WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

CWS/1/5
ORIGINAL: ENGLISH
DATE: OCTOBER 13, 2010

Committee on WIPO Standards (CWS)

First Session
Geneva, October 25 to 29, 2010

PROPOSAL FOR THE PREPARATION OF A NEW WIPO STANDARD ON THE PRESENTATION OF NUCLEOTIDE AND AMINO ACID SEQUENCE LISTINGS USING EXTENSIBLE MARKUP LANGUAGE (XML)

Document prepared by the Secretariat

1. On September 17, 2010, the European Patent Office (EPO) sent a document to the Secretariat requesting the initiation by the Committee on WIPO Standards (CWS) at its first session in October 2010 of discussions on a proposal for a new standard on the presentation of nucleotide and amino acid sequence listings based on eXtensible Markup Language (XML). The request by the EPO is reproduced in the Annex to this document.

2. In accordance with the request by the EPO, the Secretariat proposes the following for consideration and approval by the CWS:

   (a) the creation of a new task whose description would read as follows:

   "Prepare a recommendation on the presentation of nucleotide and amino acid sequence listings based on eXtensible Markup Language (XML) for adoption as a WIPO standard. The proposal of the new WIPO standard should be presented along with a report on the impact of the said standard on the current WIPO Standard ST.25, including the proposed necessary changes to Standard ST.25";

   (b) the establishment of a new Task Force, with its corresponding Task Force Leader, to handle the new task; and

   (c) a request that the Task Force present the proposal of the new WIPO standard and necessary changes to Standard ST.25 for consideration and approval by the CWS at its second session to be held in 2011.

3. The CWS is invited to:

(a) note the request by the European Patent Office for a new WIPO standard as reproduced in the Annex to this document;

(b) consider and approve the proposal concerning the creation of the task and corresponding time frame referred to in paragraphs 2(a) and (c), above; and

(c) consider and approve the establishment of the new Task Force, with its corresponding Task Force Leader, referred to in paragraph 2(b), above;

[Annex follows]

REQUEST BY THE EPO TO ESTABLISH A WIPO TASK FORCE

EPO proposal for a new standard for the presentation of sequence listings based on XML format

BACKGROUND AND REQUEST

1. Biological sequences in patent applications are disclosed as "sequence listings". Sequence listings are currently presented in accordance with WIPO Standard ST.25 both within the framework of the PCT (Annex C of the Administrative Instructions) and under most national or regional patent procedures.

2. The EPO is of the view that, for a number of technical and practical reasons, WIPO Standard ST.25 should be replaced, or at least supplemented, by a new standard based on XML format. Such new standard would mitigate the shortcomings of WIPO Standard ST.25 and provide additional advantages for both applicants and Offices since the drafting and submission of high quality sequence listings will enable more efficient downstream processes.

3. The EPO intends to submit its proposal for a new WIPO standard by the end of 2010 for adoption in 2011. The EPO therefore hereby requests that a specific Task Force be established by the Committee on WIPO Standards (CWS) with the mandate to draft the new standard based on the EPO's proposal. In support to the present request, the EPO wishes to provide the preliminary information, below.

PROBLEMS ENCOUNTERED WITH THE CURRENT WIPO STANDARD ST.25

4. It has been years since WIPO Standard ST.25 was approved whilst life sciences is a fast changing technical domain. WIPO Standard ST.25 is a format unique to the patent world. It is unfamiliar for inventors working directly with biological sequences, who are generally more familiar with the presentation of sequences according to the formats of public repositories. Hence it is unfeasible to leverage open source tools with WIPO Standard ST.25.

5. WIPO Standard ST.25 does not express the latest scientific requirements, e.g.:

   - There are more location types supported by public database providers (European Molecular Biology Laboratory (EMBL), DNA Databank of Japan (DDBJ), National Center for Biotechnology Information (NCBI)) than by WIPO Standard ST.25. Applicants must find out unconventional and frequently inaccurate ways to annotate sequences.

   - Features keys and qualifiers (controlled vocabularies used to depict characteristics of sequences) used in WIPO Standard ST.25 are not in use by the world sequence repositories.

   - Some of the International Union of Pure and Applied Chemistry (IUPAC) standard abbreviations are also not supported.

6. WIPO Standard ST.25 is error-prone. As it is designed to be both human and computer readable, it is easy to make mistakes, and those are difficult to detect. The relative ease of creating a sequence listing by non-dedicated software, such as text editors, leads to many errors, e.g.:

    - invalid sequences,
    - wrong number of sequences,
    - invalid organism names,
    - unauthorized characters,
    - inadequate features,
    - syntax and scientific information incorrectness due to the complexity of the so called mixed mode presentation of biological sequences.

7. It follows that public database providers drop the information provided by applicants and regenerate it. Offices which provide sequence listings to public repositories spend therefore much time and efforts to correct the data.

THE PROPOSED NEW STANDARD

8. The new standard proposed by the EPO would solve the inconveniences faced with WIPO Standard ST.25 and help to foster the quality of sequences in patent and public sequence databases. It is also noted that WIPO ST.36 recommends the usage of industry standards[1].

9. The new standard will therefore have the following features:

    - universality: a common format for patent and non patent communities,
    - robustness and flexibility: an XML format based on an agreed DTD.

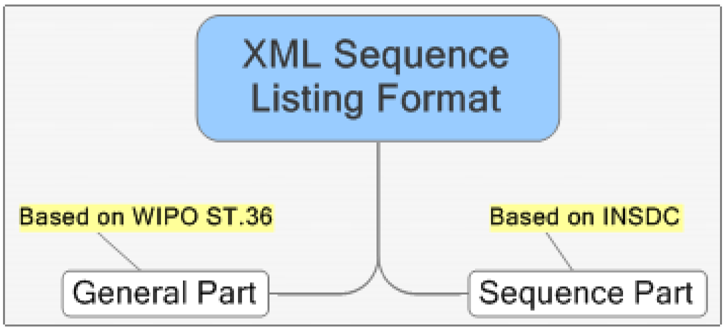
Universality

10. A sequence listing consists in a “general information part” depicting information related to the patent application and a “sequence part” consisting in a variable number of biological sequences.

11. The public data providers, which form the International Nucleotide Sequence Database Collaboration (INSDC) (consisting of the European Bioinformatics Institute (EBI), DDBJ, NCBI), have designed an XML based format for the exchange of sequences and related information called INSDSeq. This format was designed to facilitate the addition of the relevant patent related information.

[1] See § 93 of WIPO Standard ST.36: Where appropriate to the content of a document, that is, where the content is not unique to the industrial property domain, use industry-standard DTDs.

12. The new standard for the presentation of sequence listings should therefore be using INSDSeq to disclose the sequence information and amend it with patent related information (general information of sequence listing) in accordance with WIPO ST.36. The aim is to ensure the synchronization of the sequence part of WIPO's sequence listing format with INSDSeq.



13. In effect, the new standard based on XML format would be more user-friendly as scientists would be using almost the same format for both patent applications and submission to public databases with no need to make conversions. The mere familiarity with the syntax of the standard could in itself lead to fewer errors.

Robustness and flexibility

14. The syntax of the new standard provided by the DTD will be both more precise and easier to verify by means of automatic tools. It could be verified by a vast set of existing freeware.

15. The contents of an XML file, while more complicated for humans to read without a style sheet, is actually easier to manipulate by computers, and there are extensive libraries for this purpose.

STATUS

16. A draft DTD is under discussion between the Trilateral Offices. The agreed DTD will form the basis of the proposal.

17. The EPO has developed, in cooperation with the EBI, a client software for the submission of biological sequences (Biological Sequence Submission Application for Patents (BiSSAP)) which includes a verification module and which contains an option to create XML files. This software is therefore both WIPO Standard ST.25 and INSDSeq / WIPO ST.36 XML compliant. The software has been tested by European users in August with positive results. The time schedule for implementation and the making available of this new tool to the public is currently under discussion at the EPO.

[End of Annex and of document]