

产权组织标准委员会（标准委）

第十三届会议

2025 年 11 月 10 日至 14 日，日内瓦

知识产权文档的数字化

国际局编拟的文件

概 要

1. 产权组织标准委员会（标准委）成员曾表示其主要挑战之一是知识产权文档的数字化。与此同时，寻求将本国专利文档纳入《专利合作条约》（PCT）最低限度文献的专利主管机构，正在努力满足《PCT 行政规程》中所载的要求。本文件详细阐述了各主管局面临的部分挑战，并阐明制定一套数字化指南的必要性，初期将重点关注专利集。

背 景

2. 2024 年，秘书处面向标准委成员开展调查，以了解其在交换知识产权数据时可能面临的难题以及处理这些问题的潜在解决方案。调查结果显示，无论规模大小，各主管局面临的最大问题之一是知识产权公报信息未提供机器可读格式，以及支持这些活动的资源不足，包括人员技能差距和信息技术资源不足（见文件 CWS/12/23 第 8 段）。

3. 由欧洲专利局（欧专局）和美国专利商标局（美国专商局）牵头的 PCT 最低限度文献工作队（下称“工作队”），于 2005 年由 PCT 国际单位会议（MIA）设立，旨在对 PCT 最低限度文献进行全面审查。然而，相关工作直至 2017 年初才启动，当时制定的工作计划旨在实现以下四个目标（见文件 PCT/MIA/24/4 附录）：

- (a) 目标 A：针对目前 PCT 最低限度文献的专利文献和非专利文献部分，编订最新的详细目录；

- (b) 目标 B: 就国家专利集纳入 PCT 最低限度文献的条件和标准提出建议;
- (c) 目标 C: 明确规定应纳入属于 PCT 最低限度文献的专利集的专利数据著录项目和文本部分;
- (d) 目标 D: 就审查、纳入和维护非专利文献和基于传统知识的现有技术的条件和标准提出建议, 并在之后根据届时已建立的标准, 对印度主管部门关于印度传统知识数字库 (TKDL) 的经修改提案进行评估。

4. 在工作队和 PCT 工作组讨论后, 2023 年 7 月, PCT 联盟大会通过了对《PCT 实施细则》第 34 条、第 36 条和第 63 条的修正案, 自 2026 年 1 月 1 日起生效 (文件 PCT/A/55/2)。2024 年 6 月颁布了对《PCT 行政规程》的修改 (通函 C.PCT 1672), 包括新增附件 H。附件 H 详细规定了技术要求, 旨在确保所有属于 PCT 最低限度文献的专利集和实用新型文献集, 以机器可读格式免费向每个国际检索单位 (ISA) 提供, 且至少涵盖 2026 年 1 月 1 日及之后的公布, 至少追溯至 1991 年的公布应在接下来的十年内以全文格式提供。

5. 附件 H 规定文本可检索的数据有三种允许格式: 符合产权组织标准 ST. 36 的 XML 实例、符合产权组织标准 ST. 96 的 XML 实例, 或纯文本格式。各单位可选择其认为最适合的格式, 只要所选格式能支持对其专利集进行高质量检索即可。国际局推荐采用 XML 格式, 至少在提供著录项目数据时。选择将纯文本作为首选格式的主管局应考虑迁移至两项产权组织标准格式之一, 以提升专利公布质量和可检索性。

准备工作

6. 为协助各国际检索单位在 2026 年 1 月 1 日截止日期前满足技术要求, 国际局于 2025 年举办了一系列双边 PCT 最低限度文献诊断会。主管局事先准备了向国际局提出的问题, 国际局则确保有合适的专家出席诊断会解答这些问题。在此期间, 与会者指出《PCT 行政规程》未对全文文档的质量提供任何指导。应要求, 国际局向标准委权威文档工作队分享了其关于“足够优质”的机器可读文档的一套最低要求构想。

7. 对于正在将专利文档数字化的主管局, 以下内容为其符合附件 H 的要求提供了指导:

- 符合产权组织 ST. 36 的 XML 实例必须结构良好, 具有一致的根元素, 并且 OCR 错误减少到最低程度。著录项目数据至少应包含申请号、公布号和发明名称;
- 符合产权组织 ST. 96 的 XML 实例应遵循产权组织 ST. 96 文档级组件 Patent Publication, 但无需实现 MathML 或 OASIS Table;
- 纯文本文件应确保 OCR 错误减少到最低程度, 文本以规范段落形式呈现, 不含行号或页码, 且阅读顺序正确。最好同时配有 XML 格式的著录项目数据。

8. PCT 最低限度文献工作队同意采用三阶段认证流程, 面向那些寻求将其文献集纳入 PCT 最低限度文献的国际检索单位或其他专利局。第二阶段是“验证阶段”, 旨在测试一个主管局能否提供一个数据存储库, 其中的文件可供下载并符合附件 H 所列三种格式之一; 还测试主管局能否生成符合产权组织标准 ST. 37 的权威文档, 其中包括必要的文本检索标识符。

主管局面临的挑战

9. 在 PCT 最低限度文献诊断会及验证阶段期间，国际局提供了反馈意见，以协助开发机器可读专利公布并制作符合产权组织 ST. 37 的权威文档。在这一审查过程之后，主管局在制作机器可读文档时面临的共同挑战如下：

- (a) 所有格式均存在 OCR 质量低劣而导致大量错误的问题，因此污染检索索引，尤其是含化学公式的专利文档；
- (b) 所有格式均存在版面检测或处理不准确的问题，导致页眉页脚信息混入正文，或输出结果出现行号；
- (c) 所有格式均存在说明书、权利要求书和附图部分分割不当的问题，例如导致说明书文本被纳入权利要求书全文；
- (d) 所有格式均存在全文部分的语言标记错误，导致后续在检索系统中出现重大索引问题；
- (e) 对于符合产权组织 ST. 96 的实例，存在未经 PatentPublication 文档级组件验证的无效实例，且层级结构中语言代码的使用存在混淆；
- (f) 对于符合产权组织 ST. 36 的实例，对什么是“有效实例”以及什么是 XML 文档“足够优质”的最低标准存在误解。例如，是否必须使用 ST. 36 的 DTD 模型；
- (g) 对于纯文本格式，文本未按规范的段落呈现而是一行一行显示，导致内容难以检索，且著录项目数据采用非产权组织标准格式。

技术援助

10. 为应对上述部分挑战，国际局向寻求支持的主管局提供技术援助。产权组织 OCR 解决方案支持基于图像且文本可检索的书签化 PDF 文件制作符合产权组织 ST. 36 的 XML 实例。不过，标准委需注意，该工具不支持权利要求书编号检测和标记，并将表格和复杂的化学与数学公式作为嵌入式附图来处理。

11. WIPO Publish 使主管局能够以产权组织标准 ST. 36 格式提取数据并实施质量控制，其输出即将迁移至符合产权组织 ST. 96 的格式。这为众多知识产权局交换公报信息提供了便利。2025 年，国际局增加新功能，可为执行 PCT 最低限度文献要求的主管局提取专利文档。已在菲律宾知识产权局试点成功。专利集属于 PCT 最低限度文献的其他专利局可采用该同一方法；有意采用的主管局可联系国际局。

12. 国际局还向约 30 个主管局提供直接援助，支持其文献数字化工作，包括为部分主管局提供全文 OCR 服务。该流程包含文档扫描、数字化，然后对著录项目数据进行质检。在所有情况下，这项工作均外包给第三方承包商。

下一步工作建议

13. 注意到上述挑战，国际局拟提议制定一套知识产权文档数字化指南，以弥合上面指出的认知差距。为深入分析正在将其知识产权文档数字化的主管局所面临的挑战，国际局建议通过发出

标准委和 PCT 联合通函，开展问卷调查。国际局计划编拟调查问卷，并邀请两机构成员在 2026 年第一季度完成调查。

14. 虽然现阶段的讨论聚焦于专利集的数字化，但该指南的范围很可能扩展，纳入关于商标和工业品外观设计文档数字化的进一步指导。

15. 调查截止后，国际局计划于 2026 年上半年召开会议，讨论数字化指南的编拟事宜。届时将邀请标准委与 PCT 全体成员参与，不过，会议形式及日期等具体会议信息将适时宣布。在该会议上，国际局将报告调查结果以及制定知识产权文档数字化指南的可能时间表。鼓励标准委成员参与问卷调查及上述会议，因为需要其反馈意见以制定指南。国际局还计划向标准委第十四届会议汇报上述会议的成果。

16. 此外，国际局计划于 2026 年第二季度举办关于产权组织标准 ST. 36 和 ST. 96 的讲习班。该讲习班的邀请函将发送至标准委成员及观察员，活动信息将在产权组织网站上宣传。国际局同时邀请实施产权组织标准的主管局在需要时寻求技术援助。

17. 请标准委：

(a) 注意本文件的内容，特别是第 9 段所述各主管局面临的挑战；

(b) 批准上文第 13 段所述的、通过邀请标准委和 PCT 成员的联合通函开展调查的提案；

(c) 鼓励成员参与上文第 15 段所述的调查及会议，并请成员注意会议成果将提交标准委第十四届会议；并

(d) 鼓励成员和观察员参与上文第 16 段所述的、将于 2026 年第二季度举行的关于产权组织标准 ST. 36 和 ST. 96 的讲习班。

[文件完]