

Committee on WIPO Standards (CWS)

Thirteenth Session
Geneva, November 10 to 14, 2025

DIGITIZATION OF INTELLECTUAL PROPERTY DOCUMENTATION

Document prepared by the International Bureau

SUMMARY

1. The Members of the Committee on WIPO Standards (CWS) have indicated that one of their main challenges is the digitization of their intellectual property (IP) documentation. At the same time, patent authorities seeking to include their patent documents as part of the Patent Cooperation Treaty (PCT) minimum documentation are currently working on meeting the requirements set out in the PCT Administrative Instructions. This document details some of the challenges experienced by Offices and sets out the need for producing a set of digitization guidelines, focusing initially on patent collections.

BACKGROUND

2. In 2024, the Secretariat conducted a survey of CWS Members to understand the problems that they may experience exchanging intellectual property data and potential solutions to these issues. Survey results revealed that one of the biggest issues experienced by Offices, regardless of their size, relate to IP gazette information not being available in machine-readable format and that there is insufficient resourcing, both in terms of staff skill gaps and IT resourcing to support these activities (see paragraph 8 of document CWS/12/23).

3. The PCT Minimum Documentation Task Force (the “Task Force”), led by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), was established in 2005 by the Meeting of International Authorities (MIA) under the PCT to perform a comprehensive review of the PCT minimum documentation. However, work did not commence in earnest until early 2017, where a workplan was defined as achieving the following four objectives (see Appendix to document PCT/MIA/24/4):

- (a) Objective A: Create an up-to-date inventory of the patent literature and non-patent literature parts of the current PCT minimum documentation;
- (b) Objective B: Recommend criteria and standards for including a national patent collection in the PCT minimum documentation;
- (c) Objective C: Propose clearly defined bibliographic and text components of patent data that should be present in patent collections belonging to the PCT minimum documentation; and
- (d) Objective D: Recommend criteria and standards for the review, addition and maintenance of non-patent literature and traditional knowledge based prior-art and afterwards assess on the basis of the criteria that will have been established, the revised proposal from the Indian authorities on the Indian Traditional Knowledge Digital Library database.

4. Following discussions in the Task Force and the PCT Working Group, in July 2023, the PCT Union Assembly adopted amendments to Rules 34, 36 and 63 of the PCT Regulations with effect from January 1, 2026 (document PCT/A/55/2). Modifications to the PCT Administrative Instructions were promulgated in June 2024 (Circular C. PCT 1672), including the addition of new Annex H. Annex H details technical requirements which aim to ensure that all patent and utility model collections belonging to the PCT minimum documentation are available free of charge to each International Searching Authority (ISA) in a machine-readable format at least for publications on or after January 1, 2026, with publications at least back to 1991 to be made available in full text format over the following ten years.

5. Annex H specifies that there are three allowable formats for the text-searchable data: XML instances compliant with either WIPO Standard ST.36 or WIPO Standard ST.96 or plain-text format. Authorities can select the format that they determine most suitable as long as it supports a high-quality search of their patent collections. The International Bureau recommends the use of XML, at least for providing bibliographic data. Offices selecting plain-text as their preferred format are asked to consider migrating to one of the WIPO Standards, in an attempt to improve the quality of their patent publications and their searchability.

PREPARATIONS

6. To support ISAs in meeting the technical requirements in advance of the deadline of January 1, 2026, the International Bureau hosted a series of bilateral PCT minimum documentation clinics in 2025. Offices prepared in advance the questions they had for the International Bureau and the International Bureau ensured that the appropriate expert was present at the clinic to address these questions. During these sessions, it was noted that there was no guidance provided in PCT Administrative Instructions regarding the quality of full-text documents. On request, the International Bureau shared with the CWS Authority File Task Force its idea for a set of minimum requirements for what they consider a 'good-enough' machine-readable document.

7. For Offices who are digitizing their patent documents, the following provides guidance for alignment with the requirements of Annex H:

- For XML instances compliant with WIPO ST.36, these must be well formed and have a consistent root element and minimal OCR errors. Bibliographic data should include at least the application reference, the publication reference and the invention title;
- For XML instances compliant with WIPO ST.96, these should comply with the WIPO ST.96 document-level component `PatentPublication` but there is no need to implement `MathML` or `OASISTable`; and

- For plain text files, these should have minimal OCR errors, have text presented as proper paragraphs, no inclusion of line or page numbers and allow proper reading order. Preferably they should be accompanied by bibliographic data in an XML format.

8. The PCT Minimum Documentation Task Force agreed to a three-phase certification process for ISAs or other patent Offices seeking their collections to be part of the PCT minimum documentation. The second stage or "verification stage" tested whether an Office can provide a data repository containing documents compliant with one of the three formats indicated in Annex H to be downloaded and can produce an authority file compliant with WIPO Standard ST.37, including the necessary text-searchable indicators.

CHALLENGES EXPERIENCED BY OFFICES

9. During the PCT minimum documentation clinics and the verification stage, the International Bureau provided feedback to assist in the development of machine-readable patent publications and the production of WIPO ST.37-compliant authority files. Following this review process, common challenges experienced by Offices during the production of machine-readable documents included:

- (a) For all formats, poor quality OCR resulting in numerous errors, and as such polluting search indexes, especially for patent documents with chemical formulae;
- (b) For all formats, poor layout detection or handling, resulting in the inclusion of header or footer information within the text, or line numbers appearing in the output;
- (c) For all formats, poor segmentation of description, claims and drawings sections, resulting for example in description text being included in the claims full-text;
- (d) For all formats, badly tagged language for full-text sections, having the consequence of subsequent critical indexation issues in search systems;
- (e) For WIPO ST.96-compliant instances, invalid instances which have not been validated against the `PatentPublication` document-level component and confusion around the use of the language codes within the hierarchy;
- (f) For WIPO ST.36-compliant instances, a misunderstanding of what is considered to be a 'valid' instance and what is the minimum standard when considering whether the XML document is 'good enough'. For instance, if the Model DTD from ST.36 have to be used; and
- (g) For the plain text format, text not provided as proper paragraphs and instead appearing as lines, which makes the contents difficult to search and bibliographic data being provided in non-WIPO Standard formats.

TECHNICAL ASSISTANCE

10. In order to address some of the challenges raised above, the International Bureau provides technical assistance to Offices seeking support. The *WIPO OCR Solution* supports the production of WIPO ST.36-compliant XML instances based on image-based and text-searchable bookmarked PDFs. However, the CWS should note that this tool does not support claim number detection and tagging and treats tables and complex chemical and mathematical formulae as embedded drawings.

11. *WIPO Publish* enables Offices to extract and quality-control their data in WIPO Standard ST.36 format and soon the output will be migrating to be compliant with WIPO ST.96. This facilitates the exchange of gazette information for many IP Offices. In 2025, the International

Bureau added features to extract patent documents for Offices implementing the PCT minimum documentation requirements. A pilot has been successfully completed with the Intellectual Property Office of the Philippines. The same approach could be used for other patent Offices whose collections are part of the PCT minimum documentation; and any interested Offices may wish to contact the International Bureau.

12. The International Bureau also provides direct assistance to around 30 Offices to support the digitization of their documentation, including full-text OCR for some Offices. This process involves the scanning of documents, digitization and then a quality-check is undertaken for the bibliographic data. In all cases, this work is outsourced to third party contractors.

PROPOSAL FOR NEXT STEPS

13. Noting the challenges above, the International Bureau would like to propose the production of a set of IP document digitization guidelines, to assist in bridging the gap in understanding which was noted above. In order to further examine the challenges experienced by Offices who are currently digitizing their IP documentation, the International Bureau proposes to conduct a survey questionnaire by issuing a joint CWS and PCT Circular. The International Bureau intends to prepare this survey questionnaire and to invite the members of both bodies to complete the survey in the first quarter of 2026.

14. While the discussion at this stage is focused on the digitization of patent collections, it is likely that these guidelines will be expanded in scope to incorporate further guidance on the digitization of trademark and industrial design documents.

15. Following the deadline of the survey, the International Bureau intends to organize a meeting to discuss the preparation of the digitization guidelines in the first half of 2026. All Members of the CWS and PCT will be invited, however details regarding the meeting including the meeting format and date(s) will be announced in due course. At this meeting, the International Bureau intends to report the outcomes of the survey results as well as a potential timeline for the production of the IP document digitization guidelines. The CWS Members are encouraged to participate in the survey questionnaire and this meeting as their feedback is required to produce the digitization guidelines. The International Bureau also intends to report the outcome of this meeting to the fourteenth session of the CWS.

16. In addition, the International Bureau is planning to conduct a workshop on WIPO Standard ST.36 and ST.96 in the second quarter of 2026. Invitations to this workshop will be sent to the CWS Members and Observers and the event will be advertised on the WIPO website. The International Bureau also invites Offices implementing WIPO Standards to reach out for technical assistance when needed.

17. *The CWS is invited to:*

(a) note the content of the present document, in particular the challenges experienced by Offices as indicated in paragraph 9;

(b) approve the proposal for conducting the survey via the joint Circular inviting the members of CWS and PCT, as indicated in paragraph 13 above;

(c) encourage its Members to participate in the survey and the

meeting, and note that the outcomes of the meeting will be presented at the fourteenth session of the CWS, as indicated in paragraph 15 above; and

(d) encourage its Members and Observers to participate in the workshop on WIPO Standard ST.36 and ST.96 to be held in the second quarter of 2026, as indicated in paragraph 16 above.

[End of document]