From: Winkler, Michael
Sent: Wednesday, June 20, 2012 1:18 PM
To: seq_listing_xml
Cc: Wolski, Susan
Subject: ABA-IPL Comments Relating to Disclosure of Sequence Listings using XML

Please find attached comments on Disclosure of Sequence Listings using XML, submitted on behalf of ABA-IPL Section Chair Robert A. Armitage.

Please feel free to contact us if there are any questions.

Thank you.

Mike Winkler
Director, Section of Intellectual Property Law
American Bar Association
321 North Clark Street
Chicago, IL 60654
T: (312) 988-5639
F: (312) 988-6800
mike.winkler@americanbar.org

Mark your calendar:
ABA Annual Meeting
ABA-IPL Section events
August 2-5, 2012
www.americanbar.org/iplaw

Connect with ABA-IPL:

AMERICAN BAR ASSOCIATION

June 20, 2012

Via Electronic Mail
seq_listing_xml@uspto.gov
copy to Susan.Wolski@uspto.gov

The Honorable David J. Kappos
Under Secretary of Commerce for Intellectual Property
and Director of the United States Patent and Trademark Office
Mail Stop Comments - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

      Attn: Susan C. Wolski, Office of Patent Cooperation Treaty Legal
            Administration, Office of Associate Commissioner for Patent
            Examination Policy
      Re: Request for Comments on the Recommendation for the Disclosure of
            Sequence Listings Using XML (Proposed ST.26), 77 Fed. Reg.
            28541 (May 15, 2012)

Dear Under Secretary Kappos:

      I am writing on behalf of the American Bar Association Section of
Intellectual Property Law (the "Section") to provide comments in response to the
request of the United States Patent and Trademark Office ("USPTO" or "the
Office") published in the Federal Register on May 15, 2012 (PTO-P-2012-2018). In
particular, the Section submits the following comments on the *Request for
Comments on the Recommendation for the Disclosure of Sequence Listings Using
XML (Proposed ST.26)*, 77 Fed. Reg. 28541 (the "Request for Comments"). These
comments have not been approved by the American Bar Association's House of
Delegates or Board of Governors and should not be considered to be views of the
American Bar Association.

      The Section supports in principle the establishment of a standardized
electronic format for the submission of biological sequence data to the USPTO and
other patent offices. Accordingly, the Section supports the development of a tool
which enables preparation of a sequence listing for submission in the proposed
ST.26 (XML) format. To improve consistency among databases, the Section further
supports in principle submission of biological sequence data in a format that is
more consistent with the format used in public sequence databases in order to

simplify exchange of sequence data in electronic form and to facilitate searching of sequence data. Thus, the Section supports the development of a tool that accepts

biological sequence data prepared for submission in accordance with either the current sequence data format under WIPO ST.25 or the proposed ST.26 format and converts the data to the other format, which conversion tool would enable a comparison of the data submitted under both formats and lessen the burden and expense of preparing biological sequence data in two separate formats.

As an example, the Office's free PatentIn software could be updated to allow sequence listings to be saved in either ST.25 or ST.26 format. Providing such software for free, along with a user manual and technical support, would lessen the burden on practitioners as the practitioner could continue to prepare sequence listings in ST.25, delaying or perhaps even eliminating the need for current practitioners to learn new software. This would also allow the practitioners to see the two sequence listings side-by-side, helping them learn to read the XML format and helping them compare the sequence listings for obvious differences between them—potentially erroneous differences.

In response to the specific requests proffered by the Office, the Section would like to offer the following comments to requests (1), (4), and (6).

Request (1) *Comprehensiveness and Clarity.*

"The Office invites comments on whether the main body of the standard is sufficiently comprehensive and clear to achieve this goal, and in particular welcomes suggestions to add details or clarify the language as appropriate."

The overall comprehensiveness and clarity of proposed ST.26 is greatly appreciated, particularly the detailed examples of different scenarios which the practitioner will encounter. However, several areas exist where proposed ST.26 lacks clarity. Of course, the Section favors clarity in the standard which establishes and defines the format for use in electronically submitting biological sequence data to the USPTO and other patent offices. Accordingly, the Section supports clarifying the language used in sections of the main body of the proposed XML standard WIPO ST.26, including but not limited to—

(a) clarification in paragraph 20 of the sentence, "The symbol 'X' is the equivalent of only one modified amino acid" to explain its meaning in words or through examples, and specifically, to explain whether this means: 1) that no genera are allowed to describe X; 2) that genera are allowed but a genus must be a list of single amino acids only; 3) that genera are allowed but a genus may not include the term "or is absent"; 4) that X cannot represent a modification of an amino acid such as amidation or PEGylation with no amino acid named; or 5) that X cannot represent an amino acid and its modification such as Pro-NH$_2$, but instead must represent only the amino acid itself;

(b) clarification in paragraph 20 of the sentence, "Modified amino acids, including D-amino acids, should be represented in the sequence as the corresponding unmodified amino acids whenever possible" to explain its meaning in words or through examples, and specifically to explain how to list amino acids which have additions or deletions to an unmodified amino acid's core structure, such as norleucine to leucine or homoarginine to arginine;

(c) clarification in paragraphs 20 and 21 between the feature keys for "post-transitionally modified amino acids," "non-post-transitionally modified amino acids," and "other" amino acids not covered by paragraph 20, to explain the differences between these terms and the purpose for describing these amino acids using different feature keys in a sequence listing; and

(d) clarification in paragraph 58 of the term "variant" to explain its meaning in words or through examples, and specifically, to explain whether this term is used to describe a genus of possible amino acids at a particular location in a biological sequence or whether only a single amino acid may be represented at each particular location; and

(e) revising the language used in paragraphs 44 and 47 of the main body of the proposed XML standard WIPO ST.26 to replace the use of the symbols "<" and ">" with the letters "L" for the "less than" symbol and "G" for the "greater than" symbol because the symbols < and > are reserved characters of the XML format and data including these symbols will be encoded using the character sequences < and >, the use of which will reduce human-readability and increase the complexity of encoding sequence data using the proposed ST.26 standard.

To aid the Office in fully understanding the statements in subparts (a)-(e), the Section offers more detailed explanations, examples, and questions for clarification.

(a) <u>Clarification in paragraph 20 of the sentence, "The symbol 'X' is the equivalent of only one modified amino acid."</u> This statement could be understood several ways.

1) Does it mean that X cannot represent a genus of single amino acids? This would differ from the current practice, in which X (or its equivalent 3-letter code Xaa) is used to define a genus of amino acids which could occur at a specific position in a sequence. For example, "Xaa at position 5 is Leu, Ile, Val, or Ala." Under proposed ST.26, does "only one modified amino acid" mean that each possible amino acid in the genus—Leu, Ile, Val, and Ala—must be listed singly in a separate sequence with its own sequence ID number? No example is provided in proposed ST.26 which illustrates a sequence in which X defines a genus of single amino acids which could be substituted at that position.

2) Alternatively, does this statement mean that X cannot refer to a group of amino acids? When one or more amino acids may occur at the end of a sequence, some practitioners refer to X as a group of amino acids such as "X at position 31 may be Lys, Gly, or Gly-Pro." When proposed ST.26 refers to "only one modified amino acid," does that mean that a practitioner can no longer include "Gly-Pro" as a possible choice for X?

3) Another possible interpretation places the emphasis of the statement "only one modified amino acid" on the term "amino acid." Often, practitioners will include modifications which are not amino acids as moieties which may be found at a specific position. For example, a sequence may be listed as "X at position 31 may be Lys, Gly, Pro-NH$_2$, -NH$_2$, PEG, or is absent." Even under ST.25, the description of -NH$_2$ or PEG as X would draw an objection, citing to the fact that X must be an amino acid. It seems reasonable that this is the purpose for specifically using the language "only one modified amino acid" in proposed ST.26—to avoid modifications being described as X. Yet, under ST.25, the term "or is absent" in the list of possible amino acids is acceptable. Would proposed ST.26 also deem this language acceptable? The language is unclear. Further, the modified amino acid moiety "Pro-NH$_2$" is acceptable under ST.25. Yet, it is unclear whether such description would be acceptable under proposed ST.26, and given that the examples shown in proposed ST.26 only show unmodified amino acids, it seems unlikely that the amidated form of a single amino acid would be acceptable in a genus list. The literal language of proposed ST.26 is simply unclear.

This ambiguity could be resolved by revising the language to explain exactly where the emphasis lies. Clarifying language might include: a) no genera are allowed for X; b) genera are allowed but must be a list of single amino acids only; c) terms indicating deletions—viz., "or is absent"—are not allowed; d) modifications like amidation or PEGylation, without listing an amino acid, are not allowed; or e) modifications such as amidation on the listed amino acids may not be shown. Alternatively, more examples could be provided to demonstrate the intended use; consider also showing examples of non-compliant uses.

(b) Clarification in paragraph 20 of the sentence, "Modified amino acids, including D-amino acids, should be represented in the sequence as the corresponding unmodified amino acids whenever possible." This language is easily understandable when referring specifically to the D-isomer, as the amino acid is the same amino acid, just the D-form as opposed to the L-form. Similarly, if a cysteine or lysine is modified through PEGylation, or if a phenylalanine is substituted with various moieties around the phenyl ring, it is clear that the unmodified form (cysteine, lysine, or phenylalanine) is the appropriate amino acid to name. The confusion arises when the modified amino acid is an unusual amino acid, such as norleucine, beta-arginine, or homocysteine. Should those be represented as leucine, arginine, and cysteine, respectively? How does a practitioner know what the "corresponding unmodified amino acid" is? Does this term in proposed

ST.26 refer only to amino acids that have the identical core structure for the amino acid and moieties attached to that core as opposed to having a modification within the core (such as the addition or deletion of a methylene group)? Clarification is needed.

(c) <u>Clarification in paragraphs 20 and 21 between the feature keys for "post-transitionally modified amino acids," "non-post-transitionally modified amino acids," and "other" amino acids not covered by paragraph 20.</u> In the examples, it is clear that these two types of modifications require very different feature keys, yet they both describe unusual amino acids. It can readily be envisioned that errors will occur in determining under which of these two modifications certain amino acids fall. The effect of such an error could be devastating, should such an error be cause for rejecting an application as incomplete. Consider whether such a distinction is necessary, and if so, proposed ST.26 should provide the reasoning for the distinction and definitions to aid practitioners in distinguishing the amino acids in each group.

(d) <u>Clarification in paragraph 58 of the term "variant."</u> The distinction in paragraph 21 for "other" amino acids not covered by paragraph 20 is confusing, especially in light of the example shown for the "Description of an 'other' amino acid." This example demonstrates how a practitioner would describe homoserine in a sequence. However, this seems no different from how one would describe the non-post-translationally modified amino acid ornithine or the D-amino acid D-arginine, both of which are exemplified in paragraph 20. Clearly, the "unknown" description in paragraph 21 is necessary, but the description for "other" amino acids is confusing because it is not readily distinguishable from paragraph 20. Are there other amino acids which fall outside the scope of post-translationally modified amino acids and non-post-translationally modified amino acids? If so, then proposed ST.26 should explain the difference to aid practitioners in distinguishing the amino acids in each group.

(e) <u>Revising the language used in paragraphs 44 and 47 of the main body of the proposed XML standard WIPO ST.26 to replace the use of the symbols "<" and ">" with the letters "L" for the "less than" symbol and "G" for the "greater than" symbol.</u> The Section further supports crafting the ST.26 standard to maximize the human readability of sequences encoded in the ST.26 format and the ease of encoding sequences into the ST.26 format. Accordingly, the Section supports revising the language used in paragraphs 44 and 47 to replace the use of the symbols "<" and ">" with the letters "L" for the "less than" symbol and "G" for the "greater than" symbol. The symbols < and > are reserved characters of the XML format which must be encoded using the character sequences < and >. These character sequences reduce the human-readability of sequences encoded in the ST.26 format and increase the complexity of encoding sequence data using the proposed ST.26 format.

Request (4) *Definition of a Sequence for which a Sequence Listing is Required.*
(a) *Prohibited sequences.*
(b) *Modified nucleotides.*
(c) *D-amino acids.*
(d) *Variants.*
"The Office requests comments on whether these changes as set forth in paragraphs (a) through (d) above are desirable, and what difficulties, if any, are likely to be faced in complying with the definition in the XML standard."

Paragraph 1 of proposed ST.26 states purposes for revisions to the biological sequence data submission standard, such as "facilitat[ing] searching of the sequence data, and allow[ing] data to be exchanged in electronic form and introduced onto computerized databases." The Section supports in principle the disclosure of biological sequence data in a format which facilitates searching of sequence data. Accordingly, the Section opposes the prohibition by Proposed ST.26 of branched nucleotide or amino acid sequences or any sequences with fewer than ten specifically defined nucleotides or fewer than four specifically defined amino acids, and the Section supports permitting but not requiring the inclusion of such sequences in XML format. Analyses of each of the four areas for comment are provided below.

(a) <u>Prohibited sequences.</u> Given the stated purposes of facilitating searches and allowing data exchange, it is difficult to understand why proposed ST.26 would introduce a prohibition of sequence data which under ST.25 is not required but is permitted. Paragraph 4 of proposed ST.26 states: "A sequence listing shall not include any branched nucleotide or amino acid sequences or any sequences with fewer than ten specifically defined nucleotides or fewer than four specifically defined amino acids." According to the Manual for Patent Examining Procedure [hereinafter MPEP], under ST.25:

> The limit of four or more amino acids was established for consistency with limits in place for industry database collections whereas the limit of ten or more nucleotides, while lower than certain industry database limits, was established to encompass those nucleotide sequences to which the smallest probe will bind in a stable manner. The limits for amino acids and nucleotides are also consistent with those established for sequence data exchange with the Japanese Patent Office and the European Patent Office.

MPEP § 2422.01 (8th ed. rev. 8 July 2010). ST.25 did not require that certain sequences be listed; however, it also did not *prohibit* such sequences from being listed. Proposed ST.26 provides no explanation at all for why certain sequences cannot be listed—whether it be for consistency with database requirements or for some other reason. Certainly, if these sequences are to be prohibited from listing, an explanation should be provided in the standard.

Although the majority of sequences fall outside the exclusion zone, there are times when short sequences are described in a patent specification, e.g., linkers,

fragments, probes, primers, expressed sequence tags, complementarity determining regions, and epitopes may be three or fewer amino acids in length or nine or fewer nucleotides. For full disclosure, such sequences should at least be allowed in sequence listings.

Furthermore, proposed ST.26 includes the term "specifically defined" when describing amino acids and nucleotides. Amino acids represented by "X" and nucleotides represented by "n" are not specifically defined. Thus, an amino acid sequence which contains only three named amino acids (e.g., Gly-Ile-Val-) joined by peptide bonds to amino acids that are post-translationally modified, non-post-transitionally modified, unknown, or other would be prohibited from listing. This would be true no matter whether the sequence contained four amino acids or 1000 amino acids, so long as only three amino acids are specifically defined. Such a prohibition will mean that some sequences included in patents will not be included in sequence listings, and thus, will not be searchable.

Branched amino acid or nucleotide sequences are also prohibited in the sequence listing under proposed ST.26. Such sequences may be listed under ST.25 by describing the location of the branching as a modified residue, and listing the moiety or chain which modifies that residue. Even multiple branch chains can be described. However, under proposed ST.26, the addition of even one branched moiety on a sequence prohibits the listing of that sequence.

An additional concern is the effect of listing a sequence which is prohibited. One would presume that under proposed ST.26, listing a prohibited sequence would result in a Notice to Comply which would require the removal of a prohibited sequence for compliance. Many practitioners may choose to include sequences which are not clearly prohibited but are questionable, thinking that it is better to include a questionable sequence and have the Patent Office decide whether the sequence is prohibited. Under ST.25, such action would be acceptable, as "[c]ompliance is not a filing date issue" nor is it "a 35 U.S.C. 119/120 issue." MPEP § 2429. However, given that certain sequences are prohibited under proposed ST.26, compliance may have a different effect. Should the effect of listing a sequence prohibited under proposed ST.26 result in failure to obtain a filing date, this would be major cause for concern.

In view of these concerns, the Office should give careful consideration to the purpose for excluding from listing the sequences described above. If the purpose for prohibition is to provide consistency with other databases, then the prohibition should be removed, allowing but not requiring such sequences to be listed, as under the current standard.

(b) <u>Modified nucleotides.</u> The Section supports the inclusion of sequences containing any nucleotides that can be represented using any of the symbols in Annex B.1., paragraph 1, Table 1. The inclusion of such sequences comports with the Section's arguments presented above in the prohibited sequences section, advocating listing of sequences which may properly be listed in the prescribed format. Allowing such

sequences to be listed in sequence listings provides a more complete database of sequences, enabling more comprehensive searches in those databases.

(c) <u>D-amino acids.</u> The Section also supports the requirement for inclusion of sequences which contain at least one D-amino acid. This, like the permissive inclusion of modified nucleotides, will provide for a more complete database of sequences and enable more comprehensive searches.

The inclusion of modified nucleotides and the requirement to list D-amino acids under proposed ST.26 is an advance in sequence listings. Many sequences which were not required under ST.25 will now be included in the sequence database. However, these changes beg the question: why allow or even require these sequences to be listed, yet prohibit sequences which are only three amino acids in length or are branched? These requirements seem antithetical in purpose. To reconcile the purposes, the Section recommends reconsideration of the prohibition of sequences such that those sequences are permitted but not required.

(d) <u>Variants.</u> Under ST.25, confusion exists around the use of the feature "variant." Some practitioners use this term to define the variable amino acids or nucleotides which could occur in a genus sequence. For example, a sequence having ten amino acids uses the feature variant, and specifically lists an Xaa at position 2, which is defined as "Leu, Ile, or Gly," and an Xaa at position 10, which is defined as "Ser, Gly, Gly-NH$_2$ or is absent." Other practitioners list the same genus using the feature "misc_feature" as opposed to "variant." Both are allowable under ST.25, yet it is unclear which description is correct or preferred.

This ambiguity is not resolved by proposed ST.26. In fact, given the example shown for "VARIANT," more ambiguity is introduced, as it only demonstrates a sequence having one amino acid, Leucine, in the variant position 100. If the sequence represents a genus, as described above, would multiple sequence ID numbers be required under proposed ST.26, each listing a single variable at each specific position? Alternatively, would it be acceptable to have the genus sequence listed, containing X in each location in which a variable occurs, but have multiple VARIANT feature keys, each listing one of the possible amino acids which may occur at a specific position? A third alternative would be to use the VARIANT feature key, but list the multiple choices for amino acid at that position. As it is currently drafted, proposed ST.26 is ambiguous with respect to the correct means for describing a genus sequence, and whether VARIANT is the correct feature key to use for such. Further clarification, including an example, is recommended.

Request (6) *Transition Issues.*
"(a) The Office invites comments regarding how much time is likely to be needed for applicants to transition to the XML standard (with the assumption that sequence listing authoring software will be publicly available)."

The Section supports a transition period to allow patent practitioners to assess changes in format for submitting biological sequence data. Specifically, the Section supports an eighteen-month pilot period, beginning upon the introduction by the USPTO of a tool that accepts biological sequence data prepared for submission in accordance with either the current sequence data format under WIPO ST.25 or the proposed ST.26 format, during which pilot period feedback is provided to the USPTO regarding capability of the tool for sequence listing preparation and errors encountered, and at the end of the pilot period, the USPTO assesses the success of the pilot to determine whether to require submission of biological sequence data in ST.26 format.

Given the extent of the change between ST.25 and proposed ST.26, an extended transition period may be needed to determine the reliability of sequences submitted under proposed ST.26. An eighteen-month pilot period is recommended to allow the Office and practitioners (a) to receive feedback on the capability of program(s) that the Office makes publicly available to generate files in the ST.26 format, and (b) to correct any errors discovered in such program(s).

Further, the Section supports simplification of the submission of biological sequence data to a single cost-effective and reliable format. Thus, the Section supports the introduction to a tool that accepts biological sequence data prepared for submission in accordance with either the current sequence data format under WIPO ST.25 or the proposed ST.26 format and converts the data to the other format for use during the transition pilot period, during which period biological sequence data is submitted to the USPTO in both formats to enable comparison of the ST.26 formatted data for accuracy against the WIPO ST.25 formatted data, and at the end of the pilot period, the USPTO assesses the success of the pilot to determine whether to require submission of biological sequence data in ST.26 format.

Although eighteen months from the production release of the program(s) provided by the Office may be a suitable transition time period, the Office is encouraged to provide for extensions of the transition time period in the event that practitioners encounter difficulty operating the program(s) provided by the Office.

Request (6) *Transition Issues.*
"(b) Given the divergent requirements of the proposed XML standard and ST.25 as described above, the Office invites comments on what difficulties an applicant should anticipate if national or regional offices required compliance with different standards (*i.e.,* ST.25 and XML). Will the existence of separate authoring tools for each of the standards mitigate such difficulties?"

The Section supports clarification of the effect of failure to comply with the standard for submitting biological sequence data, and specifically, the Section supports clarification by the USPTO that the failure to comply with the standard will result in a

notice to comply with the standard and will not by itself prohibit the granting of a filing date for an application with a non-compliant sequence listing.

One major concern with a transition directly from ST.25 to proposed ST.26 derives from the lack of any statement by the Office as to the effect of failure to comply with proposed ST.26. The Office should proactively clarify the effect of non-compliance, specifically stating how the filing date will be affected and describing the notice which will be provided to the applicant. Additionally, allowing an applicant to submit sequences using the current ST.25 format along with an XML version under proposed ST.26 for a pilot period will avoid the lack of clarity under proposed ST.26 by protecting the applicant's filing date, as s/he will be able to rely on the ST.25 formatted data to demonstrate the data intended for submission. Of course, duplicate ST.25/proposed ST.26 submissions will increase the burden on the Office because the examiner will have to look at both formats; however, this burden will also provide data to demonstrate the accuracy of submissions under proposed ST.26 and aid to resolve issues discovered during this pilot period.

Moreover, although proposed ST.26 is an international effort to revise ST.25, it may take time for all countries in which practitioners wish to file an application to accept proposed ST.26 formatted sequence listings, and some countries may never accept proposed ST.26. If some countries were to require an ST.25 formatted sequence listing while others require XML, practitioners would need to prepare two separate sequence listings. This would certainly add significant time and effort to filing sequence listings, which is already a difficult and time-consuming task. If the Office provides a tool which allows conversion between the two forms, the burden of *preparing* two sequence listings would be lifted. Of course, the burden of reviewing both sequence listings for accuracy would still exist. Nonetheless, any reduction in burden on the patent practitioner is welcomed.

In closing, the Section recognizes and appreciates the Office's efforts to solicit public opinions regarding the WIPO standard proposed in the Request for Comments and offers the foregoing comments in an effort to help the Office implement rules that best serve the interests of the users of the patent system and the public.

If you have any questions on our comments or would wish for us to further explain any of our comments, please feel free to contact me. Either I or another member of the leadership of the Section will respond to any inquiry.

Very truly yours,

Robert A. Armitage
Section Chair
American Bar Association
Section of Intellectual Property Law