Coalition for Patent and Trademark Information Dissemination

September 17, 2012

Honorable David J. Kappos
Undersecretary of Commerce for Intellectual Property and Director of the US Patent and Trademark Office
600 Dulany Street
PS Box 1450
Alexandria, VA 22313-1450
Submitted via: fee.setting@uspto.gov

Dear Under Secretary Kappos:

We appreciate the opportunity to express preliminary views of behalf of the Coalition for Patent and Trademark Information Dissemination (CPTID) on the proposed fee schedule published by the USPTO**.** **(**http://www.gpo.gov/fdsys/pkg/FR-2012-09-06/pdf/2012-21698.pdf)

The Coalition for Patent and Trademark Information Dissemination (CPTID) would like to take this opportunity to voice its concern that whatever ultimate fees are decided upon by the agency, that such fees do **not** lead to deterioration in the quality of the USPTO's published patent information products.

The CPTID is a group of entities committed to the notion that private sector participation is essential to the quality and integrity of the US patent and trademark system. The CPTID believes that the US patent and trademark system depends on high quality raw data from the USPTO as well as the dissemination of value-added information.

Both the production of high quality raw data as well as the dissemination of value-added information can best be achieved by a public-private partnership that takes advantage of the core strengths of private sector vendors and publishers. A competitive private sector patent and trademark information industry complemented by the USPTO provides the optimal approach for meeting the broad range of user needs—from specialists to the general public.

I. During Consideration of the Appropriate Fees for Patent Application, the USPTO Should Ensure that It will Maintain the Current High Quality of Text Accuracy in its Published Raw Data

The CPTID agrees that the current IT systems at the USPTO cannot keep up with the demand of today's patent applicants and that the systems need upgrading.

It is important to our companies and organizations, as well as our clients and customers – ultimate users of the patent information - however, that while upgrading the IT systems, the USPTO does not lower the quality of the data. If USPTO captures its data with errors, converts complicated content types such as chemical structures to images rather than searchable XML or creates burdensome XML requirement on patent applicants then the potential result could be the reduction in the current high quality of patent data coming out of the agency.

a. What has raised the Coalition's Concerns about possible degrading of Patent Data at the USPTO

During PPAC and other public meetings, the agency has indicated that its information technology initiatives fall into two related but distinct categories.

First, the agency is working on its PATI ("Patent Application Text Initiative") technology to move from the current image-based content in its patent examination system to content that is comprised of editable text. As we understand it, PATI is a system that uses OCR ("Optical Character Recognition") technology to convert images to text. Although PATI may be useful for many things, Coalition members are concerned about its ability to convert image-based content to high quality, editable text . Computer-readable text has many advantages over images, including the fact that it supports text-searching as well as copying, pasting and a range of other important functionality such as text analytics. However, raw OCR generally does not produce highly accurate text. A key question for the CPTID, as well as for other users of published patent information, is whether PATI will produce results that are sufficient for the range of purposes for which the USPTO might attempt to use its "raw" OCR output and whether the USPTO is following the best course to deliver cost-effective and high quality results.

In a recent PPAC meeting, the USPTO has indicated that it intends to convert about 60,000,000 pages of patent content from images to text using PATI in fiscal year 2012. The large number of errors introduced by raw OCR – if they are not caught and corrected – could significantly impact the quality of patent examination.

The problem is greatly compounded, however, if the USPTO does not intend to regularize and clean up the patent text content to today's standard prior to final publication. Fortunately, the agency has indicated that at the present, the "raw" OCR is only going to be used for examination purposes and *not* for publication purposes. Earlier this year, Rep. Chaka Fattah, Ranking Member of the House CJS Appropriations Subcommittee, asked Director Kappos: "As PTO develops its new Patent Application Text Initiative technology, what are PTO's plans for ensuring that this new system will help maintain PTO's current high standard of content accuracy for PTO's databases and published U.S. patent content?" Director Kappos answered: "The Patent Application Text Initiative (PATI) is one project under the Patent End-to- End (PE2E) portfolio of information technology projects. Any PE2E projects that would impact the current patent publication process would meet or exceed the current standards. " Users of US patent information are working to make sure the USPTO – at a minimum - remains committed to maintaining its current quality standard for published patent information. This is good

news. If at any point in the future, however, the agency decides that the raw OCR is "good enough," the quality of U.S. published patent content is certain to suffer.

The users of this information rely on the quality of the U.S. published patent databases to make their own decisions about whether their ideas are really novel, whether to apply for a patent, to determine the state of the art in a particular field, and (in the case of global IPOs) whether the patent applications they are examining in countries outside the U.S. are truly novel. Removing a rigorous check from the process that ensures high quality US patent content would be the wrong decision, particularly at a time when the agency is raising its overall fees. Higher fees should be coupled with as good or better published patent content, not lesser quality. Patent stakeholders are keeping an eye on the USPTO to make sure it maintains its current "gold standard" for the publication of raw patent and trademark data.

Another aspect of the IT upgrade is encoding the text of patent information in XML so that there is unifying standard for all of the text that the office processes for each patent. Under Secretary Kappos has made a priority for XML to be incorporated as part of the larger technological upgrades at the agency and reiterated at a recent House Appropriations Committee hearing that it remains a goal of the agency.

XML markup is today applied by the USPTO to documents primarily at the point of publication. However, the agency has long had a vision of shifting the burden to applicants and requiring them to submit documents in XML at the time of application. By doing so, the USPTO hopes to reduce the agency's costs while at the same time improving the utility of patent application content for patent examination. The goals are worthy ones. The USPTO's past attempts at achieving these outcomes, however, have received less support than anticipated from the applicant community, and the experience of other government agencies with similar (but smaller scale and less complex) initiatives suggests that the path to successful applicant implementation of XML is long, expensive, burdensome for applicants and fraught with significant risks.

The USPTO hopes to achieve cost savings at the agency by receiving applications in XML. This is unlikely to happen for two major reasons. First, the agency is highly unlikely to get, from most applicants, XML documents constructed in a manner that actually provides full information value to the office because of improper or inconsistent use of XML tags. Even in today's far simpler e-filing environment, which simply involves submitting PDF documents, applicants struggle to code documents correctly and their submissions are frequently not compliant with the agency's document indexing system. Consider the likely error rate, then, in an environment that is far more complex such as XML. Applicants will almost certainly submit a large amount of improperly or inconsistently tagged content, and it will be a monumental and very expensive task for the agency to find and correct those applicant-generated errors. In fact, the course suggested by the agency could actually paralyze the automated tools intended to support the examining system by clogging it with error-filled and improperly tagged documents. Contrast this with the highly consistent and accurate data capture process producing the XML data used by the agency and distributed to the world today.

The problem is not that applicants are careless. Rather, the problem stems from three major causes: the task is complex, the vast majority of applicants won't submit nearly enough applications to become expert at filing compliant XML, and those who, collectively, provide relatively well constructed documents will certainly introduce variability due to differing capture processes. In fact, there are over 30,000 registered patent attorneys in the U.S. and over 19,500 different agents submit patent applications annually. 16,000 of them submit fewer than 10 applications per year. It is unrealistic to expect those who file so few applications in a given year to become expert at the complex task of XML filing. Given these factors, any cost savings for the agency are like to be illusory, and in fact any shift to applicant-generated XML is likely to result in a deterioration of patent data published by the agency.

In summary, implementing a text environment for patent examiners is a laudable goal, but it should not be done at the expense of the quality of published U.S. patent content. Users of patent information in the US and around the world have a right to expect that the quality of patent information being published by the USPTO will not deteriorate at a time when the agency is supposed to be implementing improved technology solutions, and particularly at a time when the agency is adjusting its fees.

b. Specific Examples of why it matters to our Coalition that the USPTO not degrade the quality of its data

Commercial providers add value by enhancing raw data to create new features and functionality. An example of added value is LexisNexis's Semantic Search feature. This feature analyzes a phrase or paragraph of text entered by the searcher. Behind the scenes, it makes intelligent connections to other words and phrases that may or may not have occurred to the searcher. After analysis of the phrase or paragraph, Semantic Search offers the searcher the additional terms and phrases it identified as related. The intelligent connections made by Semantic Search are supported by a massive database that was created by a proprietary algorithm. This algorithm examined the text of millions of USPTO patents and Elsevier scientific journals to find connected terms and phrases. It then stored the connected terms in a database. This database is used to provide the searcher with a more intelligent search experience. LexisNexis Semantic Search incorporates "Dynamic Learning". That is, in order to make intelligent connections, our technology has "learned" from the USPTO patent database and the Elsevier journals. Furthermore, it is constructed to continue learning over time. As new patents and journal articles are published, Semantic Search will learn from them. If the quality of the USPTO's data is degraded, this will likely impact the quality of our feature.

Similarly, in the new ProQuest Dialog interface, there are three new auto functions to improve a user's search experience: (1) Suggested terms, which are common search terms and phrases. Users have the ability to select terms from a drop list that appears as they type. (2) Lemmatization is the process of grouping together inflected forms of words. ProQuest Dialog looks for the following forms of your search word: British and American spelling, singular, plurals, past/future tenses and different forms of the verb, e.g., boils, boil, boiling. (3) "Did you mean?" assists with misspelled words. If the user gets no results, the search engine searches for one alternative spelling and will include a statement showing you what was searched. If the original search term has some results, those are presented along with

suggested alternate spellings. Similar to Lexis Nexis' Semantic Search feature, **w**hen one runs a search in ProQuest Dialog, the search engine automatically evaluates the user's search terms to provide "Suggested subjects." Click a suggested subject to retrieve a new list of results. The suggested subjects displayed are based on a combination of factors that relate to term frequency and controlled vocabulary within the indexing of the databases being searched. As shown in the following example, the search engine it will provide suggested subjects that include synonyms of the terms you are searching. For the term **mrsa**, the search engine suggests searching full subject names such as "Staphylococcus aureus AND Methicillin Resistance" and more.

As another example, the Thomson Reuters expert analysts that abstract, index and classify US patents for inclusion in various value-added patent information databases and products rely on high quality, error-free patent raw data to help them in these activities. Ensuring that the raw data is clean, correct, and complete is an important responsibility of the USPTO.

It is obvious that these programmatic algorithms that analyzes data ultimately depends on that data's syntax. Correct spelling, spacing, punctuation, use of special characters, etc., are all elements in the success of the algorithm. Ensuring that the raw data is clean, correct, and complete is an important responsibility of the USPTO.

c. Specific Examples of why it matters to our customers – the end users - that the USPTO not degrade the quality of its data

Users of Coalition members' patent databases - those from Dialog, Thompson Reuters and LexisNexis - tend to be highly skilled in Boolean searching. They know the effect of poor quality data on search results. In other words, they notice immediately when data is missing or corrupted. OCR is widely known to produce data that is, at best, challenging to search and, at worst, impossible to search. For example, NewsinHistory.com noted that OCR commonly misinterprets the characters a, o, e, r, i, and n. Searchers have adapted to this problem by using wildcard searches. Wildcard searching allows for the widest variety of spelling variations. For example, when searching for "majesty", the searcher would search "m?j?sty". While this workaround is useful, it is time consuming and very inconvenient for searchers to formulate this type of search. Most searchers today use every technique possible to save time, such as copying and pasting text into the user interface. Also, this type of searching obviously produces mis-hits that have to be manually examined. Manual examination of a results list is very expensive. Searchers expect, and deserve, technology that frees them from manual labor. The use of OCR creates more manual labor.

We hope that these examples from members of our coalition underscore the importance of the quality of USPTO data to companies and end users. We also hope the agency will take these concerns into consideration when the office considers how to utilize the new fee schedule for its IT upgrade in a way that will maintain the current high quality standards for patent data from the USPTO.