UNITED STATES PATENT AND TRADEMARK OFFICE

Recommendation for the Disclosure of Sequence Listings using XML (ST.26)

Sue Wolski

Office of PCT Legal Administration

Overview

- Background on revision of ST.25
- Transition from ST.25 to ST.26
- Request for Comments
  - Overview
  - Comments specifically requested
  - Differences between ST.26 and ST.25
  - Documents
    - Main Body
    - Annex B.1 Controlled Vocabulary
    - Document Type Definition (DTD for ST.26)

Revision of WIPO ST.25
Standard for the Presentation of Nucleotide and Amino Acids Sequence Listings in Patent Applications

- ST.25 became effective in 1998, has not been revised since that time
- Committee on WIPO Standards (CWS) established a Task Force in October 2010 to revise ST.25
  - European Patent Office (EPO) designated as Task Force Leader
  - Each Office designated 1 or 2 persons to participate in the Task Force
- New Format - Extensible Markup Language (XML)
  - To enhance the accuracy and quality of the presentation of sequences
- Updates requirements of the standard to reflect advancements in sequence-related technology
- More closely aligns requirements with those of the database providers

Progress to Date

- Five rounds of comments and revisions since March 2011
- Final draft documents agreed by Task Force Members end March 2012
- Expect to adopt the new standard at the CWS in early 2013
- Effective date not yet determined
  - Will depend upon transition provisions
  - Task Force to investigate feasibility of a tool that would allow for easy and complete "conversion" between ST.25/26
- Effective date of U.S. regulation changes will coincide with the ST.26 effective date

Transition from ST.25 to ST.26 (1)

- The 19th Session of the MIA (Meeting of International Authorities) held in February 2012 touched on transition
  - Some advocated use of ST.26 as soon as possible; noting that EPO BISSAP software allows selection of either ST.25/26 output format
  - Others were concerned that a "clean" transition was needed; noting that "conversion" between ST.25/26 may not always be possible/reliable
  - The MIA agreed that
    - The task force would investigate the feasibility of a tool that would allow for easy and complete "conversion" between ST.25/26
    - Based upon conclusions reached, appropriate PCT bodies would commence discussion on the most appropriate transition mechanism
    - See paragraphs 82-87 of PCT/MIA/19/14 PROV. http://www.wipo.int/edocs/mdocs/pct/en/pct_mia_19/pct_mia_19_14_prov.doc

Transition from ST.25 to ST.26 (2)

- The 5th Session of the PCT Working Group held 29 May – 1 June 2012
  - IB *proposed* an effective date 2-3 years after adoption
    - All sequence listings would then be required in ST.26
    - ST.25 would exist only for applications already filed
  - IB *proposed* issuance of a Circular to all ROs, ISAs and DOs requesting an estimate as to a date upon which receipt, processing and searching of ST.26 sequence listings would be possible
  - EPO took over assessment of the feasibility of a "conversion" tool
    - evaluation not yet completed, results expected this summer
  - http://www.wipo.int/edocs/mdocs/pct/en/pct_wg_5/pct_wg_5_14.doc
  - http://www.wipo.int/edocs/mdocs/pct/en/pct_wg_5/pct_wg_5_14_add.doc

Request for Comments on the Recommendation for the Disclosure of Sequence Listings Using XML (Proposed ST.26)

- Published in the Federal Register at 77 Fed. Reg. 28541 on May 15, 2012
- http://www.uspto.gov/patents/law/comments/sequence_listings.jsp
- Three documents have been posted
  - Main body of the standard (22 pages)
    - Primary document to which comments should be directed
  - Annex B.1 Controlled vocabularies (57 pages)
  - Document Type Definition (DTD for Proposed ST.26) (5 pages)
- Written comments must be received on or before July 16, 2012 to ensure consideration
- No public hearing will be held

Comments Specifically Requested (1)

- Comments may be offered on any aspect of the proposed standard or Annexes, transition issues, or expected implementation in the United States
- Comments are specifically requested on the following issues:
  1. Is the main body of the standard sufficiently clear and comprehensive?
  2. Does the standard include any unnecessary procedural requirements, or exclude any procedural requirements that should be retained?
  3. Are there any feature keys or qualifiers that are not clear, or that are optional and should be mandatory (or vice versa)?

Comments Specifically Requested (2)

4. Are the major changes made in ST.26 (see next slide) desirable and what difficulties, if any, are likely to be faced in complying with the XML standard?
5. ST.26 does not provide for the inclusion of prior publications of references in the sequence listing. Is there any perceived detriment to this?
6. How much time is likely to be needed for applicants to transition to the XML standard, and what difficulties should an applicant anticipate if some national or regional offices required compliance with ST.25 while others required compliance with ST.26?

Major Differences Between ST.26 and ST.25

- ST.26
  - XML rather than text format
    - Numeric identifiers replaced by XML tags
  - Contains only technical requirements; PCT Administrative Instructions (AIs) and national/regional laws will contain procedural requirements
  - Requires inclusion of modified nucleic acids and amino acids not previously provided for (e.g. D-amino acids, PNAs, morpholinos)
  - Requires inclusion of variations (i.e. deletions, additions, substitutions)
  - Uses feature keys and qualifiers (limited set of controlled vocabulary) to annotate regions or sites of interest in sequences (replaces and significantly enhances ST.25 Tables 5 and 6)

Example SEQ ID NO 3 from an ST.25 Sequence Listing

<210> 3
<211> 11
<212> PRT
<213> Artificial Sequence

<220>
<223> Designed peptide based on size and polarity to act as a linker between the alpha and beta chains of Protein XYZ.

<400> 3

Met	Val	Asn	Leu	Glu	Pro	Met	His	Thr	Glu	Ile
1				5					10	

Example SEQ ID NO 3 from an ST.26 Sequence Listing

```
<SequenceData sequenceIDNumber="3">
    <INSDSeq>
        <INSDSeq_length>11</INSDSeq_length>
        <INSDSeq_moltype>AA</INSDSeq_moltype>
        <INSDSeq_division>PAT</INSDSeq_division>
        <INSDSeq_feature-table>
            <INSDFeature>
                <INSDFeature_key>SOURCE</INSDFeature_key>
                <INSDFeature_location>1..11</INSDFeature_location>
                <INSDFeature_quals>
                    <INSDQualifier>
                        <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                        <INSDQualifier_value>Synthetic Construct</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                        <INSDQualifier_value>protein</INSDQualifier_value>
                    </INSDQualifier>
                    <INSDQualifier>
                        <INSDQualifier_name>NOTE</INSDQualifier_name>
                        <INSDQualifier_value>designed peptide based on size and
        polarity to act as a linker between the alpha and beta chains of Protein
        XYZ</INSDQualifier_value>
                    </INSDQualifier>
                </INSDFeature_quals>
            </INSDFeature>
        </INSDSeq_feature-table>
        <INSDSeq_sequence>MVNLEPMHTEI</INSDSeq_sequence>
    </INSDSeq>
</Sequence Data>
```

Main Body – Scope (paragraphs 1-8)

- High level overview
- Defines for purposes of the standard what is meant by the expressions “nucleotide” and “amino acid”
- Specifies sequences for which a sequence listing is required AND specifies what a SL shall not include
  - A SL 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

Main Body – Presentation of Sequences (paragraphs 9-22)

- All nucleotides must be represented by a one-letter lower case symbol (as defined in Annex B.1)
  - Representation of modified nucleotides explained and examples provided
- All amino acids must be represented by a one-letter UPPER case symbol (as defined in Annex B.1)
  - Representation of modified amino acids explained and examples provided
- Representation of sequences with gaps explained

Main Body – Structure of the SL in XML (paragraphs 23-59)

- Root Element
  - Basic information about the SL file
    - the DTD version used to create the SL, the SL file name, the name and version of the software used to generate the file, the date the SL file was produced
- General Information Part
  - Bibliographic information that relates to the patent application
    - Applicant name, name of first mentioned inventor, file reference number, application number if known, filing date, etc.
  - Used solely for association of the sequence listing to the patent application for which the SL is submitted

Main Body – Structure of the SL in XML (paragraphs 23-59)

- Sequence Data Part – Includes sequence data elements each of which contain information relating to one sequence
  - Mandatory elements: sequence length, molecule type, indication that sequence is related to a patent application, feature table, and sequence
  - Description of how to use feature keys and qualifiers
    - Feature name, location, source of sequenced molecule, free text, coding sequences, and variants
    - Feature keys and qualifiers are set forth in Annex B.1

Annex B.1 Controlled Vocabulary

- Tables 1-4 (pages 2-6) - *replaces ST.25 Tables 1-4*
  - Nucleic acids, modified nucleic acids
  - Amino acids, modified and unusual amino acids
- Feature keys for nucleic acid sequences (pages 7-23)
  - *replaces ST.25 Table 5*
- Qualifiers for nucleic acid sequences (pages 24-41) – *new to ST.26*
- Feature keys for amino acid sequences (pages 42-48)
  - *replaces ST.25 Table 6*
- Qualifiers for amino acid sequences (page 49) – *new to ST.26*
- Genetic Code Tables – (pages 50-52) - *new to ST.26*
- "Country" qualifier values – (pages 53-57) - *new to ST.26*

Feature Keys and Qualifiers for Nucleic Acid Sequences

- Agreed upon by the International Nucleotide Sequence Database Collaboration (INSDC)
  - Used for submissions to the National Center for Biotechnology Information (NCBI) for entries into the GenBank database
  - Familiar to inventors who have supplied sequences to NCBI
- 60 feature keys and 81 qualifiers for nucleic acid sequences
  - INSDC feature keys/qualifiers not relevant for patent data not included
  - The "source" feature key is mandatory for all nucleic acid sequences
    - Remaining feature keys are optional
  - 8 optional feature keys require the presence of another "Parent Key"
    - E.g., C_region feature key requires the CDS feature key
  - 7 feature keys require the presence of one or more qualifiers
    - Remaining qualifiers are optional

Feature keys and Qualifiers for Amino Acid Sequences

- Agreed upon by the UniProt Consortium
- 44 feature keys and 3 qualifiers for amino acid sequences
  - The "SOURCE" feature key is mandatory for all amino acid sequences
    - Remaining feature keys are optional
  - 11 feature keys require the presence of one or more qualifiers
    - Remaining qualifiers are optional

Qualifiers and Free Text

- There are 3 types of value formats for qualifiers
  - Free text
  - Controlled vocabulary or enumerated values (e.g. number or date)
  - Sequences
- Free text is limited to 255 characters
  - Any further information may be included in the application body
- ST.25 requires that any free text be repeated in the main part of the application description
  - To avoid translation of the entire sequence listing
  - Procedural requirement not contained in ST.26, but will be contained in PCT AIs and national/regional law
  - Should be cognizant of this requirement for lengthy sequence listings

Document Type Definition (DTD)

- Document (5 pages) contains
  - General information part
    - Elements intentionally depart from the two WIPO XML standards, ST.36 and ST.96
  - Sequence data part
    - Subset of the INSDC DTD

UNITED STATES PATENT AND TRADEMARK OFFICE

Thank You!

Sue Wolski

Office of PCT Legal Administration

571-272-3304

Susan.Wolski@USPTO.GOV