

Digitization Guidelines Survey

Response ID:224 Data

2. Identification page

1. Please enter the ST.3 code of the member state or inter-governmental organization you represent.

EA - Eurasian Patent Organization (EAPO)

2. Please enter the name of the Office or Organization you represent. For instance, International Bureau of WIPO.

If your organization name is the same as your ST.3 code name, you may put "n/a".

Eurasian Patent Organization

2. General questions

3. Is your national patent collection part of the PCT minimum documentation or does your Office contribute your IP documentation to one of WIPO's databases? Please select all that apply.

Patentscope

Global Design Database

PCT minimum documentation

4. What file or document format(s) does your Office accept at filing for IP applications? Please select all that apply.

For industrial design collections:

Image-based

Paper

For patent collections:

PDF

DOCX

Paper

For trademark collections:

Not applicable

5. In what language(s) do you publish your IP documents in? Please provide a full list of 2-letter language codes.

RU

6. In what format(s) do you publish or disseminate your IP collection? Please select all that apply.

For industrial design collections:

WIPO Standard ST.96

Image-based

3D formats

For patent collections:

WIPO Standard ST.36

WIPO Standard ST.96

PDF

For trademark collections:

Not applicable

7. In what format(s) does your Office publish the bibliographic data associated with the format(s) indicated in question 8? Please select all that apply.

For industrial design collections:

XML

Other (please specify): HTML (Publication Server at Web Site)

For patent collections:

XML

PDF

Other (please specify): HTML (Publication Server at Web Site)

For trademark collections:

Not applicable

8. If your Office produces XML or JSON instances and does not validate these instances before publication, please provide your reasoning below.

9. Does your Office publish the color drawings they receive, if they are permitted during the filing of a patent application?

Yes

3. Digitization activities

10. Does your Office perform digitization of documents you receive before publication or data dissemination, for one or more IP rights, or has experience of digitization of documents associated with IP rights?

Yes

10.1. How are or were these documents or associated data transformed from the format you received? e.g., paper document manually scanned and OCR'd or DOCX converted to XML.

Paper document manually scanned and OCR'd -> normalized DOCX (DOCX with correct formatting, having only technically allowed content and markup of parts, claims, main drawing etc.)

Image-only pdf manually OCR'd -> normalized DOCX

Text pdf is processed to extract the text layer (operator controlled) -> normalized DOCX

Docx is processed and normalized (operator controlled) -> normalized DOCX

Normalized DOCX is then converted to publication quality XML

10.2. If you rely or relied on a third party to perform the digitization of your documents, why did you reach this decision?

10.3. Does your Office use or did you use any IT tools, in particular AI tools, in the digitization process that you find particularly useful?

Currently Paddle OCR shows most promising results

10.4. What are or were the biggest technical challenges you have experienced in digitizing your documents? e.g., extracting drawings or chemical formulae from a patent application. Please provide further details in the comments box below, including how you overcame these challenges and the relevant IP right type.

Incorrect block markup (detection of images in text, including inline images)

Incorrect recognition of letters, digits, math and other special characters

Mixture of Latin and same looking Cyrillic letters

Incorrect word/sentence/paragraph breaks

Excessive merging of paragraphs

Unpredictable changes to document when using AI tools like PaddleOCR (hallucination effects): parts of text may be removed, some digits and characters (not present in the original document) may be added, replacement of Russian words with English ones.

4. Digitization activities

11. If your Office is considering future digitization activities, please provide details regarding the tools you are considering, in particular AI tools, and include your reasons for selecting them for potential use in digitization.

Automatic (AI-based) detection of the «complexity» and "type" of document

Depending on the complexity and type we plan to apply a proper tool or several tools (classic OCR, AI-OCR (now PaddleOCR is most promising), text layer extraction from PDF, math formulae detection using machine learning methods etc.)

We plan to continue testing all emerging AI OCR Tools as the technology is progressing rapidly

12. Please provide any other feedback regarding capacity building for the digitization of IP document collections.