

Digitization Guidelines Survey

Response ID:158 Data

1. Identification page

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

EP - European Patent Office (EPO)

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".

n/a

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

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:

Not applicable

For patent collections:

XML

PDF

Image-based

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.

EN, FR, DE

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

For industrial design collections:

Not applicable

For patent collections:

WIPO Standard ST.36

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 6? Please select all that apply.

For industrial design collections:

Not applicable

For patent collections:

XML

PDF

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.

For an OCR text quality of >99.5% (internal use)

- Specifications documents in paper are scanned and OCR'd internally
- Specifications documents in image or PDF are OCR'd internally
- Specification documents in DOCX are converted to ST36 internally
- Forms are OCR'd by an external contractor and converted to XML
- Citations in application documents and reports are extracted by an external contractor and converted to XML

For an OCR text quality of 99.995% (Publication)

- Specifications documents are OCR'd by an external contractor.

Images of pages/pages of PDFs are OCR'd to be structured into ST.36.

DocX are transformed into ST.36

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

It is difficult to reach high quality and the third party had the expertise. But we are getting there with AI. In some digitization steps, third parties are used. This is for flexibility reasons as well as to harness existing expertise.

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?

OCR:

- Abby Fine reader
- Mistral OCR

- DOCX:

Aspose

Yes, LLMs are starting to prove a viable option to other external or traditional digitization systems/services

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.

Applications documents: properly reflecting the original structure (paragraphs numbering, claims numbering, lists, ...) in the XML document is a challenge because many different formats/layouts of documents received.

We are still working on improving this.

For the extraction of citations, the non patent literature is more difficult to extract as generally not following any standard (contrary to patent literature that comprises clear patent numbers with country code etc).

In both case we are working on it in training an LLM.

For patent applications, extracting information from complex structures such as tables, mathematical formulas, chemical compounds, images embedded in text is the most challenging part. They can either be kept as image or, with more recent developments, structured and extracted as rich information.

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.

Further work with Mistral OCR are on going. Our goal is to reach publication quality (99.995%).

We also have activities on internalising the Form OCR and the Citation extraction.

Exploring the potential of additional AI solutions/models is on our roadmap of further developments. Those solutions will improve quality and timeliness during the patent grant process, also allowing for increased usage of fully digital data end-to-end.

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