

Digitization Guidelines Survey

Response ID:136 Data

1. Identification page

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

AU - Australia

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".

IP Australia

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 Brands 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:

PDF

Image-based

JSON

Paper

For patent collections:

XML

JSON

PDF

Image-based

DOCX

Paper

For trademark collections:

JSON

PDF

Image-based

Media-based formats (e.g., MP3)

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

EN

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

For industrial design collections:

Image-based

PDF

Other (please specify): API/JSON (Australian Design Search API)

For patent collections:

WIPO Standard ST.36

WIPO Standard ST.96

PDF

JSON (e.g., WIPO Standard ST.97)

For trademark collections:

Image-based
Media formats (e.g., MP3)
PDF
Other (please specify): MECA; API/JSON (Australian Trade Mark Search API)

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:

JSON
Other (please specify): HTML (web-based search system)

For patent collections:

XML
JSON
Plain text
Other (please specify): HTML (web-based search system)

For trademark collections:

XML
JSON
Other (please specify): HTML (web-based search system)

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

Testing was conducted on the process to build the XML for PCT Minimum Documentation requirements. Validation of every instance is not performed due to resource constraints.

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

No

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.

To align with PCT Minimum Documentation requirements, OCR'd PDFs are converted to XML.

Any paper documentation received is digitised by scanning to PDF. Not all documents are OCR'd, but anything forming part of a patent's specification is OCR'd.

All IP rights documentation stored as PDF; data is converted to formats indicated above as needed.

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

N/A

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?

We run custom image AI models to detect mathematical formulas and recognise chemical structures. We run them on the AWS SageMaker platform.

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.

Main challenges for PCT Minimum Documentation conversion of documents to XML came from document formatting such as:

- chemical formulas within sentences (not on a separate line)
- heavily condensed text
- identification of math compared to inline text
- Claim documents with paragraph markers in similar position to claim numbers
- Claim documents with multiple numbered lists within a claim
- Drawings that overlap (multiple to a page)
- Drawings with figure reference not in alignment with the image
- Descriptions with sequence lists included in the document
- Special formatting characters used by some jurisdictions
- Superfluous text provided with documents, e.g. cover pages

- Tables over hundreds of pages
- Tables with misaligned columns

These were identified by testing. Some but not all were overcome or improved by OCR coding enhancements.

These general document formatting issues and lack of standardisation diminish output quality, given all the possible edge case variations that cannot be accounted for in the build.

The most challenging conversions involved math and chem formula detection, as there was no initial guarantee that a feasible solution could be found. We overcame these challenges by training our own image recognition and classification AI models for the purpose.

Tables have also been a challenge. Textract handled them well initially, but performed worse after an update was done by AWS. We haven't found a solution for that and our table quality has decreased.

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.

N/A at this stage.

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

Key capabilities involved in the digitisation for the PCT Minimum Documentation work:

- Data science – AI/ML, image processing, NLP
- Software engineering – Java in our case
- Cloud infrastructure and DevOps – AWS in our case