can come with more complex searching methodologies based on custom and less comprehensive Abstract/Index databases.

Full-text documents are now accessible much faster than at any time since the dawn of online patent research, which enables researchers to consider more current information in their assignment. Patent records are now posted into databases rapidly upon publication, with LexisNexis providing access to our IP customers within hours of their release around the world.

Moreover, there are more documents available in full text than ever before. Eleven patent authorities worldwide now publish some form of full-text patent records. Beyond the patent offices, LexisNexis provides access to full-text patent records from 20 additional countries, including Canada, Italy, Netherlands, France, Portugal, Sweden and many more. LexisNexis now offers patent researchers access to more than 38 million full-text records from 30 different patent authorities worldwide, the widest catalog of full-text documents available in the world.

Machine Translation

The principal way that the international IP community has sought to overcome language barriers in patent research is with the use of Machine Translation (sometimes known as just MT). Machine translation is simply a process in which a computer program analyzes a source text and then produces a translated text without human intervention.

With careful preparation of the search query and the source text, commercial machine translation tools can produce very useful results for patent researchers. They allow users to search multiple collections of patent records simultaneously with just a single search query.

There are actually two primary types of machine translation that are used in the marketplace: Rules-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT). The nuances of each type of machine translation means that researchers will experience better results from RBMT at times and SMT will be superior at other times. For example, RBMT tends to have better precision on grammar, but SMT tends to provide the most extensive language system. Given this, there is no “one-size-fits-all” MT solution. Rather, each language pair needs to be assessed within a content domain. For example, our feasibility research into machine translation systems for Japanese to English MT revealed that SMT produced the best search results, but the grammar advantage of RBMT was often of great consequence in the results we obtained.

A recent patent database study (H. Cunningham, A. Hanbury, and S. Rüger: IRFC 2010, LNCS 6107, pp. 161, 2010. © Springer-Verlag Berlin Heidelberg 2010) confirmed that the details are found in the full text. Among the findings of the study:

- “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.”
- “Only up to four percent of all entities are unique to the claim section in comparison to the description section. At least 79 percent of all entities are unique to the description section in relation to the claim section.”
- “… between 92 percent and 97 percent of the entities are unique to the sections claims plus description.”
The obvious solution was to combine the strengths of each approach into a single hybrid solution for machine translation of this language pair. As the chart below indicates, the proof is in the pudding: a hybrid approach of RBMT and SMT machine translation approaches yielded the highest-quality search results, markedly better than either one on their own.

**Average—Case Insensitive BLEU Score** (See sidebar for BLEU description.)

The point is that “one size” of machine translation does not “fit all” patent research needs. LexisNexis is committed to finding the right solution for each language pair and to improve the system as new developments are made.
Why semantic searching is important:

- Exposes “intelligence” (complete transparency) of a massive corpus of learning content for application to the search process (full text of U.S. 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; and
- It can rank the documents, raising the most relevant to the top of the list and thus saving time in culling the answer set.

By deploying the power of semantic search technology, LexisNexis leverages the intelligence from millions of full-text technical publications, organizes that intelligence, and then makes it accessible to the experience and expertise of the searcher. The result is a unique and valuable tool for patent and non-patent literature.

The New World of Semantic Search

There is a third dimension to the LexisNexis approach that offers a glimpse into the future of patent research: the ability to search with concepts, not just words or phrases.

Semantic search is a process used to produce search results of similar documents based on the relationship between concepts contained within the material. It’s an approach best used when you’re trying to locate a number of documents that will help you locate the information you want. Rather than use ranking algorithms (e.g., Google™ methodology) to predict relevancy, semantic search uses the science of meaning in language (“semantics”) to produce highly relevant search results.

Traditional semantic search solutions suffer from Black Box syndrome—that is, they can’t reveal what it going on within the search algorithm and thus they require the user to simply “trust” they are getting a good result. Further, they can only search content indexed by the semantic engine.

The new LexisNexis® semantic search solution accomplishes three breakthrough objectives in online search, all aimed at giving the user new insights and control over how they are used in the search process:

1. It is transparent: No Black Box here. Each query is enhanced by the machine intelligence and the inferred terms/relationships are shown to the user for their complete understanding.

2. It gives control to the user: Not only is the semantic search transparent, but users are given the ability to add, delete, increase or decrease the importance of all query words in a unique visual query interface.

3. It utilizes the most scalable semantic technology in the world: The semantic query can be applied against any number of indexed content sets, within LexisNexis (patent and non-patent scientific information) and even on the Web. We started building the LexisNexis semantic intelligence with the USPTO patent index of more than 7 million documents and millions of Elsevier journal articles. Unlike other semantic technologies that must sample smaller sets of documents for learning, the LexisNexis semantic index includes the learning from more than 10 million documents and counting. This vast body of intelligence can now be applied to multiple data sets without further indexing.

Why semantic searching is important:

- Exposes “intelligence” (complete transparency) of a massive corpus of learning content for application to the search process (full text of U.S. 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; and
- It can rank the documents, raising the most relevant to the top of the list and thus saving time in culling the answer set.

By deploying the power of semantic search technology, LexisNexis leverages the intelligence from millions of full-text technical publications, organizes that intelligence, and then makes it accessible to the experience and expertise of the searcher. The result is a unique and valuable tool for patent and non-patent literature.

The New World of Semantic Search

There is a third dimension to the LexisNexis approach that offers a glimpse into the future of patent research: the ability to search with concepts, not just words or phrases.

Semantic search is a process used to produce search results of similar documents based on the relationship between concepts contained within the material. It’s an approach best used when you’re trying to locate a number of documents that will help you locate the information you want. Rather than use ranking algorithms (e.g., Google™ methodology) to predict relevancy, semantic search uses the science of meaning in language (“semantics”) to produce highly relevant search results.

Traditional semantic search solutions suffer from Black Box syndrome—that is, they can’t reveal what it going on within the search algorithm and thus they require the user to simply “trust” they are getting a good result. Further, they can only search content indexed by the semantic engine.

The new LexisNexis® semantic search solution accomplishes three breakthrough objectives in online search, all aimed at giving the user new insights and control over how they are used in the search process:

1. It is transparent: No Black Box here. Each query is enhanced by the machine intelligence and the inferred terms/relationships are shown to the user for their complete understanding.

2. It gives control to the user: Not only is the semantic search transparent, but users are given the ability to add, delete, increase or decrease the importance of all query words in a unique visual query interface.

3. It utilizes the most scalable semantic technology in the world: The semantic query can be applied against any number of indexed content sets, within LexisNexis (patent and non-patent scientific information) and even on the Web. We started building the LexisNexis semantic intelligence with the USPTO patent index of more than 7 million documents and millions of Elsevier journal articles. Unlike other semantic technologies that must sample smaller sets of documents for learning, the LexisNexis semantic index includes the learning from more than 10 million documents and counting. This vast body of intelligence can now be applied to multiple data sets without further indexing.
Peter J. Vanderheyden  
**Vice President & Managing Director, Global Intellectual Property, LexisNexis**

As Vice President and Managing Director, Global Intellectual Property for LexisNexis, Peter Vanderheyden brings more than twenty-five years of leadership in the industry. In his role, Mr. Vanderheyden is responsible for the growth and development strategy of the LexisNexis services and solutions for intellectual property professionals.

Prior to LexisNexis, Mr. Vanderheyden held a number of leadership positions in the industry including vice president of marketing and business development for IP.com and consulting engagements helping companies develop strategic business plans for both start-up and product re-launched businesses.

Mr. Vanderheyden spent more than 16 years with IBM in a variety of executive positions including serving as a solutions development executive in the government sector. During this time, he founded the Delphion Company as a spin-off from IBM focusing on an innovative and efficient method for researching intellectual property information.

He is a published author of journal articles and has spoken at industry conferences on intellectual property.

What makes the LexisNexis semantic search experience so unique is that we have reinvented semantic search to occur NOT as a transaction, but rather as a conversation between the semantic brain and the user of the system. It engages the user in a search experience that encourages interaction with the machine intelligence. To make this possible, LexisNexis had to create the interaction as a visual experience. We accomplished this by transforming each query into what we call a QueryCloud, a visual representation of the newly generated semantic query.

As the graphic on the previous page illustrates, QueryClouds create the “visual workbench” that allows users to transform a query into a cloud showing the top-20 most-relevant words and phrases (concepts) in varying font sizes. By increasing or decreasing the font size of terms and phrases (drag and drop), and even adding and subtracting words and phrases, the user is completely engaging the artificial intelligence to focus in on highly relevant results.

**Conclusion**

The global economy in which we all participate creates special challenges for those of us who are charged with helping protect patents and other forms of intellectual property. We not only have the routine duties of conducting thorough online research, but we have a special challenge of incorporating documents written in foreign languages—and often with entirely different alphabets—than our own.

This is a big challenge—and big challenges require big solutions.

Providers of information services, such as LexisNexis, need to be bold in our efforts and investments; doing technical due diligence and investing selectively when we believe we can reduce risk for our customers. Researchers need to be curious and creative, showing a willingness to experiment with new tools in order to find the right combination of approaches that will yield the best results.

The message from LexisNexis is this: We’re committed to finding and combining the right tools and technologies to help our customers leverage their own intellect and experience to deliver a better outcome. Professional user expertise combined with our three-pronged approach is truly a winning combination.

The best way to overcome language barriers in patent research is to give researchers the tools they need and turn them loose to find the approach that works best for each unique research challenge.

---

**About LexisNexis**

LexisNexis® ([www.lexisnexis.com](http://www.lexisnexis.com)) is a leading global provider of content-enabled workflow solutions designed specifically for professionals in the legal, risk management, corporate, government, law enforcement, accounting and academic markets. LexisNexis originally pioneered online information with its Lexis® and Nexis® services. A member of Reed Elsevier [NYSE: ENL; NYSE: RUK] ([www.reedelsevier.com](http://www.reedelsevier.com)), LexisNexis serves customers in more than 100 countries with 18,000 employees worldwide.

---

The opinions expressed within this case study represent customer opinions. LexisNexis believes this case study experience generally represents the experience found with other similar customer situations. However, each customer will have its own subjective goals and requirements and will subscribe to different combinations of LexisNexis services to suit those specific goals and requirements. This case study may not be deemed to create any warranty or representation that any other customer’s experience will be the same as the experience identified herein. LexisNexis uses the customer’s trademarks herein with the customer’s permission.

All information provided in this document is general in nature and is provided for educational purposes only. It should not be construed as legal advice. For legal advice applicable to the facts of your particular situation, you should obtain the services of a qualified attorney licensed to practice in your state.

LexisNexis, Lexis, Nexis, lexis.com and the Knowledge Burst logo are registered trademarks of Reed Elsevier Properties Inc., used under license. TotalPatent and PatentOptimizer are trademarks of LexisNexis, a division of Reed Elsevier Inc. REEDFAX is a registered trademark of Reed Technology and Information Services Inc. Other products or services may be trademarks or registered trademarks of their respective companies. © 2010 LexisNexis, a division of Reed Elsevier Inc. All rights reserved. NXC02483-0 1210