

**UNITED STATES
PATENT AND TRADEMARK OFFICE**



The background of the slide is a complex technical drawing, likely a cross-section of a mechanical assembly. It features numerous hatched areas representing different materials and various geometric shapes like cylinders, spheres, and rectangular blocks. Numerous numerical callouts are scattered throughout the drawing, such as 80, 84, 86, 82, 24, 30, 86, 34, 32, 52, and 50. The drawing is rendered in a light gray color against a dark purple background.

ST.26 Sequence Annotation: Qualifier “Free Text”

Susan Wolski

July 2019

WIPO standard ST.26

- Title

- Recommended Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings using XML (Extensible Markup Language)

- Definition of a sequence listing

- “A part of the description of the patent application...which includes the disclosed nucleotide and/or amino acid sequence(s), along with any further description, as prescribed by this Standard.”

Description of an invention

- Patent applications describe inventions in the
 - Description, claims, and drawings
- Computer inventions may also contain
 - A computer program listing
 - Written in programming language for understanding by the relevant scientific community
- Biotechnology inventions having sequence disclosures also contain
 - A sequence listing
 - Written in a language for understanding by the relevant scientific community

WIPO Standard ST.26 Goals

- The four goals of ST.26 are the same as those of ST.25, to:
 - Allow applicants to draw up a single sequence listing acceptable for the purposes of both international and national or regional procedures,
 - Enhance the accuracy and quality of presentations of sequences for easier dissemination, benefiting applicants, the public, and examiners,
 - Facilitate searching of the sequence data, and
 - Allow sequence data to be exchanged in electronic form and introduced into computerized databases.

Ongoing effort to meet ST.26 goals

- To advance the four ST.26 goals, the European Patent Office (EPO) proposed:
 - Use of the International Nucleotide Sequence Data Collaboration (INSDC) system of sequence annotation and XML format for ST.26
 - INSDC is a collaboration of database providers
 - DDBJ: DNA Data Bank of Japan
 - EMBL-EBI: European Molecular Biology Laboratory, European Bioinformatics Institute
 - NCBI: National Center for Biotechnology Information

INSDC databases

- Accept sequence submissions from
 - The scientific community worldwide, and
 - Patent applications filed with:
 - JPO: Japan Patent Office
 - KIPO: Korean Intellectual Property Office
 - EPO: European Patent Office
 - USPTO: United States Patent and Trademark Office
- Provide free search of sequence data to the public

INSDC publication of sequence data

- ST.25
 - Publication of sequences only; no annotations
 - Since annotations are not published
 - Character set used for annotations is inconsequential
- ST.26 – on/after effective date
 - Expected publication of sequences, as well as annotations
 - For publication annotations must be INSDC-compliant
 - Character set used for annotations is Unicode Basic Latin
 - Advances the second, third, and fourth ST.26 goals
 - Enhances the accuracy and quality of sequence data
 - Facilitates searching of sequence data
 - Allows exchange of sequence data in electronic form and introduction into computerized databases

ST.26 and INSDC annotations

- Feature keys and qualifiers
- Qualifier values are one of:
 - Controlled vocabulary or enumerated values
 - Sequences
 - “Free text”

“Free text”

- 51 qualifiers have a “free text” value format
- INSDC requires Unicode Basic Latin
 - ST.26 likewise limits “free text” to Unicode Basic Latin
- **INSDC submissions should also be in English**
 - To allow content to be understandable to the broadest audience possible
 - Scientific community overwhelmingly uses English for annotation and description of sequences
 - ST.26 states that free text “should preferably be in the English language”

ST.26 applicability

- International applications (IAs) under the PCT – both international and national stage
- Likely many IAs will enter national/regional phase in JPO, KIPO, EPO, or USPTO
 - Sequence data is provided to DDBJ, EBI, or NCBI
- Applications filed in national and regional offices outside the PCT
- Data in INSDC-compliant format advances the first ST.26 goal
 - For a single sequence listing to be acceptable for both international and national or regional procedures

Language aspects of “free text”

- Discussed by the Sequence Listings (SEQL) Task Force since 2016
- All “free text” qualifier values in ST.26 annex I
 - Include: “NOTE: This value may require translation for National/Regional procedures”
- Recent effort to identify all “free text” as either
 - Language dependent, or
 - Language independent

“Language dependent”

- Certain qualifiers values are “language dependent” and cannot be further limited.

6.21. Qualifier	function
Definition	function attributed to a sequence
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>essential for recognition of cofactor </INSDQualifier_value>
Comment	The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

6.40. Qualifier	note
Definition	any comment or additional information
Value format	free text (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>A comment about the feature</INSDQualifier_value>

“Note” qualifier in ST.26 XML sequence listing

```
<INSDFeature>  
  <INSDFeature_key>REGION</INSDFeature_key>  
  <INSDFeature_location>&gt;5</INSDFeature_location>  
  <INSDFeature_qualifiers>  
    <INSDQualifier>  
      <INSDQualifier_name>NOTE</INSDQualifier_name>  
      <INSDQualifier_value>The entire sequence of amino acids 1-5 can be repeated one or more  
times</INSDQualifier_value>  
    </INSDQualifier>  
  </INSDFeature_qualifiers>  
</INSDFeature>
```

Limitation to “language independent”

- Some qualifier values can be effectively limited to “language independent.”

6.20. Qualifier	frequency
Definition	frequency of the occurrence of a feature
Value format	free text <u>limited to representing numeric representation of</u> the proportion of a population carrying the feature expressed as a fraction (NOTE: this value may require translation for National/Regional procedures)
Example	<INSDQualifier_value>23/108</INSDQualifier_value> <INSDQualifier_value>1 in 12</INSDQualifier_value> <INSDQualifier_value>0.85</INSDQualifier_value>

PCT/WG/12/13 proposal

- To modify ST.26 *or* the PCT
 - To clearly identify those “free text” values always considered to be “language dependent”
 - Repetition in application description required by PCT
 - “Free text” values not identified as “language dependent”
 - Repetition in application description **not** required by PCT

Proposed ST.26 Annex I revision to identify “language dependent” qualifiers

- Identification in ST.26, rather than the PCT would ensure harmonization within and outside the PCT.
- Annex I already warns that “free text” values may require translation.

Sequence Listing Task Force

- To propose draft revision of ST.26 Annex I
 - By the end of 2019 Q3
 - To include limitations on certain qualifier values to achieve “language independence”
 - To identify “language dependent” qualifiers
- To propose CWS adoption
 - In 2020



Thank you.

Susan Wolski

susan.wolski@uspto.gov

www.uspto.gov