The Revolution in Patent Information Publishing: Making International Full-Text Databases a Reality

By Richard Garner
Product Director, IP Research Solutions
LexisNexis

LexisNexis®

Introduction

As with many areas of life in the Information Age, over the past 25 years there has been nothing less than a revolution in the field of Intellectual Property management. The rapid growth in the availability of patent records from increasingly sophisticated databases has been a key driver in this revolution by enabling business executives and IP attorneys to be more precise in the way they manage their IP portfolios.

At the same time, some patent issuing authorities around the world have been slow to join the revolution. Of the 185 World Intellectual Property Organization (WIPO) member states, it is estimated that less than 20 percent publish regular updates of their full-text patent data. Fewer still have published their complete back files. The result of this disparity between authorities that are driving the train toward global full-text databases and those that are lagging behind is that the patent information publishing industry is still dependent on image data to create searchable full-text data.

The purpose of this white paper is to provide some context for better understanding the different approaches being used to solve the industry conundrum of converting image data to searchable text. We will take a walk back through history in order to shine light on how we got to this point in time, explore the various systems and processes currently being used to build databases of enhanced first-level patent data, discuss the importance of establishing a true international full-text patent database in the marketplace, and offer some insights on how to carry on the revolution with the aid of machine translation.

A Walk Through History

Patent abstracts first appeared online when SDC Information Services (Orbit) loaded the Derwent World Patents Index file from 1976, but subject-based text was usually limited to the title of the record. There were no full-text databases that existed in machine-readable form and the cost of computer storage made it prohibitively expensive even to go down this road.

Things first began to change when LexisNexis (formerly Mead Data Central) launched Lexpat® in March 1983. This was the first example of a full-text patent database, populated with U.S. patents that had been granted for the period from 1976 onwards. Each year's specifications were loaded into a separate file, which generated enormous volumes of available data. The slow search speeds available at that time made analyzing the documents on-screen relatively easy, as the text literally built up as you watched.

By the mid-1980s, information systems had improved sufficiently to be able to add clipped images from the drawing pages of patent specifications. This was quite a breakthrough at the time for IP professionals, but its utility was modest since there were still very limited means for viewing them online.

Then in June 1986, the Administrative Council of the European Patent Organisation (EPO) awarded a contract for the capture of the complete PCT minimum documentation—contained in the EPO search collection —to a consortium of three firms. The BACON (BAckfile CONversion) project covered the digitization and capture of the complete EPO back file.

This historic joint initiative between the USPTO, JPO and the EPO involved scanning the full text and drawings of the first publication of patent documents by the major industrial property offices, back to 1920 or earlier. Then they loaded the image data, in facsimile form, onto magnetic tape so as to permit the subsequent creation of an image retrieval system.

The first phase of the BACON project (1920 –1987) comprised 125 million pages and resulted in 12 million documents stored on 65,000 magnetic tapes. The project took roughly three years to complete and produced roughly 13 terabytes of data. This backfile data was made available to the examiners and other national offices for their internal use, and was eventually disseminated for public information.

The BACON Project provided a much-needed boost to the movement toward the construction of a full-text database of patent records. Unfortunately, there was minimal progress in the industry for the next two decades—a time many of us have come to know as The Wilderness Years.

Systems and Processes to Build Databases

In the last several years, the pace of innovation in the patent information space has accelerated beyond what any IP professional could have imagined. We now have state-of-the-art databases providing access to full-text patent records from patent authorities around the world. These databases are tapped by researchers using intuitive software applications that often feature powerful search engines.

The vexing problem still facing our industry, though, is a challenge that civilization has confronted for millennia: language translation. The simple fact is that a variety of languages are used in the creation of patent records and technical documentation. How do we build patent databases that integrate complex documents from countries all over the world into a single online resource, particularly given the non-uniform alphabets and the use of different characters in the world's languages?

There are a couple of complicating factors that make this challenge even more daunting. First, most patent researchers simply can't be expected to be familiar with the vocabulary of languages from multiple continents and cultures, which is especially critical in the world of patent literature. Second, patent records and other prior art are important in courts of law, regardless of their place of origin and the language in which they were composed, so it's not an option to conveniently bypass documents that aren't written in English or that contain characters not included in the Western alphabet.

Patent information vendors have put in place competing systems and processes to address this crucial need of translating patent records from their native languages into the global business language of English. Some have addressed the challenge with lots of human resources, other have approached it with technology.

Manual Translation

Manual translation is the translation of documents from one language to another by humans through a manual "one sentence at a time" means. The job is done by experienced professionals and native speakers who convert the original documents into meaningful and accurate texts in the desired language. The primary benefit of manual translation is that grammatical and structural errors can be kept to a minimum. However, the most commonly cited problem with manual translation is that it is substantially more expensive as it relies on human capital to perform the work.

Machine Translation

Machine translation is simply a process in which a computer program analyzes a source text and then produces a translated text in another language, all without human intervention. The software program has its own set of rules for delivering the translation, and the analysis of the text and grammatical structure are completed by this artificial intelligence. The principal benefit of machine translation is its speed—which is now almost instantaneous. The primary objection to machine translation is the potential for uncertainties in translation without human involvement in the process.

An International Full-Text Database

In the world of online research, full-text search refers to the use of a search engine that examines all of 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, yet 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 and not found in the title and abstract. So if patent researchers want to search where they're most likely to find the answers they need, they'll need access to full-text records.

There are a number of important user benefits to the construction of an international database of full-text patent records:

1. **Simplicity**: any user who has experimented with a simple search engine is able to make sense of full-text searching approaches.
2. **Cost-efficiency**: the overheads involved in creating new abstracts and indexing make it possible to deliver a full-text database at more affordable prices.
3. **Reduction in indexing errors**: with little or no manual indexing in a full-text database, the scope for database construction errors is reduced.
4. **Comprehensive retrieval**: full-text retrieval draws from the entire disclosure and ensures a much better recall of unusual terms that may only occur rarely in a document.
5. **Fewer skills to maintain**: with full-text searching, there is no need for the searcher to be continually updating their knowledge of subject heading lists or similar controlled vocabulary.

The good news is that full-text patent records are now accessible much faster than in the past, which enables researchers to consider more current information in their assignment. Patent records are now posted into databases soon after publication, with LexisNexis providing access to our IP customers within hours of their release around the world.

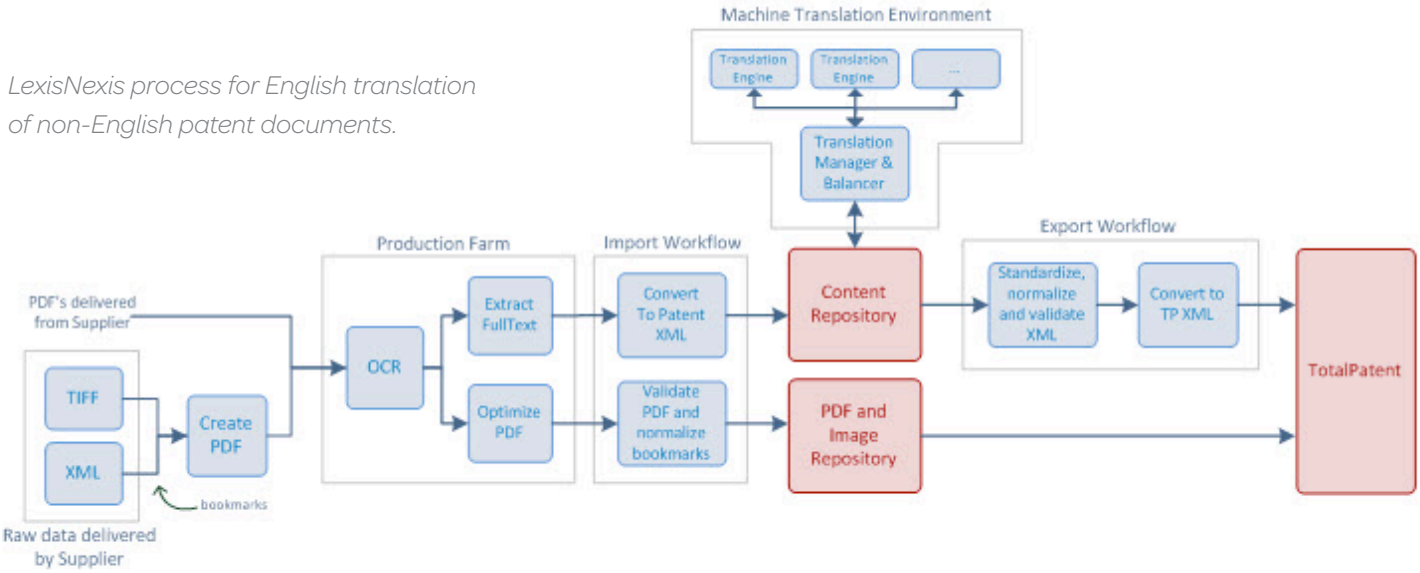
Moreover, there are more documents available in full text than ever before, with 31 patent authorities worldwide now publishing some form of full-text patent records. Beyond the patent offices, LexisNexis provides access to full-text patent records from a number of additional countries, including Italy, Netherlands, Portugal, Sweden, Russia, Asian countries and many more.

First, technology makes the translation of international patent records much less costly than human translation. By using advanced software and sophisticated computer networks to do the work, all sorts of possibilities become feasible that would otherwise simply be cost-prohibitive with human capital.

Second, technology delivers translation of patent records in a timelier manner. Machine translation is significantly faster than human translation, which makes it possible for international patent records to be translated into English and uploaded to a database on a 24-hour turnaround basis.

Third, technology is much more scalable than human translation. Machine translation has made it possible for the patent information industry to provide international patent records all the way back to the 19th century in some cases, a historical backfile that would not be achievable for years to come if we relied exclusively on human translation.

LexisNexis process for English translation of non-English patent documents.



Driving the Revolution Forward with Machine Translation

The principal way the international IP community has sought to overcome language barriers in patent research is with the use of machine translation. With careful preparation of the search query and the source text, commercial machine translation tools can produce very useful results for patent researchers. Moreover, they allow users to search multiple collections of patent records simultaneously with just a single search query.

There are three central considerations that we believe make machine translation the right technology to drive the patent information revolution forward.

Customers want help from vendors in finding a more comprehensive result by translating "foreign" full text to English. There are a few ways that the industry is working to meet this expectation:

- Getting the concepts right by training translation engines with IP and other scientific material
- Getting the language right by using patent records to create the language model
- Fine-tuning the search engines by "manually" adding new words
- Implementing a continuous cycle of measuring, training, improving software and re-translating the database

Conclusion

Business executives and IP attorneys need access to relevant patents, patent applications and corresponding patent documents from jurisdictions all over the world. Too often, their research pursuits are impeded by language barriers.

At LexisNexis, we've been at the forefront of patent information research from the earliest days of the industry. LexisNexis was the first company to put USPTO full-text records into an online database environment back in 1983 and we blazed new trails in patent research in 2005 by creating a standalone Intellectual Property business unit. Since that time, we've rolled out a number of new products and have now established ourselves as the world's largest provider of full-text patent records in a single online database.

Our track record has taught us that the best way to overcome challenges in patent research is to give customers the tools they need to be successful and then support them as they bring their expertise to bear in the search process. We believe that technology will once again be the key in continuing the revolution in patent information publishing, this time by using machine translation to make international full-text patent databases a reality.

About LexisNexis:

LexisNexis® (*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*), LexisNexis serves customers in more than 100 countries with 15,000 employees worldwide.

LexisNexis®