Final Draft - 28.03.2012

Recommendation for the disclosure of sequence listings using XML (Proposed ST.26)

SCOPE

1. The purpose of this Standard for the presentation of nucleotide and amino acid sequence listings in patent applications is to: allow applicant to draw up a single sequence listing in a patent application 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 onto computerized databases.

2. A sequence listing complying with this standard (hereinafter sequence listing) contains a general information part and a sequence data part. The sequence listing must be presented as a single file in Extensible Markup Language (XML) using the Document Type Definition (DTD) presented in Annex B.3. The DTD for the general information part intentionally departs from WIPO Standards ST.36 and ST.96 to ensure management independently of those standards. The purpose of the bibliographic information contained in the general information part is solely for association of the sequence listing to the patent application for which the sequence listing is submitted. The sequence data part is composed of one or more sequence data elements each of which contain information about one sequence. The sequence data elements include various feature keys and subsequent qualifiers agreed upon by the International Nucleotide Sequence Database Collaboration (INSDC) and the UniProt Consortium.

3. For the purpose of this Standard, the expression “sequence listing” means a part of the description of the patent application as filed or a document filed subsequently to the application, which gives a detailed disclosure of nucleotide and/or amino acid sequences and other available information.

4. For the purpose of this Standard, a sequence for which a sequence listing is required is one that is disclosed by enumeration of its residues and is either i) an unbranched sequence containing ten or more specifically defined nucleotides, wherein adjacent nucleotides are joined 3’ to 5’ (or 5’ to 3’), or ii) an unbranched sequence containing four or more specifically defined amino acids, wherein adjacent amino acids are joined by peptide bonds. 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. “Specifically defined” means any nucleotide other than those represented by “n” and any amino acid other than those represented by “X”.

5. For the purpose of this Standard, the expression “nucleotide” means any nucleotide that can be represented using any of the symbols set forth in Annex B.1, paragraph 1, Table 1. Nucleotides may contain, *inter alia*:

- a modified or synthetic purine or pyrimidine base;
- a modified or synthetic ribose or deoxyribose or
- a modified or synthetic 3’ to 5’ internucleotide linkage, i.e., any chemical moiety that provides the same structural function as the phosphate moiety of DNA or RNA.

6. For the purpose of this Standard, the expression “amino acid” means any amino acid that can be represented using any of the symbols set forth in Annex B.1, paragraph 3, Table 3. Such amino acids include, inter alia, D-amino acids and amino acids containing modified or synthetic side chains.

7. For the purpose of this Standard, “sequence identification number” means a unique number (integer) assigned to each sequence in the sequence listing. The sequence identification numbers

shall begin with 1, and increase consecutively by integers.

8. For the purpose of this Standard, the expression "controlled vocabulary" is the terminology contained in this standard that must be used when describing the features of a sequence, i.e., annotations of regions or sites of interest. For example, the feature keys and qualifiers set forth in Annex B.1 are controlled vocabulary.

PRESENTATION OF SEQUENCES

9. Each sequence shall be assigned a separate sequence identification number. The sequence identification numbers shall begin with 1, and increase consecutively by integers. Where no sequence is present for a sequence identification number, i.e. an intentionally skipped sequence, "000" must be used in place of a sequence (*see* paragraph 37). The total number of sequences must be indicated in the sequence listing and must equal the total number of sequence identification numbers, whether followed by a sequence or by "000."

Nucleotide sequences

10. A nucleotide sequence shall be presented only by a single strand, in the 5'-end to 3'-end direction from left to right. The designations 5' and 3' shall not be present in the sequence. A double-stranded nucleotide sequence disclosed by enumeration of the residues of both strands shall be presented i) as a single sequence or as two separate sequences, each assigned its own sequence identification number, where the two separate strands are fully complementary to each other, or ii) as two separate sequences, each assigned its own sequence identification number, where the two strands are not fully complementary to each other.

11. Numbering of the nucleotide positions shall start at the first base of the sequence with number 1. It shall be continuous through the whole sequence in the direction 5' to 3'.

12. The above numbering method for nucleotide sequences is also applicable to nucleotide sequences that are circular in configuration. In this case, the applicant must choose the nucleotide with which numbering begins.

13. All nucleotides in a sequence shall be represented using the symbols set forth in Annex B.1, paragraph 1, Table 1. Only lower case letters shall be used. Any symbol used to represent a nucleotide is the equivalent of only one nucleotide. Where an ambiguity symbol (representing two or more bases in the alternative) is appropriate, the most restrictive symbol should be used. For example, if a base in a given position could be "a or g," then "r" should be used, rather than "n." The symbol "n" will be construed as "a or c or g or t/u" when it is used with no further description as required by paragraph 14 or 15.

14. Modified nucleotides should be represented in the sequence as the corresponding unmodified bases, i.e., a, c, g or t whenever possible. Any modified nucleotide in a sequence that cannot otherwise be represented by any other symbol in Annex B.1, paragraph 1, Table 1, such as non-naturally occurring nucleotides, must be represented by "n." Where the symbol "n" is used to represent a modified nucleotide it is the equivalent of only one modified nucleotide.

A modified nucleotide or special features of a modified nucleotide must be further described in the feature table (*see* paragraph 38 *et seq.*) using the feature key "modified_base" and the mandatory qualifier "mod_base" in conjunction with an abbreviation contained in Annex B.1, paragraph 2, Table 2 as the qualifier value; if the abbreviation is "OTHER," the complete unabbreviated name of the modified base must be provided in a "note" qualifier. The abbreviations (or full names) provided in Table 2 must not be used in the sequence itself. A region containing a known number of contiguous "n" residues for which the same description applies may be jointly described using the syntax "x..y" as the location descriptor in the element `INSDFeature_location` (*see* paragraphs 42-47).

Examples:

Modified nucleotide using an abbreviation from Annex B.1, paragraph 2, Table 2

```
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>15</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>i</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Modified nucleotide using “OTHER” from Annex B.1, paragraph 2, Table 2

```
<INSDFeature>
    <INSDFeature_key>modified_base</INSDFeature_key>
    <INSDFeature_location>4</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>mod_base</INSDQualifier_name>
            <INSDQualifier_value>OTHER</INSDQualifier_value>
        </INSDQualifier>
        <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>xanthine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

15. Any “unknown” or “other” nucleotide must be represented by “n” in the sequence and must be further described in the feature table (see paragraph 38 *et seq*). The symbol “n” is the equivalent of only one “unknown” or “other” nucleotide. An “unknown” nucleotide designated as “n” must be further described using the feature key “unsure;” otherwise, that “unknown” nucleotide will be construed as “a or c or g or t/u” according to paragraph 13. An “other” nucleotide designated by “n,” i.e. not listed in the tables set forth in Annex B.1, paragraphs 1 and 2, must be further described using the feature key “misc_feature” together with the qualifier “note” identifying the “other” nucleotide by its complete, unabbreviated name. See Annex B.1 for descriptions of feature keys and qualifiers. A region containing a known number of contiguous “n” residues for which the same description applies may be jointly described using the syntax “x..y” as the location descriptor in the element `INSDFeature_location` (*see* paragraphs 42-47).

Example:
A region of “other” nucleotides for which the same description applies

```
<INSDFeature>
    <INSDFeature_key>misc_feature</INSDFeature_key>
    <INSDFeature_location>358..485</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>guanosine
tetraphosphate</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Amino acid sequences

16. The amino acids in a protein or peptide sequence shall be listed in the amino to carboxy direction from left to right. The amino and carboxy groups shall not be represented in the sequence.

17. Numbering of amino acid positions shall start at the first amino acid of the sequence, with the number 1, including amino acids preceding the mature protein, for example, pre-sequences, pro-sequences, pre-pro-sequences and signal sequences.

18. All amino acids in a sequence shall be represented using the symbols set forth in Annex B.1, paragraph 3, Table 3. Only upper case letters shall be used. Any symbol used to represent an amino acid is the equivalent of only one amino acid.

19. A disclosure of amino acids separated by one or more blank spaces or internal terminator symbols (for example,"Ter" or "*" or ".") represents multiple separate amino acid sequences. Where such separate amino acid sequences contain at least four specifically defined amino acids and are encompassed by paragraph 4, each such separate sequence shall be presented in the sequence listing as one or more separate amino acid sequences, each with its own sequence identification number, using only the symbols set forth in Table 3 of Annex B.1. Sequences separated by spaces or terminator symbols shall not be presented as a single amino acid sequence in the sequence listing. Terminator symbols and spaces must not be used in sequences in a sequence listing.

20. Modified amino acids, including D-amino acids, should be represented in the sequence as the corresponding unmodified amino acids whenever possible. Any modified amino acid in a sequence that cannot otherwise be represented by any other symbol in Annex B.1, paragraph 3, Table 3, must be represented by "X." The symbol "X" is the equivalent of only one modified amino acid. A modified amino acid or special features of a modified amino acid must be further described in a feature table (see paragraph 38 et seq.) using the feature key "MOD_RES" for post-translationally modified amino acids and the feature key "SITE" for other modified amino acids, and the qualifier "NOTE." The value for the qualifier "NOTE" must either be an abbreviation set forth in Annex B.1, paragraph 4, Table 4, or the complete, unabbreviated name of the modified amino acid. The abbreviations set forth in Table 4 (or full names) must not be used in the sequence itself. A region containing a known number of contiguous "X" residues for which the same description applies may be jointly described using the syntax "x..y" as the location descriptor in the element INSDFeature_location (see paragraphs 42-47).

Examples:
Post-translationally modified amino acid

```
<INSDFeature>
    <INSDFeature_key>MOD_RES</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>aIle</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Non post-translationally modified amino acid

```
<INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Orn</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

D-amino acid

```
<INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>9</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>D-Arginine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

21. "Unknown” or “other” amino acids not covered by paragraph 20, must be represented by “X” in the sequence and must be further described in the feature table (*see* paragraph 38 *et seq.*). The symbol “X” is the equivalent of only one “unknown” or “other” amino acid. An “unknown” amino acid designated as “X” must be further described using the feature key “UNSURE” and optionally the qualifier “NOTE.” An “other” amino acid designated as “X” must be further described using the feature key “SITE” and the qualifier “NOTE” with the complete, unabbreviated name of the “other” amino acid. A region containing a known number of contiguous “X” residues for which the same description applies may be jointly described using the syntax “x..y” as the location descriptor in the element `INSDFeature_location` (*see* paragraphs 42-47).

Examples:

Description of an “unknown” amino acid

```
<INSDFeature>
    <INSDFeature_key>UNSURE</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Alanine or Valine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Description of an “other” amino acid

```
<INSDFeature>
    <INSDFeature_key>SITE</INSDFeature_key>
    <INSDFeature_location>3</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Homoserine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Presentation of Special Situations

22. A sequence disclosed by enumeration of its residues that is constructed from one or more non-contiguous segments of a larger sequence or of segments from different sequences must be included in the sequence listing as a single sequence with a single sequence identification number.

A sequence disclosed by enumeration of its residues that contains regions of specifically enumerated residues separated by one or more regions of contiguous “n” or “X” residues, wherein the exact number of residues in each region is disclosed, must be included in the sequence listing as a single sequence with a single sequence identification number.

A sequence disclosed by enumeration of its residues that contains regions of specifically enumerated residues separated by one or more gaps of an unknown or undisclosed number of residues must be included in the sequence listing as a plurality of separate sequences. Each such separate sequence shall contain one region of specifically enumerated residues with its own sequence identification number, wherein the number of separate sequences is equal to the number of regions of specifically enumerated residues. Sequences containing gaps of an unknown or undisclosed number of residues must not be included in the sequence listing as a single sequence.

STRUCTURE OF THE SEQUENCE LISTING IN XML

23. An XML instance of a sequence listing file according to this standard is composed of:

(a) General information part
The general information part contains information concerning the patent application to which the sequence listing is directed;

(b) Sequence data part
The sequence data part contains one or more sequence data elements, each of which, in turn contain information about one sequence.

24. The sequence listing must be presented in XML using the DTD presented in Annex B.3.

25. The entire electronic sequence listing shall be contained within one contiguous file, encoded using Unicode UTF-8.

26. In an XML instance of a sequence listing, the following reserved characters must be replaced by the corresponding predefined entities (*see e.g.*, paragraphs 44 and 47) when used in a value or content of an attribute or element:

Reserved Character	**Predefined Entities**
<	&lt;
>	&gt;
&	&amp;
"	&quot;
'	&apos;

Root element

27. The root element of an XML instance according to this standard is the element `ST26SequenceListing`. The following attributes should be provided:

Attribute	**Description**
`dtdVersion`	Version of the DTD used to create this file in the format "#.#", e.g. "1.0".
`fileName`	Filename of the sequence listing file.
`softwareName`	Name of the software that generated this file.
`softwareVersion`	Version of the software that generated this file.
`productionDate`	Date of production of the sequence listing file (format "yyyy-mm-dd").

Example:

```
<ST26SequenceListing dtdVersion="1.0" fileName="US11-405455-SEQL.xml"
softwareName="BISSAP" softwareVersion="1.0" productionDate="2006-05-10">

</ST26SequenceListing>
```

General information part

28. The elements of the general information part relate to patent application information, and contain the following concepts:

Element	**Description**	**Mandatory/Optional**
`ApplicantFileReference`	Unique identifier assigned by applicant to identify a particular application	Mandatory when a sequence listing is furnished at any time prior to assignment of the application number; otherwise, Optional

Element	Description	Mandatory/Optional
`ApplicationNumber`	Information concerning the patent application for which the sequence listing is submitted	Mandatory when a sequence listing is furnished at any time following the assignment of the application number
The `ApplicationNumber` is composed of:		
`FilingOfficeCode`	WIPO ST.3 Code of the office of filing	Mandatory
`ApplicationNumberText`	The application number as provided by the office of filing	Mandatory
`ApplicationFilingDate`	Date of filing of the patent application for which the sequence listing is submitted	Mandatory
`PriorityApplicationNumber`	Application Number of the earliest priority claim (also contains `FilingOfficeCode` and `ApplicationNumberText` elements, see `ApplicationNumber` above)	Mandatory where priority is claimed
`EarliestPriorityDate`	Date of filing of the earliest priority claim (format "yyyy-mm-dd")	Mandatory where priority is claimed
`ApplicantName`	Name of the first mentioned applicant in the Latin alphabet	Mandatory
`ApplicantNameCharacters`	Name of the first mentioned applicant in characters other than the Latin alphabet	Optional
`InventorName`	Name of the first mentioned inventor in the Latin alphabet	Mandatory
`InventorNameCharacters`	Name of the first mentioned inventor in characters other than the Latin alphabet	Optional
`InventionTitle`	Title of the invention in the Latin alphabet	Mandatory
`InventionTitleCharacters`	Title of the invention in characters other than the Latin alphabet	Optional
`SequencesTotalNumber`	The total number of all sequences in the sequence listing including intentionally skipped sequences (see paragraph 9).	Mandatory

Example:

```
<ST26SequenceListing fileName="US11-405455-SEQL.xml"
productionDate="2006-05-10" softwareName="BISSAP" softwareVersion="1.0"
dtdVersion="1.0">
    <ApplicantFileReference>AB/123</ApplicantFileReference>
    <ApplicationNumber>
        <FilingOfficeCode>US</FilingOfficeCode>
        <ApplicationNumberText>11/405,455</ApplicationNumberText>
    </ApplicationNumber>
    <ApplicationFilingDate>2006-04-16</ApplicationFilingDate>
    <PriorityApplicationNumber>
        <FilingOfficeCode>JP</FilingOfficeCode>
        <ApplicationNumberText>2001-209712</ApplicationNumberText>
    </PriorityApplicationNumber>
    <EarliestPriorityDate>2001-07-10</EarliestPriorityDate>
    <ApplicantName>Ajinomoto Co., Inc.</ApplicantName>
    <InventorName>Yasuhiro Takenaka</InventorName>
    <InventionTitle>DNA FOR ENCODING D HYDANTOIN HYDROLASES
</InventionTitle>
    <SequencesTotalNumber>9</SequencesTotalNumber>
    <SequenceData sequenceIDNumber="1"> {...}* </SequenceData>
    <SequenceData sequenceIDNumber="2"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="3"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="4"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="5"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="6"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="7"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="8"> {...} </SequenceData>
    <SequenceData sequenceIDNumber="9"> {...} </SequenceData>
</ST26SequenceListing>

*{. . .} signifies relevant information for each sequence that has not
been included in this example.
```

29. The name of the applicant and the name of the inventor shall be indicated in the elements ApplicantName and InventorName, respectively, in characters of the Latin alphabet. The name of the applicant and the name of the inventor may also be indicated in characters other than those of the Latin alphabet in the element ApplicantNameCharacters or InventorNameCharacters, respectively, provided the element ApplicantName or InventorName, as appropriate, contains a transliteration or translation of the name into English. The title of the invention shall be indicated in the element InventionTitle in characters of the Latin alphabet. The title of the invention may also be indicated in characters other than those of the Latin alphabet in the element InventionTitleCharacters, provided the element InventionTitle contains a translation of the title into English.

Example:

```
<ApplicantName>Shutsugan Pharmaceuticals Kabushiki Kaisha</ApplicantName>
<ApplicantNameCharacters>出願製薬株式会社</ApplicantNameCharacters>
<InventorName>Taro Tokkyo</InventorName>
<InventorNameCharacters>特許　太郎</InventorNameCharacters>
<InventionTitle>Mus musculus abcd-1 gene for efg protein</InventionTitle>
<InventionTitleCharacters>efg タンパク質のためのマウス abcd-1 遺伝子
</InventionTitleCharacters>
```

Sequence data part

30. The sequence data part is composed of one or more SequenceData elements, each element containing information about one sequence.

31. Each SequenceData element has a mandatory attribute sequenceIDNumber, in which the

sequence identification number (see paragraph 9) for each sequence is contained.

Example:

```
<SequenceData sequenceIDNumber="1">
```

32. The SequenceData element contains a dependent element INSDSeq, which in turn contains further dependent elements as follows:

Element	Description	Mandatory*/Optional
INSDSeq_length	Length of the sequence	Mandatory
INSDSeq_moltype	Molecule type	Mandatory
INSDSeq_division	Indication that a sequence is related to a patent application	Mandatory with the value "PAT"
INSDSeq_feature-table	List of annotations of the sequence	Mandatory, except for intentionally skipped sequence(s)
INSDSeq_sequence	Sequence	Mandatory

*See paragraph 37 for intentionally skipped sequences.

33. All mandatory elements must be populated, other than the elements INSDSeq_length and INSDSeq_moltype when provided for intentionally skipped sequences as described in paragraph 37. Optional elements for which content is not available should not appear in the XML instance of the sequence listing.

34. The element INSDSeq_length must disclose the number of nucleotides or amino acids of the sequence contained in the INSDSeq_sequence element, except for any intentionally skipped sequences, *see* paragraph 37.

Example :

```
<INSDSeq_length>8</INSDSeq_length>
```

35. The element INSDSeq_moltype must disclose the type of molecule that is being represented. For nucleotide sequences, the molecule type must be indicated as DNA or RNA. For protein or polypeptide sequences, the molecule type must be indicated as AA.

Example :

```
<INSDSeq_moltype>AA</INSDSeq_moltype>
```

Where a nucleotide sequence contains both DNA and RNA fragments, the value for INSDSeq_moltype shall be “DNA.” The combined DNA/RNA molecule must be further described in the feature table, using the feature key “source” and the mandatory qualifier “organism” with the value “Synthetic Construct” and the mandatory qualifier “mol_type” with the value “other DNA.” Each DNA and RNA fragment of the combined DNA/RNA molecule should be further described with the feature key “misc_feature.”

Example:
Description of a nucleotide sequence containing both DNA and RNA fragments

```
<INSDSeq>
    <INSDSeq_length>120</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
        <INSDFeature>
            <INSDFeature_key>source</INSDFeature_key>
            <INSDFeature_location>1..120</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>other DNA</INSDQualifier_value>
                </INSDQualifier>
            </INSDFeature_quals>
        </INSDFeature>
        <INSDFeature>
            <INSDFeature_key>misc_feature</INSDFeature_key>
            <INSDFeature_location>1..60</INSDFeature_location>
            <INSDFeature_quals>
                <INSDQualifier>
                    <INSDQualifier_name>note</INSDQualifier_name>
                    <INSDQualifier_value>DNA
fragment</INSDQualifier_value>
                </INSDQualifier>
            </INSDFeature_quals>
        </INSDFeature>
        <INSDFeature>
            <INSDFeature_key>misc_feature</INSDFeature_key>
            <INSDFeature_location>61..120</INSDFeature_location>
            <INSDFeature_quals>
                <INSDQualifier>
                    <INSDQualifier_name>note</INSDQualifier_name>
                    <INSDQualifier_value>RNA
fragment</INSDQualifier_value>
                </INSDQualifier>
            </INSDFeature_quals>
        </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>
 cgacccacgcgtccgaggaaccaaccatcacgtttgaggacttcgtgaaggaattggataatacccgtccct
accaaaatggcgagcgccgactcattgctcctcgtaccgtcgagcggc
    </INSDSeq_sequence>
</INSDSeq>
```

36. The element `INSDSeq_sequence` must disclose the sequence. For intentionally skipped sequences, *see* paragraph 37. The residues in the sequence must be presented contiguously using only the appropriate symbols set forth in Annex B.1, paragraph 1, Table 1 and paragraph 3, Table 3. The sequence must not contain numbers, punctuation or whitespace characters.

37. The following elements must be provided for intentionally skipped sequences to preserve the numbering of subsequent sequences:

- the element SequenceData and its attribute `sequenceIDNumber`, with the sequence identification number of the skipped sequence provided as the value;

- the elements `INSDSeq_moltype`, `INSDSeq_length`, `INSDSeq_division`, present with no value provided;

- INSDSeq_sequence with the string "000" as the value.

Example :

```
<SequenceData sequenceIDNumber="3">
    <INSDSeq>
        <INSDSeq_length />
        <INSDSeq_moltype />
        <INSDSeq_division />
        <INSDSeq_sequence>000</INSDSeq_sequence>
    </INSDSeq>
</SequenceData>
```

Feature Table

38. The element INSDSeq_feature-table contains information on the location and roles of various regions within a particular sequence using controlled vocabulary. A feature table is required for every sequence, except for any intentionally skipped sequences, in which case it must not be included. The feature table is contained in the element INSDSeq_feature-table, which contains one or more INSDFeature elements.

39. Each INSDFeature element describes one feature, and contains the following further elements:

Element	Description	Mandatory/Optional
INSDFeature_key	A word or abbreviation indicating a feature	Mandatory
INSDFeature_location	Region of the presented sequence which corresponds to the feature	Mandatory
INSDFeature_quals	Qualifier containing auxiliary information about a feature	Mandatory where the feature key requires one or more qualifiers, e.g. source; otherwise, Optional

Feature keys

40. Annex B.1 contains an exclusive listing of feature keys that must be used under this standard, together with further information concerning mandatory and optional qualifiers. The exclusive listing of feature keys for nucleotide sequences is set forth in paragraph 5 and for amino acid sequences is set forth in paragraph 7.

Mandatory Feature Keys

41. The “source” feature key is mandatory for all for nucleotide sequences and the “SOURCE” feature key is mandatory for all amino acid sequences, except for any intentionally skipped sequences.

Certain feature keys require the presence of another feature key referred to as a “Parent Key”, e.g. the C_region feature key requires the CDS feature key (see Annex B.1).

Feature Location

42. The mandatory element INSDFeature_location contains at least one location descriptor, which defines a site or a region corresponding to a feature of the INSDSeq_sequence, and may contain one or more location operator(s) (see paragraphs 45 – 46).

43. The location descriptor can be a single residue number, a site between two adjacent residue numbers, a region delimiting a contiguous span of residue numbers, or a site or region that extends beyond the specified residue or span of residues. Multiple location descriptors must be used in conjunction with a location operator when a feature corresponds to discontinuous sites or regions of the sequence (see paragraphs 45 – 46). The location descriptor must not include numbering for residues beyond the range of the sequence in the `INSDSeq_sequence` element.

44. The syntax for each type of location descriptor is indicated in the table below, where x and y are residue numbers, indicated as non-negative integers, contained in the `INSDSeq_sequence` element, and x is less than y. In an XML instance of a sequence listing, the syntax characters "<" and ">" must be replaced by the appropriate predefined entities (*see* paragraph 26).

Location descriptor type	**Syntax**	**Description**
Single residue number	`x`	Points to a single residue in the presented sequence.
Residue numbers delimitating a sequence span	`x..y`	Points to a continuous range of residues bounded by and including the starting and ending residues.

Residues before the first or beyond the last specified residue number	`<x, >x,`
	`<x..y`
	`x..>y`
		Points to a region including a specified residue or span of residues and extending beyond a specified residue. The '<' and '>' symbols may be used with a single residue or the starting and ending residue numbers of a span of residues to indicate that a features extends beyond the specified residue number.

A site between two adjoining residue numbers	`x^y`	Points to a site between two adjoining residues, e.g. endonucleolytic cleavage site, The position numbers for the adjacent residues are separated by a carat (^). The permitted formats for this descriptor are x^x+1 (for example 55^56), or, for circular nucleotides, x^1, where "x" is the full length of the molecule, i.e. 1000^1 for circular molecule with length 1000.

45. A location operator is a prefix to either one location descriptor or a combination of location descriptors corresponding to a single but discontinuous feature, and specifies where the location corresponding to the feature on the indicated sequence is found or how the feature is constructed. A list of location operators is provided below with their definitions.

Location operator for nucleotides and amino acids:

Location syntax	**Location description**
`join(location,location, ... location)`	The indicated locations are joined (placed end-to-end) to form one contiguous sequence.
`order(location,location, ... location)`	The elements are found in the specified order but nothing is implied about the reasonableness of joining them.

Location operator for nucleotides only:

Location syntax	**Location description**
`complement(location)`	Indicates that the feature is located on the strand complementary to the sequence span specified by the location descriptor, when read in the 5' to 3' direction.

46. The location operator "complement" can be used for nucleotides only. "Complement" can be used in combination with either "join" or "order" within the same location. Combinations of "join" and "order" within the same location are not permitted.

47. Examples of feature locations are given in the table hereunder. In an XML instance of a sequence listing, the characters "<" and ">" must be replaced by the appropriate predefined entities (see paragraph 26).

Locations for nucleotides and amino acids:

Location Example	Description
`467`	Points to residue 467 in the sequence.
`123^124`	Points to a site between residues 123 and 124.
`340..565`	Points to a continuous range of residues bounded by and including residues 340 and 565.
`<1`	Points to a feature location before the first residue.
`<345..500`	Indicates that the exact lower boundary point of a feature is unknown. The location begins at some residue previous to 345 and continues to and includes residue 500.
`<1..888`	Indicates that the feature starts before the first sequence residue and continues to and includes residue 888.
`1..>888`	Indicates that the feature starts as the first sequenced residue and continues beyond residue 888.
`join(12..78,134..202)`	Indicates that regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.

Locations for nucleotides only:

Location example	Description
`complement(34..126)`	Start at the base complementary to 126 and finish at the base complementary to base 34 (the feature is on the strand complementary to the presented strand).
`complement(join(2691..4571, 4918..5163))`	Joins bases 2691 to 4571 and 4918 to 5163, then complements the joined segments (the feature is on the strand complementary to the presented strand).
`join(complement(4918..5163),complement(2691..4571))`	Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the feature is on the strand complementary to the presented strand).

Feature Qualifiers

48. Qualifiers are used to supply information about features in addition to that conveyed by the feature key and feature location. There are three types of value formats to accommodate different types of information conveyed by qualifiers, namely:

- free text (see paragraphs 53 - 54);
- controlled vocabulary or enumerated values (e.g. a number or date); and
- sequences.

49. The exclusive listing of qualifiers and their specified value formats, if any, for each nucleotide feature key is contained in Annex B.1, paragraph 6, and the exclusive listing of qualifiers for amino acid feature keys is contained in Annex B.1, paragraph 8.

Mandatory Feature Qualifiers

50. One mandatory feature key, i.e. “source” for nucleotide sequences and “SOURCE” for amino acid sequences, requires two mandatory qualifiers, “organism” and “mol_type” for nucleotide sequences and “ORGANISM” and “MOL_TYPE” for amino acid sequences. Some optional feature keys also require mandatory qualifiers.

Qualifier Elements

51. The element INSDFeature_quals contains one or more INSDQualifier elements. Each INSDQualifier element represents a single qualifier and contains two further elements:

Element	Description	Mandatory/Optional
INSDQualifier_name	Name of the qualifier (see Annex B.1, paragraphs 6 and 8)	Mandatory
INSDQualifier_value	Value of the qualifier, if any, in the specified format (see Annex B.1, paragraphs 6 and 8)	Mandatory, when specified (see Annex B.1, paragraphs 7 and 8

52. The qualifier “organism” for nucleotide sequences and “ORGANISM” for amino acid sequences (see Annex B.1) must disclose the source, i.e., organism or origin, of the sequence that is being represented.

If the sequence is naturally occurring and the source organism has a Latin genus and species designation, that designation must be used as the qualifier value. The preferred English common name may be included in parentheses, following the Latin genus and species designation.

Example for a nucleotide sequence:

```
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>1..5164</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>organism</INSDQualifier_name>
                <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>mol_type</INSDQualifier_name>
                <INSDQualifier_value>genomic DNA</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
```

Example for a protein sequence:

```
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..174</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>ORGANISM</INSDQualifier_name>
                <INSDQualifier_value>Homo sapiens</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>MOL_TYPE</INSDQualifier_name>
                <INSDQualifier_value>protein</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
```

If the sequence is naturally occurring and the source organism does not have a Latin genus and species designation, such as a virus, then another acceptable scientific name must be used, e.g. “Canine adenovirus type 2” as the qualifier value.

Organism designations should be selected from a taxonomy database.

If the source of the sequence is natural, but the organism species is unknown, then the qualifier value must be indicated as “unclassified,” followed by any known taxonomic information. For example, an unknown bacterium “B8” could be indicated as: “unclassified, unidentified bacterium B8.”

Example for identification of an unknown bacterium:

```
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>unclassified, unidentified bacterium
B8</INSDQualifier_value>
```

If the sequence is not naturally occurring, the qualifier value must be indicated as “Synthetic Construct.” Further information with respect to the way the sequence was generated may be specified using the qualifier “note” for nucleotide sequences and the qualifier “NOTE” for amino acid sequences.

Example for a protein sequence that is not naturally occurring:

```
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>SOURCE</INSDFeature_key>
        <INSDFeature_location>1..40</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>synthetic peptide used as assay for
antibodies</INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
```

Free text

53. Free text is a type of value format for qualifiers, presented in the form of a descriptive text phrase that shall be composed of characters from the UNICODE Basic Latin code chart and should preferably be in the English language.

54. The use of free text shall be limited to a few short terms indispensible for the understanding of the sequence. For each qualifier, the free text shall not exceed 255 characters. Any further information may be included in the main part of the application in the language thereof.

Coding sequences

55. The “CDS” feature key may be used to identify coding sequences, i.e. sequences of nucleotides which correspond to the sequence of amino acids in a protein and the stop codon. The element `INSDFeature_location` should identify the location of the CDS feature and must include the stop codon.

56. The "transl_table" and "translation" qualifiers may be used with the “CDS” feature key (see Annex B.1). Use of the Standard Code Table (see Annex B.1, paragraph 9, Table 5) is assumed where the "transl_table" qualifier is not used.

57. A disclosed protein sequence that is encoded by the coding sequence and encompassed by paragraph 4 must be assigned its own sequence identification number and be presented in the sequence listing. The “source” feature key and “organism” qualifier for the protein sequence must correspond to that of its coding sequence.

Example:

```
<INSDSeq_feature-table>
    <INSDFeature>
        <INSDFeature_key>CDS</INSDFeature_key>
        <INSDFeature_location>1..507</INSDFeature_location>
        <INSDFeature_quals>
            <INSDQualifier>
                <INSDQualifier_name>transl_table</INSDQualifier_name>
                <INSDQualifier_value>11</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
                <INSDQualifier_name>translation</INSDQualifier_name>
                <INSDQualifier_value>
MLVHLERTTIMFDFSSLINLPLIWGLLIAIAVLLYILMDGFDLGIGILLPFAPSDKCRDHMISSIAPFWDGNE
TWLVLGGGGLFAAFPLAYSILMPAFYIPIIIMLLGLIVRGVSFEFRFKAEGKYRRLWDYAFHFGSLGAAFCQG
MILGAFIHGVEVNGRNFSGGQLM
                </INSDQualifier_value>
            </INSDQualifier>
        </INSDFeature_quals>
    </INSDFeature>
</INSDSeq_feature-table>
```

Variants

58. A variant sequence disclosed by enumeration of its residues and encompassed by paragraph 4 must be assigned its own sequence identification number and be presented in the sequence listing. A specific variant, i.e., deletion, addition, or substitution, disclosed only by reference to a disclosed primary sequence in the sequence listing, must be presented in the sequence listing either as a separate sequence assigned its own sequence identification number or by annotation of the primary sequence with appropriate feature keys and qualifiers. A specific variant containing multiple variations from the primary sequence at distinct locations, where the variations at each location only occur together, must be presented in the sequence listing as a separate sequence assigned its own sequence identification number.

Type of sequence	**Feature Key**	**Use**
Nucleic acid	variation	Naturally occurring mutations and polymorphisms, eg. Alleles, RFLPs
Nucleic acid	misc_difference	Variability introduced by genetic manipulation, e.g. site directed mutagenesis
Amino acid	VARIANT	Any type of variant
Amino acid	VAR_SEQ	Variant produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting

Example of a “variation” (substitution):
Feature key “variation” for a nucleotide sequence:
A cytosine replaces the nucleotide given in position 413 of the sequence.

```
<INSDFeature>
    <INSDFeature_key>variation</INSDFeature_key>
    <INSDFeature_location>413</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>replace</INSDQualifier_name>
            <INSDQualifier_value>c</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example of a “mis_difference” (deletion):
Feature key “misc_difference” for a nucleotide sequence containing a deletion:
The nucleotide at position 413 of the sequence is deleted.

```
<INSDFeature>
    <INSDFeature_key>misc_difference</INSDFeature_key>
    <INSDFeature_location>413</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>replace</INSDQualifier_name>
            <INSDQualifier_value></INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example of a “mis_difference” (addition):
Feature key “misc_difference” for a nucleotide sequence containing an addition:
An adenine is added between positions 100 and 101 of the sequence.

```
<INSDFeature>
    <INSDFeature_key>misc_difference</INSDFeature_key>
    <INSDFeature_location>100^101</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>replace</INSDQualifier_name>
            <INSDQualifier_value>a</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Example of a “VARIANT” (substitution):
Feature key “VARIANT” for an amino acid sequence:
A Leucine replaces the amino acid given in position 100 of the sequence .

```
<INSDFeature>
    <INSDFeature_key>VARIANT</INSDFeature_key>
    <INSDFeature_location>100</INSDFeature_location>
    <INSDFeature_quals>
        <INSDQualifier>
            <INSDQualifier_name>NOTE</INSDQualifier_name>
            <INSDQualifier_value>Leucine</INSDQualifier_value>
        </INSDQualifier>
    </INSDFeature_quals>
</INSDFeature>
```

Fields for Patent Offices Use only

59. In the context of data exchange with database providers, the Patent Offices should populate for each sequence the element `INSD_other-seqids` with one `INSDSeqid` containing a reference to the corresponding published patent and the sequence identification number in the following format:

```
PAT|{country code}|{publication number}|{sequence identification number}
```

REFERENCES

The following standards and documents are of relevance to this Standard:

(a) WIPO Standard ST.3 – Two-Letter Codes for the Representation of States, Other Entities and Intergovernmental Organizations;
(b) WIPO Standard ST.25 – Presentation of nucleotide and amino acid sequence listings
(c) WIPO Standard ST.36 – Processing of Patent Information Using XML
(d) WIPO Standard ST.96 – Processing of Industrial Property Information Using XML
(e) International Nucleotide Sequence Database Collaboration (INSDC): http://www.insdc.org/
(f) UniProt Consortium: http://www.uniprot.org/