# ST.26 - ANNEX II

DOCUMENT TYPE DEFINITION FOR SEQUENCE LISTING (DTD)

Final Draft

*Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/5*

<?xml version="1.0" encoding="UTF-8"?>

<!--Annex II of WIPO Standard ST.26, Document Type Definition (DTD) for Sequence Listing

This entity may be identified by the PUBLIC identifier:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

PUBLIC "-//WIPO//DTD SEQUENCE LISTING 1.~~0~~1//EN" "ST26SequenceListing\_V1\_~~0~~1.dtd"

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* PUBLIC DTD URL

\* http://www.wipo.int/standards/DTD/ST26SequenceListing\_V1\_~~0~~1.dtd

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

~~WIPO Standard ST.26, version 1.0, Recommended Standard for the presentation of nucleotide and amino acid sequence listings using XML (eXtensible Markup Language), adopted by the Committee on WIPO Standards (CWS) at its reconvened fourth session on March 24, 2016~~

Revision of Annex II to WIPO Standard ST.26 is submitted for approval by the Committee on WIPO Standards (CWS) at its fifth session.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* CONTACTS

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

xml.standards@wipo.int

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* NOTES

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

The sequence data part is a subset of the complete INSDC DTD V.1.5 that only covers

the requirements of WIPO Standard ST.26.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* REVISION HISTORY

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2017-06-02: Version 1.1 (if it is approved by the CWS)

Changes:

Comments added to <INSDSeq\_length>, <INSDSeq\_division> and <INSDSeq\_sequence> to clarify the reason of the differences between the INSDC DTD v.1.5 and ST26 Sequence Listing DTD V1\_1.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2016-03-24: Version 1.0 adopted by the CWS/4Bis

2014-03-11: Final draft for adoption.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ST26SequenceListing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* ROOT ELEMENT

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

-->

<!ELEMENT ST26SequenceListing ((ApplicantFileReference | (

 ApplicationIdentification,ApplicantFileReference?)),

 EarliestPriorityApplicationIdentification?,(ApplicantName,

 ApplicantNameLatin?)?,(InventorName,InventorNameLatin?)?,

 InventionTitle+,SequenceTotalQuantity,SequenceData+) >

<!--The elements ApplicantName and InventorName are optional in this DTD to facilitate

the conversion between various encoding schemes-->

<!ATTLIST ST26SequenceListing

 dtdVersion CDATA #REQUIRED

 fileName CDATA #IMPLIED

 softwareName CDATA #IMPLIED

 softwareVersion CDATA #IMPLIED

 productionDate CDATA #IMPLIED >

<!--ApplicantFileReference

Applicant's or agent's file reference, mandatory if application identification not provided.

-->

<!ELEMENT ApplicantFileReference (#PCDATA) >

<!--ApplicationIdentification

Application identification for which the sequence listing is submitted, when available.

-->

<!ELEMENT ApplicationIdentification (IPOfficeCode,ApplicationNumberText,

 FilingDate?) >

<!--EarliestPriorityApplicationIdentification

Application identification of the earliest claimed priority, which contains IPOfficeCode, ApplicationNumberText and FilingDate elements.

For details, please see ApplicationIdentification.

-->

<!ELEMENT EarliestPriorityApplicationIdentification (IPOfficeCode,

 ApplicationNumberText,FilingDate?) >

<!--ApplicantName

The name of the first mentioned applicant in characters set forth in paragraph 40 (a) of the ST.26 main body document.

-->

<!--languageCode: Appropriate language code from ISO 639-1 – Codes for the representation of names of languages - Part 1: Alpha-2

-->

<!ELEMENT ApplicantName (#PCDATA) >

<!ATTLIST ApplicantName

 languageCode CDATA #REQUIRED >

<!--ApplicantNameLatin

Where ApplicantName is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the name of the first mentioned applicant must also be typed in characters as set forth in paragraph 40 (b) of the ST.26 main body document.

-->

<!ELEMENT ApplicantNameLatin (#PCDATA) >

<!--InventorName

Name of the first mentioned inventor typed in the characters as set forth in paragraph 40 (a).-->

<!--languageCode: Appropriate language code from ISO 639-1 – Codes for the representation of names of languages - Part 1: Alpha-2

-->

<!ELEMENT InventorName (#PCDATA) >

<!ATTLIST InventorName

 languageCode CDATA #REQUIRED >

<!--InventorNameLatin

Where InventorName is typed in characters other than those as set forth in paragraph 40 (b), a translation or transliteration of the first mentioned inventor may also be typed in characters as set forth in paragraph 40 (b).

-->

<!ELEMENT InventorNameLatin (#PCDATA) >

<!--InventionTitle

Title of the invention typed in the characters as set forth in paragraph 40 (a) in the language of filing. A translation of the title of the invention into additional languages may be typed in the characters as set forth in paragraph 40 (a) using additional InventionTitle elements. Preferably two to seven words.

-->

<!--languageCode: Appropriate language code from ISO 639-1 - Codes

for the representation of names of languages - Part 1: Alpha-2

-->

<!ELEMENT InventionTitle (#PCDATA) >

<!ATTLIST InventionTitle

 languageCode CDATA #REQUIRED >

<!--SequenceTotalQuantity

Indicates the total number of sequences in the document.

Its purpose is to be quickly accessible for automatic processing.

-->

<!ELEMENT SequenceTotalQuantity (#PCDATA) >

<!--SequenceData

Data for individual Sequence.

For intentionally skipped sequences see the ST.26 main body document.

-->

<!ELEMENT SequenceData (INSDSeq) >

<!ATTLIST SequenceData

 sequenceIDNumber CDATA #REQUIRED >

<!--IPOfficeCode

ST.3 code. For example, if the application identification is PCT/IB2013/099999, then IPOfficeCode value will be International Bureau of WIPO.

-->

<!ELEMENT IPOfficeCode (#PCDATA) >

<!--ApplicationNumberText

The application identification as provided by the office of filing (e.g. PCT/IB2013/099999)

-->

<!ELEMENT ApplicationNumberText (#PCDATA) >

<!--FilingDate

The date of filing of the patent application for which the sequence listing is submitted in ST.2 format "CCYY-MM-DD", using a 4-digit calendar year, a 2-digit calendar month and a 2-digit day within the calendar month, e.g., 2015-01-31. For details, please see paragraphs 7 (a) and 11 of WIPO Standard ST.2.

-->

<!ELEMENT FilingDate (#PCDATA) >

<!--\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* INSD Part

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

The purpose of the INSD part of this DTD is to define a customized DTD for sequence listings to support the work of IP offices while facilitating the data exchange with the public repositories.

The INSD part is subset of the INSD DTD v1.~~4~~5 and as such can only be used to generate an XML instance as it will not support the complete INSD structure.

This part is based on:

The International Nucleotide Sequence Database (INSD) collaboration.

INSDSeq provides the elements of a sequence as presented in the GenBank/EMBL/DDBJ-style flatfile formats. Not all elements are used here.

-->

<!--INSDSeq

Sequence data. Changed INSD V1.5 DTD elements, INSDSeq\_division and INSDSeq\_sequence from optional to mandatory per business requirements.

-->

<!ELEMENT INSDSeq (INSDSeq\_length,INSDSeq\_moltype,INSDSeq\_division,

 INSDSeq\_other-seqids?,INSDSeq\_feature-table?,INSDSeq\_sequence) >

<!--INSDSeq\_length

The length of the sequence. INSDSeq\_length allows only integer.

-->

<!ELEMENT INSDSeq\_length (#PCDATA) >

<!--INSDSeq\_moltype

Admissible values: DNA, RNA, AA

-->

<!ELEMENT INSDSeq\_moltype (#PCDATA) >

<!--INSDSeq\_division

Indication that a sequence is related to a patent application. Must be populated with the value PAT.

-->

<!ELEMENT INSDSeq\_division (#PCDATA) >

<!--INSDSeq\_other-seqids

In the context of data exchange with database providers, the Patent Offices should populate for each sequence the element INSDSeq\_other-seqids with one INSDSeqid containing a reference to the corresponding published patent and the sequence identification.

-->

<!ELEMENT INSDSeq\_other-seqids (INSDSeqid?) >

<!--INSDSeq\_feature-table

Information on the location and roles of various regions within a particular sequence. Whenever the element INSDSeq\_feature-table is used, it must contain at least one feature.

-->

<!ELEMENT INSDSeq\_feature-table (INSDFeature+) >

<!--INSDSeq\_sequence

The residues of the sequence. The sequence must not contain numbers, punctuation or whitespace characters.

-->

<!ELEMENT INSDSeq\_sequence (#PCDATA) >

<!--INSDSeqid

Intended for the use of Patent Offices in data exchange only.

Format:

pat|{office code}|{publication number}|{document kind code}|{Sequence identification number}

where office code is the code of the IP office publishing the patent document, publication number is the publication number of the application or patent, document kind code is the letter codes to distinguish patent documents as defined in ST.16 and Sequence identification number is the number of the sequence in that application or patent

Example:

pat|WO|2013999999|A1|123456

This represents the 123456th sequence from WO patent publication No. 2013999999 (A1)

-->

<!ELEMENT INSDSeqid (#PCDATA) >

<!--INSDFeature

Description of one feature.

-->

<!ELEMENT INSDFeature (INSDFeature\_key,INSDFeature\_location,INSDFeature\_quals?) >

<!--INSDFeature\_key

A word or abbreviation indicating a feature.

-->

<!ELEMENT INSDFeature\_key (#PCDATA) >

<!--INSDFeature\_location

Region of the presented sequence which corresponds to the feature.

-->

<!ELEMENT INSDFeature\_location (#PCDATA) >

<!--INSDFeature\_quals

List of qualifiers containing auxiliary information about a feature.

-->

<!ELEMENT INSDFeature\_quals (INSDQualifier\*) >

<!--INSDQualifier

Additional information about a feature.

For coding sequences and variants see the ST.26 main body document.

-->

<!ELEMENT INSDQualifier (INSDQualifier\_name,INSDQualifier\_value?) >

<!--INSDQualifier\_name

Name of the qualifier.

-->

<!ELEMENT INSDQualifier\_name (#PCDATA) >

<!--INSDQualifier\_value

Value of the qualifier.

-->

<!ELEMENT INSDQualifier\_value (#PCDATA) >

[Annex VI to ST.26 follows]