WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

PCT/MIA/27/8
ORIGINAL: ENGLISH ONLY
DATE: JANUARY 17, 2020

Meeting of International Authorities under the Patent Cooperation Treaty (PCT)

Twenty-Seventh Session
Gatineau, February 6 and 7, 2020

IMPLEMENTATION OF WIPO STANDARD ST.26

Document prepared by the International Bureau

SUMMARY

1. Most of the work required for the implementation of WIPO Standard ST.26 is progressing well towards the target date for use in all national and international applications filed on or after January 1, 2022. Discussions are under way concerning a suitable technical means for handling language-dependent free text. Legal proposals for amendments to the PCT Regulations and modifications to the PCT Administrative Instructions will be based on ensuring an effective framework for submission and use of both listings and associated free text.

2. International Authorities should indicate whether any new special factors have been identified that imply a need for further developments to ensure that they are able to use ST.26 sequence listings effectively for search, including both understanding and searching sequences as part of the search and examination of the particular application, and in ensuring that sequences can be added effectively to databases for more effective disclosure and later searches.

BACKGROUND

3. The Committee on WIPO Standards (CWS), at its fourth reconvened session in March 2016, adopted WIPO Standard ST.26 “Recommended Standard for the Presentation of Nucleotide and Amino Acid Sequence Listings using XML (eXtensible Markup Language)”. The CWS further revised ST.26 at its fifth session in May/June 2017, at its sixth session in October 2018 and its seventh session in July 2019. The most recent version of WIPO Standard ST.26 (version 1.3) is available on the WIPO website in Part 3 of the *Handbook on Industrial*

Property Information and Documentation.[1] In order to support the implementation of WIPO Standard ST.26, the International Bureau develops the WIPO Sequence tools. The beta version of tools is available for testing in WIPO website at: https://www.wipo.int/standards/en/sequence/index.html.

4. At the fifth session of the CWS in May/June 2017, the CWS agreed on January 1, 2022, as the date for transition from WIPO Standard ST.25 to ST.26, to be applicable to any national or international application filed on or after that date , and requested the Sequence Listings (SEQL) Task Force (see paragraphs 44 and 45 of the Report of the session, document CWS/5/22):

(a) to support the International Bureau by providing users' requirements and feedback on the authoring and validation tool;

(b) to support the International Bureau in the consequential revision of the PCT Administrative Instructions; and

(c) to prepare necessary revisions of WIPO Standard ST. 26 upon request by the CWS.

5. Implementation of WIPO Standard ST.26 in respect of international applications filed on or after January 1, 2022, will require amendments to the PCT Regulations to be adopted by the PCT Assembly in September 2020, to enter into force on January 1, 2022. Major modifications to the Administrative Instructions will also be required.

6. Preliminary draft proposals for such amendments and modifications were considered at the twelfth session of the PCT Working Group (see document PCT/WG/12/13). Most of the general principles were considered acceptable, but it was apparent that further work was required in relation to the handling of language-dependent free text within the sequence listings.

LANGUAGE-DEPENDENT FREE TEXT

7. WIPO Standard ST.26 is an XML format, designed for effective interactions with the databases used for search and information dissemination. Most of the content is thus codes designed to be read by machines, representing nucleotide and amino acid sequences, or providing markup of features of the sequences. While many of the codes are made up from words in the English language, they are in fact language-neutral descriptors of the technical content or related information, such as numbers and dates.

8. Some of the "qualifiers" that provide information on the features contain content described by the Standard as "free text". The free text content contains material in a less structured format. Some of this free text represents data such as reference numbers, which are always language-independent. However, other parts of the free text information represents general descriptive text, which most national patent laws will expect to be provided in the language of the description, as used for national processing. The CWS SEQL Task Force is currently considering a proposal to distinguish between qualifiers whose free text is always language-independent and ones whose free text is sometimes or always language-dependent. This proposal is expected to be adopted by the CWS at its eighth session in July 2020.

9. WIPO Standard ST.25 allows for free text to be associated with sequences. This information is not imported into search databases, in part because the purpose of the free text and the region of the sequence with which it is associated is not coded effectively, but mainly because the information may be provided in any language and is often not easily read because of differences in "code pages" used to encode the files. The database suppliers currently being transmitted sequence listings by major national Offices (from both English-speaking and other

[1] https://www.wipo.int/export/sites/www/standards/en/pdf/03-26-01.pdf

countries) have indicated that they have no use for annotations that are not in English and do not currently load the free text from ST.25 files, irrespective of language.

10. One of the key aims of introducing WIPO Standard ST.26 was to enrich the databases used for the search and dissemination of sequence information by ensuring that free text information is imported along with the sequences. This will provide both more effective disclosure of sequences for use by researchers and improved searches on later patent applications. To this end, the structures used to represent sequences and the associated features and qualifiers were taken directly from an industry standard. This standard allows the use of only "printable characters (including the space character) from the Unicode Basic Latin code table". WIPO Standard ST.26 adopted that limitation, but then only states that free text in a listing should "preferably" be in the English language.

11. Because accented and non-Latin characters cannot appear in a valid ST.26 file, most annotations would be impossible to present effectively in a language other than English. Nevertheless, discussions in the PCT Working Group (see paragraphs 136 and 137 of document PCT/WG/12/24) implied that it was not likely to be acceptable to work with listings in English only for the international phase of international applications.

12. Consequently, the International Bureau has informally presented a technical proposal to the CWS SEQL Task Force, whereby an international application related to sequence listings could, where appropriate, include three parts within the description:

(a) a sequence listing meeting the current requirements of WIPO Standard ST.26 (potentially, with a minor variation to add "id" attributes to certain elements, which would be entirely transparent to applicants but allow simpler and more reliable technical systems to be used);

(b) the main body of the description, referring to sequences within the sequence listing but not normally repeating any of the technical content; and

(c) a "language file", containing the free text of the sequence listing in a machine-readable form in the language of the main body of the application.

13. The proposal for a language file uses the XLIFF[2] 2.0 file specification. XLIFF is a standard approved as ISO 21720:2017 and recognized by a wide range of professional translation software. The purpose of that standard is to provide a consistent way of taking language-dependent text from a file or set of files and to provide equivalent text in a second language in a manner that allows new files to be automatically constructed, representing the information "localized" in the second language.

14. WIPO Sequence would generate language files for use in two different ways:

(a) For translation, a file would be prepared containing all the free text from a sequence listing that appeared within qualifiers identified as "language dependent" as "source" elements. This would be passed to a translator, who would add equivalent "target" texts in the second language. The proposal provides for including "notes" to give additional contextual information to the translator if so required. The XLIFF standard itself includes rules for compliant "agents" stating how the translations should be added, in particular

[2] XML Localization Interchange File Format: http://docs.oasis-open.org/xliff/xliff-core/v2.0/os/xliff-core-v2.0-os.html. XLIFF version 2.1 was adopted by OASIS in 2018 and is backwards compatible with XLIFF 2.0, but the proposal uses only features from version 2.0 in order to maximize compatibility with translation systems currently in use.

requiring that the metadata indicating how the source text corresponds to the original file from which it is taken should not be modified.

(b) For transmission to an Office as part of the filing of an international application or of a translation thereof, a similar file would be provided containing both the “source” and the target” texts, but not including any notes that might cause uncertainty over whether they formed part of the disclosure that was being submitted.

15. The sequence listing and language files would contain information embedded within them to ensure that Offices’ systems would have a simple way of checking that a language file uploaded to an Office system correctly corresponded to a sequence listing received either at the same time or earlier, without needing to check item-by-item that the source texts corresponded correctly and completely with the expected qualifier texts within the listing (though WIPO Sequence Validator might be enhanced to offer such functionality where desired, as noted below).

16. Once received by an Office, the language file would be treated as a part of the description (whether that of the international application as filed, or of a translation into another language for publication, search or preliminary examination, or national phase entry). It would meet the need to provide the free text in the language of the main body of the description as if the text were directly included in the sequence listing file. From the perspective of the examiner, the tools provided to allow a human view of the sequences would allow the free text annotations to be shown in the language of the sequence listing (normally English), the language from the language file, or possible both simultaneously.

17. Some of the likely technical implications of this proposal include developments to WIPO Sequence and related tools as follows:

(a) Allow WIPO Sequence to export XLIFF files with validations and contents appropriate to: (i) transmission to translators; and (ii) submission to Offices; allow re-import and appropriate validation of XLIFF files received back from translators, with meaningful warnings where inconsistencies are found.

(b) Improve the language features of WIPO Sequence to allow the direct entry of appropriate language information in cases where the applicant or agent preparing the file is able to supply the appropriate language text themselves.

(c) Allow WIPO Sequence Validator to provide effective cross-validation of sequence listings and language files to ensure that they properly match.

(d) Allow the Display Sequence Listing function of WIPO Sequence to take language files into account.

(e) Provide the Display Sequence Listing functionality as a module that can be integrated into other systems (potentially including internal Office systems, ePCT and PATENTSCOPE), allowing for consistent displays of sequences in different language versions into account, without depending on a full installation of WIPO Sequence.

18. In addition, consideration will be needed of whether to provide additional import-export options for translation, for example using spreadsheet files. These would have significantly greater opportunities for accidentally introducing errors by changing the structure of the spreadsheet or editing cells containing reference information and would be more difficult to provide effective feedback to applicants and agents in cases where the language information did not properly correspond with the information in the sequence listing. However, this may be necessary in order to allow use by the fullest range of translators. Alternatively, WIPO Sequence might be accompanied by an associated XLIFF editing tool, allowing easy

modification of the target elements of a language file without giving direct access to the sequence listing itself. The new required features will be further evaluated later once Offices have agreed on the proposals presented in this document, in particular paragraph 12 mentioned above.

OTHER ISSUES

19. Informal discussions are under way concerning whether it would be appropriate to allow continued use of WIPO Standard ST.25 for certain special cases filed after January 1, 2022, such as divisional applications based on earlier applications containing ST.25 listings. Such applications would not be used as the basis for claiming priority. Moreover, the PCT has no provision for divisionals and only limited options for indicating that the international application is based on an earlier "parent" application other than through claiming priority. Consequently, this issue does not appear significant for the PCT.

20. Sequence listings in accordance with WIPO Standard ST.26 can only be submitted as electronic files. Consequently, it will be essential to find ways to deal effectively with a small number of special cases, notably:

(a) international applications with sequence listings larger than the maximum file size permitted for upload to Offices' filing and document upload services; and

(b) international applications where the main body of the application is submitted on paper.

21. Both of these are expected to involve extremely small numbers of cases, provided that applicants have the option of submitting their sequence listings as compressed ZIP files. Few applicants preparing international applications requiring a sequence listing to be filed would willingly file the application on paper or submit a later required listing other than by document upload when online services are available and effective for the relevant purpose. The IT systems and legal framework will need to ensure that such exceptional cases can be handled effectively (both at the time of submission by the applicant and with regard to later processing by the various Offices concerned) without adding provisions that encourage inefficient behaviors.

22. The main further issue for consideration by International Authorities at this stage is whether any practical issues have been identified in the process of searching an international application containing sequence listings in ST.26 format that may require the development of tools not yet envisaged. The schedule for development of WIPO Sequence and related software is such that any additional functionality will need to be agreed before the eighth session of the CWS in July 2020 if it is to be developed and properly tested in time for release before January 1, 2022.

23. The Meeting is invited to comment on the issues set out in this document.

[End of document]