GSMA and partners

Automated Extraction of 3GPP-citing patent claim mappings

End-of-Project Report

Date: 31 Mar 2022

1 Executive summary

This project includes a solution for performing the digitisation, and chunking of PDF documents into relevant pieces of information. Chunks can be filtered by regions with relevant mappings where mappings between patents and technical standards are extracted using a rule-based extraction logic. A dataset of 22,905 European Search Opinions documents was processed by our pipeline resulting in a total of 187,382 mappings. It was identified that 56,994 out of the total number of mappings refer to 3GPP citing documents. Additionally, the performance of the proposed solution was measured against manual annotations and showed a precision of over 80% for most of the extracted fields.

2 Goal

The goal of this project is to demonstrate the ability to scalably build a data set of mappings that can be used to build models for matching patents and technical standards. European search opinions (ESOs) contain mappings of patents to technical standards performed by domain experts. In this project, we will make use of ESOs documents to build the dataset of mappings. The work performed so far to extract relevant mappings has been performed manually, which is not feasible because it is a time-consuming and expensive process. Domain experts believe that the structure of most mappings follow a well defined pattern, thus it is believed that a rules-based approach performed on correctly digitised and filtered ESO documents would provide enough coverage of the to-date scraped ESO documents to form the foundation of a patent-standards matching dataset.

3 Technical Approach

The implemented technical approach is mainly composed of four different phases - Digitisation, Filtering, Extraction, Mappings (Fig. 1).

Fig.1: Conceptual pipeline for mappings extraction.

A total of 22,904 ESO documents were digitised and further processing included in the implemented technical approach was applied for the extraction of mappings.

Each phase is explained with more detail below.

3.1 Digitisation

Digitisation is the process of transforming the information from unstructured data that a computer can't process, into structured data that is in a format that computers can process.

Because ESOs are provided as unstructured data, digitisation approaches were considered as a starting point and the data was converted into a machine-readable format. For this digitisation process we selected Textract as our de facto tool.

3.1.1 Textract

Initial testing to OCR tools were performed using ESO PDF documents, and textract revealed the most impressive results. For this reason, textract was the chosen tool for the digitisation process.
Textract, is an AWS service that performs digitisation of documents and images. It uses a PDF file as input and outputs the text within the document and bounding boxes for the tokens within the text. Thus, besides the text information, Textract can also provide layout information. The layout information is important as it allows us to infer on the structure of the document such as lines, paragraphs and sections.

Fig.2. illustrates an example of the bounding boxes (identified regions of interest) captured by textract. It provides bounding boxes for individual tokens and also lines recognised within the document.

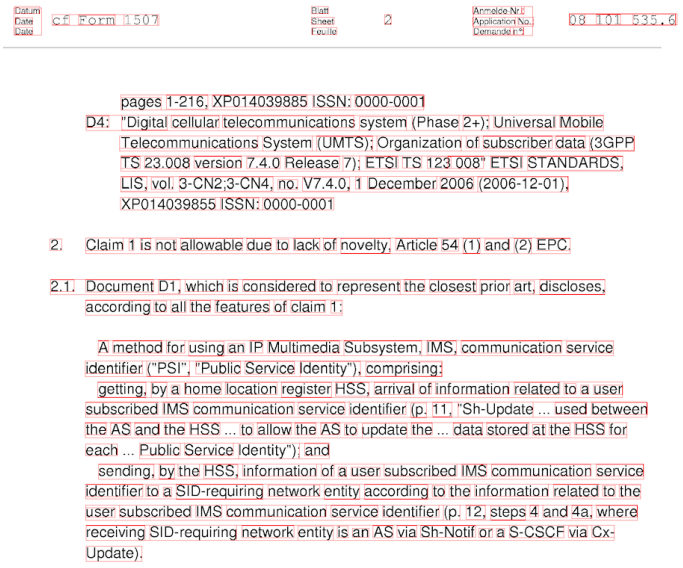


Fig.2: Textract bounding boxes output represented in red.

The initial dataset is composed of multiple PDF documents which required digitisation. Thus, AWS Textract was used and produced results structured as presented on Fig.3 , where:

1. ***Bucket***: The bucket where the file is located.
2. ***UploadedFileName***: The name of the original file.
3. ***DocumentMetadata***: Contains relevant information about the document (such as the number of pages).

4. ***JobStatus***: The status of the textract job.
5. ***Blocks***: Contains the digitised information extracted from the original document. Textract is able to identify the type of block (with tags such as Page, Word, Line, etc.), and relevant geometric information. The geometric information contains information about the bounding boxes used in the digitization process. It also includes information such as the randomly generated ID of the block, its relationship with other blocks and the actual text contained in the block.


```
{
  "Bucket": "textract-console-us-east-1-91aa8dcb-7224-49d4-ae92-0745a21ed041",
  "UploadedFileName": "19a13cf9_e826_45fb_8182_881a6c722a91_ep1933498_eogbz7ho2729fi4_european_search_opinion.pdf",
  "DocumentMetadata": {
    "Pages": 9
  },
  "JobStatus": "SUCCEEDED",
  "Blocks": [
    {
      "BlockType": "PAGE",
      "Geometry": {
        "BoundingBox": {
          "Width": 1,
          "Height": 1,
          "Left": 0,
          "Top": 0
        },
        "Polygon": [
          {
            "X": 1.73299076420408e-16,
            "Y": 0
          },
          {
            "X": 1,
            "Y": 8.654170694555159e-17
          },
          {
            "X": 1,
            "Y": 1
          },
          {
            "X": 0,
            "Y": 1
          }
        ]
      },
      "Id": "36303b9a-a6bf-4f1a-b87a-1e8489f7dd53",
      "Relationships": [
        {
          "Type": "CHILD",
          "Ids": [
            "360db52a-cd81-4086-86fd-2be68855aa4f",
            "ae12aadf-7760-4ee9-b9f6-478e91472eec",
            "2d7b20d9-de7a-430f-b7af-a577ceee3317",
          ]
        }
      ],
      "Page": 1,
      "childText": "Datum Blatt Anmelde-Nr. Date cf Form 1507 Sheet 1 Application No.: 06 790 947.3 Date Feuille Dema
      "SearchKey": "Datum Blatt Anmelde-Nr. Date cf Form 1507 Sheet 1 Application No.: 06 790 947.3 Date Feuille Dema
    },
```


Fig.3: Textract output information.

3.2 Filtering

The filtering phase includes loading the data (Data Loader), creating the representation of the document and extracting statistics (Document Feeder), chunking the data into meaningful chunks of information (Chunker) and building rules for filtering the relevant documents and chunked sections for further processing (Filtering Rules).

3.2.1 Data Loader

The data loader is responsible for providing extensible methods for loading data from multiple sources. In this work, data can be made available by loading it from a local filesystem or from a remote S3 bucket.

3.2.2 Document Feeder

The document feeder defines the representation of the document by translating the textract output into a data dictionary that groups data by block type. This document feeder can also compute document statistics. Lastly, the logic for extracting the header and footer information is defined.

3.2.2.1 Document Statistics

Before building the logic for the line, section and paragraph chunkers, document statistics were extracted. These include:

1. ***Height average of tokens***: this is used to check if the tokens share similar Y coordinates (where Y is the distance from the top border of the document).
2. ***Max left indent:*** this is used to check whether a token (smaller form of group of characters) is aligned to the far left of the document. It is important to determine whether it is the initial token of a section.
3. **Token spacing**: the spacing between tokens is used to check whether a new paragraph should be started or not.
4. **Top boundary footer:** the last bin (group) of the histogram for vertical differences between tokens. As it is the group with further distances to the top, it will correspond to the top boundary of the footer (see Fig.4).

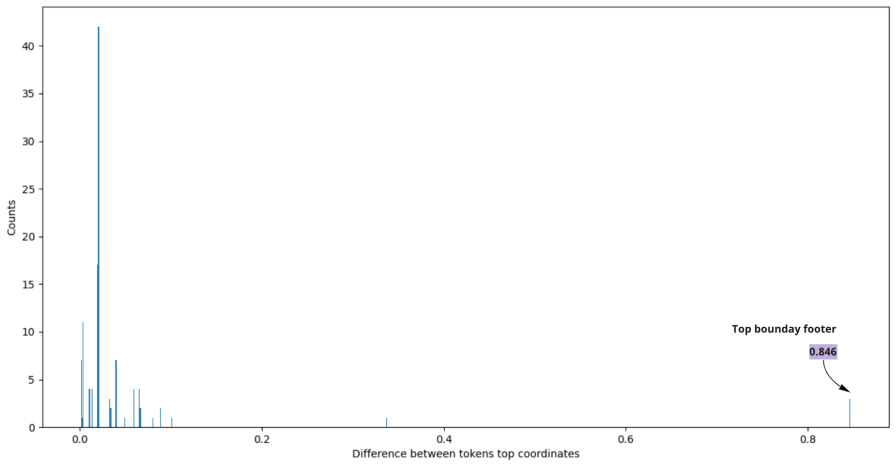


Fig. 4: Histogram of the difference between tokens' top coordinates. It is used for finding the top boundary footer. In this case the value is 0.846.

3.2.2.2 Header and Footer Extraction

The header and footer information is not relevant and can cause noise in our results. For this reason, the logic for extracting the header and footer from our documents is described below.

After analysing some documents, we concluded that the best approach for removing the header was to search for the token "*Demande nº:*" and extract all tokens before that. This logic is applied to all pages, and the header successfully removed from each document sample. As for the footer, we took advantage of the extracted statistics for removing it. The top boundary footer matches the top of the bounding box of the footer, so every token below that value is extracted as the footer.

3.2.3 Chunker

Chunking the document into meaningful pieces of text was one of the most important deliverables of this project. The textract output does not provide information about groups of words or the meaning behind them (such as paragraphs or sections). For the development of this project this information is fundamental as rules shall only be applied to target sections/paragraphs of the document, avoiding a high number of false positive results. Hence, it is important to chunk the document into lines, paragraphs and sections so that we can more easily navigate into the document, and filter the relevant portions of text.

By taking advantage of the layout information provided in the textract output, a chunker class was implemented that receives as input the data collated from the dataloader, parses it and chunks the input document into meaningful portions. The current version of the chunker is able to chunk documents into lines, paragraphs and sections. Additionally, these chunks can be combined to form a defined structure.

The logic built for the section chunker uses the information of the vertical spacing between tokens, left indent, and the section numbers (e.g. *1*, *1.1*, *2.* etc) for defining a section.

Fig.5. illustrates the bounding boxes automatically generated by the section chunker. In this example, it is possible to visualise that both sections and subsections are being correctly identified.

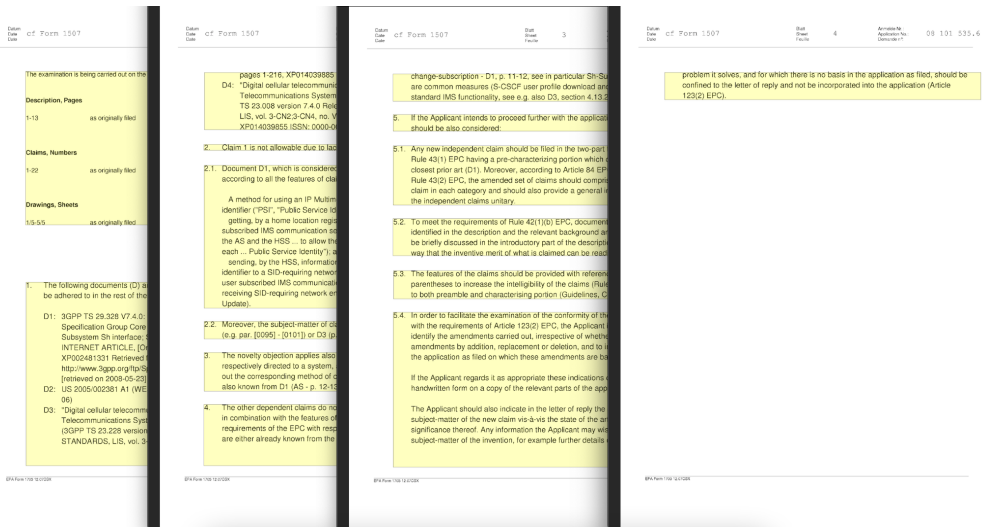


Fig.5: Section chunker

The paragraph chunker logic uses the information of the vertical spacing between tokens for defining a paragraph.

Fig.6. illustrates the bounding boxes automatically generated by the paragraph chunker.

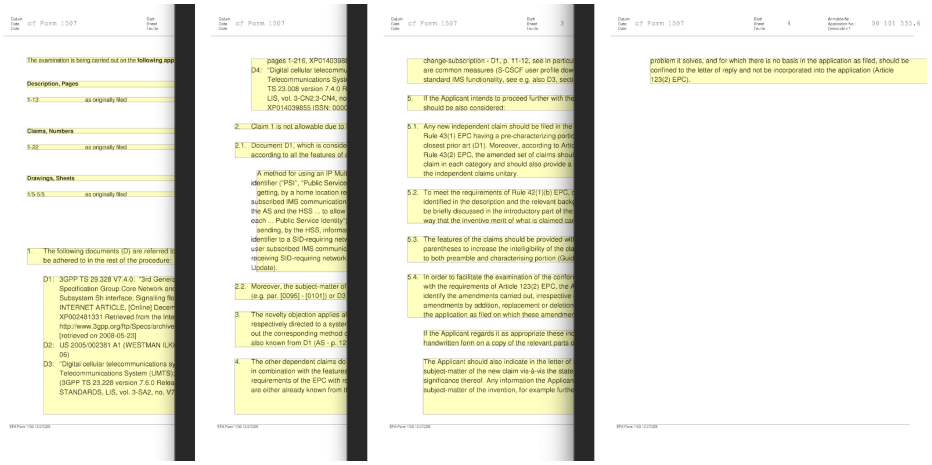


Fig.6 : Paragraph chunker

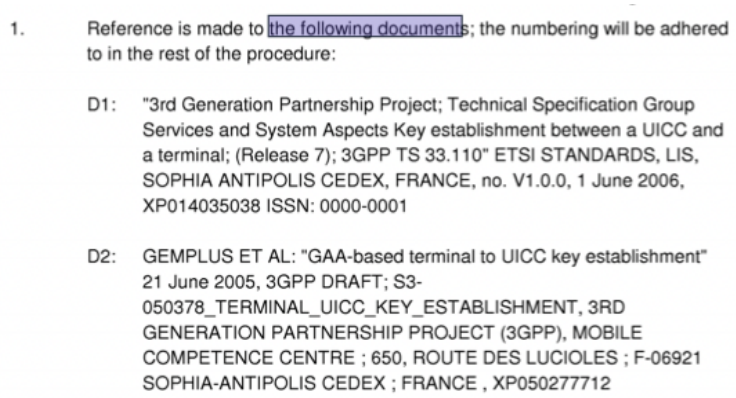
3.2.4 Filtering Rules

Three filtering rules were implemented. The first filtering rule was implemented to filter the document description section. The remaining rules filter out non-relevant documents and operate on the level of filtering non-3GPP citing documents and/or documents which contain no mappings. These rules can be chained together or run independently.

3.2.4.1 Document Description Filter

The document description filter filters the section containing the document descriptions. It is important to filter this section so that we can further extract the document numbers and corresponding text.

This filter works by searching for the "*the following document*" string. The result of the filter will be the section that matches the filtering rule (see Fig.7).



1. Reference is made to the following documents; the numbering will be adhered to in the rest of the procedure:

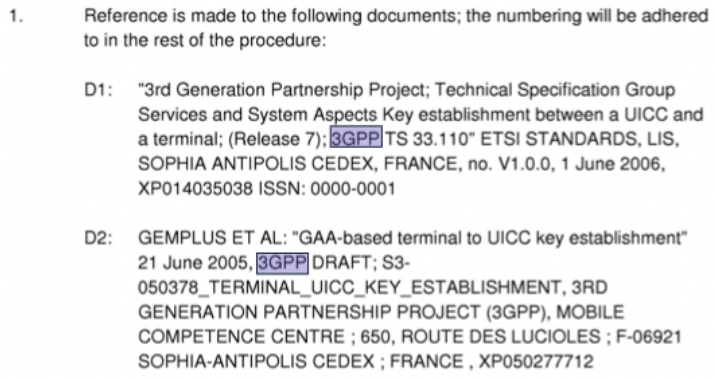
D1: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects Key establishment between a UICC and a terminal; (Release 7); 3GPP TS 33.110" ETSI STANDARDS, LIS, SOPHIA ANTIPOLIS CEDEX, FRANCE, no. V1.0.0, 1 June 2006, XP014035038 ISSN: 0000-0001

D2: GEMPLUS ET AL: "GAA-based terminal to UICC key establishment" 21 June 2005, 3GPP DRAFT; S3-050378_TERMINAL_UICC_KEY_ESTABLISHMENT, 3RD GENERATION PARTNERSHIP PROJECT (3GPP), MOBILE COMPETENCE CENTRE ; 650, ROUTE DES LUCIOLES ; F-06921 SOPHIA-ANTIPOLIS CEDEX ; FRANCE , XP050277712

Fig. 7: Document Description section filter

3.2.4.2 3GPP citing filter

The 3GPP citing filter filters documents that are 3GPP citing. This filter is applied to the filtered document description section. The filter works by searching for the "3GPP" string (see Fig. 8). It will output a flag in the output results that checks if the document is or not 3GPP citing.



1. Reference is made to the following documents; the numbering will be adhered to in the rest of the procedure:

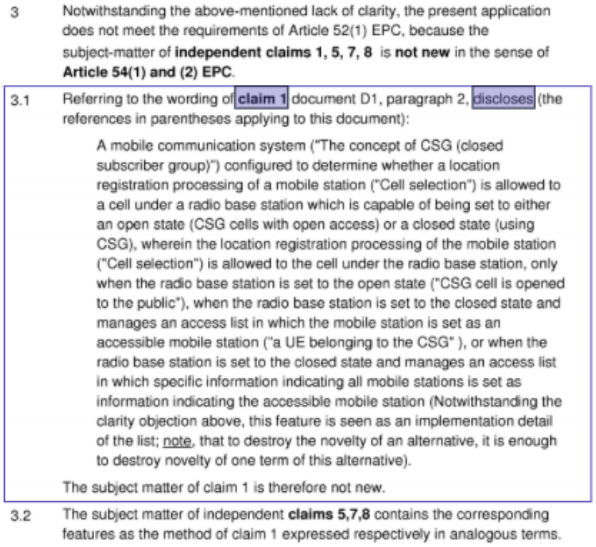
D1: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects Key establishment between a UICC and a terminal; (Release 7); 3GPP TS 33.110" ETSI STANDARDS, LIS, SOPHIA ANTIPOLIS CEDEX, FRANCE, no. V1.0.0, 1 June 2006, XP014035038 ISSN: 0000-0001

D2: GEMPLUS ET AL: "GAA-based terminal to UICC key establishment" 21 June 2005, 3GPP DRAFT; S3-050378_TERMINAL_UICC_KEY_ESTABLISHMENT, 3RD GENERATION PARTNERSHIP PROJECT (3GPP), MOBILE COMPETENCE CENTRE ; 650, ROUTE DES LUCIOLES ; F-06921 SOPHIA-ANTIPOLIS CEDEX ; FRANCE , XP050277712

Fig. 8: 3GPP citing filter

3.2.4.3 Claim X discloses filter

The claim X discloses filter searches for the sections containing both words “claim X[1]” and one of the words - “disclose” or “disclosure”. According to domain experts, when the attorney starts mapping a claim to a standard it starts by using these two words. This way, we can filter out the sections that match this filtering rule for further processing on the extraction of the mappings, or to filter out non-relevant documents as it doesn’t contain mappings. Fig.9 contains an example of a section containing mappings, and it matches the filtering rule defined.



3 Notwithstanding the above-mentioned lack of clarity, the present application does not meet the requirements of Article 52(1) EPC, because the subject-matter of **independent claims 1, 5, 7, 8** is **not new** in the sense of **Article 54(1) and (2) EPC**.

3.1 Referring to the wording of **claim 1** document D1, paragraph 2, discloses (the references in parentheses applying to this document):

A mobile communication system ("The concept of CSG (closed subscriber group)") configured to determine whether a location registration processing of a mobile station ("Cell selection") is allowed to a cell under a radio base station which is capable of being set to either an open state (CSG cells with open access) or a closed state (using CSG), wherein the location registration processing of the mobile station ("Cell selection") is allowed to the cell under the radio base station, only when the radio base station is set to the open state ("CSG cell is opened to the public"), when the radio base station is set to the closed state and manages an access list in which the mobile station is set as an accessible mobile station ("a UE belonging to the CSG"), or when the radio base station is set to the closed state and manages an access list in which specific information indicating all mobile stations is set as information indicating the accessible mobile station (Notwithstanding the clarity objection above, this feature is seen as an implementation detail of the list; note, that to destroy the novelty of an alternative, it is enough to destroy novelty of one term of this alternative).

The subject matter of claim 1 is therefore not new.

3.2 The subject matter of independent **claims 5,7,8** contains the corresponding features as the method of claim 1 expressed respectively in analogous terms.

Fig. 9: Claim X discloses filter

Different variations of the word discloses were identified as the most relevant trigger for detecting sections with mappings. However, different variations of the word “teaches” can also be used to detect relevant mappings. The implemented filter was adapted so that the word “teaches” and “teaching” are also considered.

3.3 Extraction

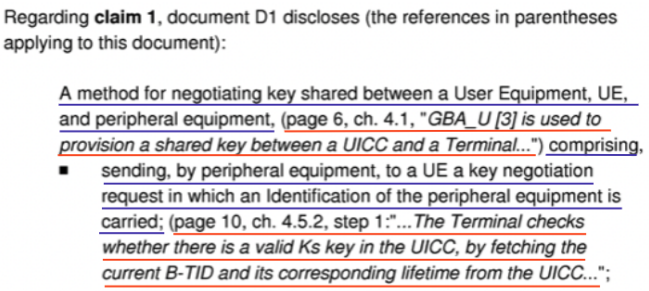
After chunking the document into its relevant portions, and filtering the most meaningful sections, we are now in position to extract information from the document. Mapping rules were built to extract the mappings between patent claims and standard documents. Additional information, that might be relevant for a follow-up project, was also extracted from specific paragraphs, results, metadata and standard description.

3.3.1 Mapping Rules

In order to capture the relevant mappings from the documents a single rule was implemented. Given an input sentence, this rule extracts the information before and in between parentheses. This mapping rule is applied to the paragraphs within the section filtered by the Claim X discloses filtering rule. Fig. 10 illustrates an example where this mapping rule is correctly applied. The text before parentheses corresponds to the claim text

[1] X can be any digit.

and the text within parentheses is the document reference text (i.e. mapping of the standard document).


Regarding **claim 1**, document D1 discloses (the references in parentheses applying to this document):

A method for negotiating key shared between a User Equipment, UE, and peripheral equipment, (page 6, ch. 4.1, "*GBA_U [3] is used to provision a shared key between a UICC and a Terminal...*") comprising,

- sending, by peripheral equipment, to a UE a key negotiation request in which an Identification of the peripheral equipment is carried; (page 10, ch. 4.5.2, step 1:"*...The Terminal checks whether there is a valid Ks key in the UICC, by fetching the current B-TID and its corresponding lifetime from the UICC...*";

Fig. 10: In purple is represented the claim text followed by the corresponding document reference mapping in red.

3.3.2 Paragraph Extractor

The paragraph extractor was built to extract the document and claim number from paragraphs within the mappings section.

3.3.2.1 Document Number Extractor

Extracts the document number from the extracted paragraph. The extractor supports the following pattern:

1. "Document D1 (see in particular citations …" → extracts "D1"

3.3.2.2 Claim Number Extractor

Extracts the claim number from the extracted paragraph. The extractor supports the following patterns:

1. "discloses according to all the features of claim 1, a method to" → extracts "1"
2. "discloses from claim 1-9" → extracts "1-9"
3. "discloses from claim 1 to 9" → extracts "1;9"
4. "discloses according to features of claim 1 and 2"→ extracts "1;2"
5. "referring to claims 4-6 and 8-10 D1 discloses"→ extracts "4-6;8-10"
6. " discloses according to features of claims 1, 8, 9"→ extracts "1, 8, 9"

3.3.3 Result Extractor

The result extractor was designed to extract additional information from our results. In this case, the quoted text from within the resulting document reference text. The document reference text is the result of applying the mapping rules previously described.

3.3.3.1 Standard Quotation Transformer

Extracts the standard quotation text from the document reference text (i.e. standard text). After the analysis of some documents, we concluded that a single pattern would not be enough to correctly extract the standard quotation text. We identified some edge cases which can be covered by the following rules:

1. text within quotation marks or single quote
2. start with **:** and ending with **;**
3. start with **“** or **‘** and ending with **;**
4. start with **‘** and ending with **”** (reverse also applies)
5. start with **:** and ends with **”** or **’**
6. all before **“** or **‘**
7. all before **:**

The previous patterns are applied sequentially, so if the one pattern matches the following patterns will not be applied.

3.3.4 Metadata Extractor

The metadata extractor was built to extract metadata from the document description section, such as the standard text, version number, category, release, publication date, and whether it is 3GPP citing or not.

3.3.4.1 Standard Text Extractor

Extracts the standard document text. It is assumed that the text referring to the standard document starts, in the same line, and after mentioning the document number e.g. D1. It ends before the next mention of a document number or until the end of the section if no more documents are followed. A concrete example is illustrated in Fig. 11.

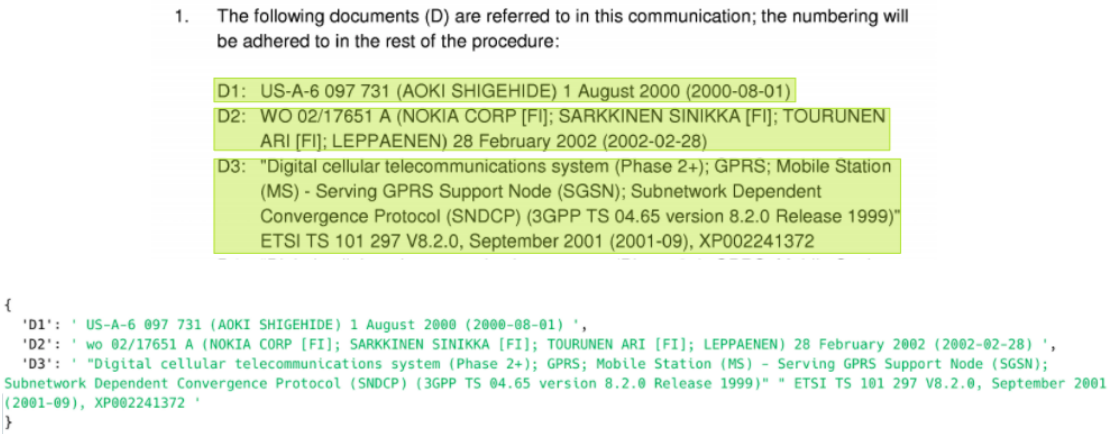


Fig. 11: Top figure is an excerpt of the document description section. The bottom figure is the output of the standard text extractor.

3.3.4.2 Standard Version Extractor

Extracts the standard version from the standard document text. It currently supports the following patterns:

1. V1.0.0
2. v 1.0.0
3. version 1.0.0
4. Version 1.0.0

3.3.4.3 Document Category Extractor

Extracts the document category from the standard document text. It covers the following patterns:

1. TS (e.g. TS 33.110)
2. CR (e.g. CR 33.110)
3. TR (e.g. TR 33.821)
4. Tdoc (e.g. Tdoc SA-WG3 SECURITY)
5. TSG (e.g. TSG SA WG3 SECURITY)

3.3.4.4 Document D 3GPP Extractor

Extracts the 3GPP flag and returns whether a specific document is 3GPP citing or not. In the example below, the document is flagged as 3GPP citing.

Example: *"**3GPP** TS 29.328 V7.4.0: "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; IP Multimedia (IM) Subsystem Sh interface; Signalling flows and message contents (Release 7)' 3GPP TS 29.328 V7.4.0: "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; IP Multimedia (IM) Subsystem Sh interface; Signalling flows and message contents (Release 7)'""*

3.3.4.5 Release Extractor

Extracts the release number from the standard document text. It supports the following patterns:

1. **Release** followed by a letter or digits (e.g. Release 1)
2. **RELEASE** followed by a letter or digits (e.g. RELEASE 1)

3.3.4.6 Publication Date

Extracts the publication date from the standard document text. Currently this extraction supports the following patterns:

1. YYYY-MM-DD

2. DD/MM/YYYY
3. DD-MM-YYYY
4. DD month (e.g. January or Jan) YYYY

3.3.5 Standard Extractor

The standard extractor was built to extract the document references from the mapping of standard documents.

3.3.5.1 Document References Extractor

The document references extractor extracts the passage type and corresponding reference (Fig. 12). The passage types supported by this extractors are included in the following dictionary:

```
{
  "page": "Page",
  "Page": "Page",
  "section": "Section",
  "Section": "Section",
  "ch\.": "Chapter",
  "Ch\.": "Chapter",
  "paragraph": "Paragraph",
  "Paragraph": "Paragraph",
  "par\.": "Paragraph",
  "Par\.": "Paragraph",
  "Figure": "Figure",
  "figure": "Figure",
  "fig\.": "Figure",
  "Fig\.": "Figure",
  "step": "Step",
  "Step": "Step",
  "p\.": "Page",
  "P\.": "Page",
  "column": "Column",
  "Column": "Column",
  "col\.": "Column",
  "line": "Line",
  "item": "Item",
  "formula": "Formula"
}
```

Fig. 12: Included keywords in the document reference extractor

The extracted types are the values in the dictionary, and these are normalised to the corresponding key in the output results.

Some examples of extracted document passages supported:

1. page 6, ch. 4.1 → extracts “Page 6” and “Chapter 4.1”
2. In Page 7.1 and something else → extracts “Page 7.1”
3. Ch.3 and 7 → extracts “Chapter 3” and “Chapter 7”
4. section 2.1 and section A.3 → extracts “Section 2.1” and “Section A.3”
5. section 4, 5, 6-9 → extracts “Section 4”, “Section 5” and “Section 6-9”
6. figure A-5-1 and A-5-2 → extracts “Figure A-5-1” and “Figure A-5-2”
7. From page 6 to 10 there are mappings → extracts “Page 6 to 10”

3.4 Mappings

The mappings phase comprises the consolidation and storage of the results into the expected output format.

3.4.1 Result Writer

The result writer collates all information produced by the previous stages and combines it into a single csv file such as the one presented in Fig. 13.

patent_no	feature_number	feature_text	document_passage_text	document_reference_text	document_passage_type	document_passage_extracted	quoted_text	d_number	version	standard_text	parsed_standard_and_version	three_gpp_citing	release	publication_date
EP1725065	claim 1	a method	"roaming", paragraph 2.4	" ", paragraph 2.4	Paragraph	2.4	roaming	D1	version 6.5	"Organisation o	TS 23.008	yes	Release 6	
EP1725065	claim 1	receiving a re	"location updating", paragr	" ", paragraph 2.4.10	Paragraph	2.4.10	location upda	D1	version 6.5	"Organisation o	TS 23.008	yes	Release 6	
EP1725065	claim 1	if the user is d	"in the course of roaming c	" ", paragraph 2.4.15	Paragraph	2.4.15	in the course	D1	version 6.5	"Organisation o	TS 23.008	yes	Release 6	
EP1725065	claim 1	determining tc	"setting network induced n	" ", paragraph 2.4.15	Paragraph	2.4.15	setting netwo	D1	version 6.5	"Organisation o	TS 23.008	yes	Release 6	
EP1725065	claim 1	access to rega	"sending network induced "	" ", paragraph 2.4.15	Paragraph	2.4.15	sending netw	D1	version 6.5	"Organisation o	TS 23.008	yes	Release 6	
EP1933498	claim 1	A method for n	page 6, ch. 4.1, 'GBA_U[3]	page 6, ch. 4.1, ' "	Page	6	GBA_U[3] is u	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	A method for n	page 6, ch. 4.1, 'GBA_U[3]	page 6, ch. 4.1, ' "	Chapter	4.1	GBA_U[3] is u	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	figure 4.3) calc	page 10, ch. 4.5.2, step 5:"	page 10, ch. 4.5.2, step 5:"	Page	10	The UICC ret	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	figure 4.3) calc	page 10, ch. 4.5.2, step 5:"	page 10, ch. 4.5.2, step 5:"	Chapter	4.5.2	The UICC ret	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	figure 4.3) calc	page 10, ch. 4.5.2, step 5:"	page 10, ch. 4.5.2, step 5:"	Figure	4.3	The UICC ret	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	figure 4.3) calc	page 10, ch. 4.5.2, step 5:"	page 10, ch. 4.5.2, step 5:"	Step	5	The UICC ret	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	receiving, by t	page 11, ch. 4.5.2, step 11	page 11, ch. 4.5.2, step 11:	Page	11	The NAF Key	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	receiving, by t	page 11, ch. 4.5.2, step 11	page 11, ch. 4.5.2, step 11:	Chapter	4.5.2	The NAF Key	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	receiving, by t	page 11, ch. 4.5.2, step 11	page 11, ch. 4.5.2, step 11:	Step	11	The NAF Key	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	which is calcul	page 11, ch. 4.5.2, steps 8	page 11, ch. 4.5.2, steps 8-	Page	11	8- The NAF K	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	which is calcul	page 11, ch. 4.5.2, steps 8	page 11, ch. 4.5.2, steps 8-	Chapter	4.5.2	8- The NAF K	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06
EP1933498	claim 1	which is calcul	page 11, ch. 4.5.2, steps 8	page 11, ch. 4.5.2, steps 8-	Step	08-Sep	8- The NAF K	D1	V1.0.0	"3rd Generatior	TS 33.110	yes	Release 7	01-Jun-06

Fig.13: Example of the output results in a CSV file.

3.5 Pipeline

The process of extracting mappings and metadata from a single document is started by loading its digitised representation from the local file system or an S3 bucket. Such behaviour is conducted by a data loader which is a building block of the document feeder. The document feeder computes relevant document statistics, removes its header and footer, and separates the digitised blocks by type. The chunker receives blocks of the 'WORD' type and starts the chunking process. Every time a line, paragraph or section is identified an event is flagged and the aggregator collates the event data. On completion the chunked document is passed to the next stage: the section filter. During this stage, sections are filtered based on their relevance to this work. Sections containing mappings are passed on to the rule runner while the section containing the document metadata is passed to the metadata extractor. The rule runner iterates over each paragraph and extracts mappings. The metadata extractor, on the other hand, executes a sequence of instructions which will extract information such as the standard version, document category, among others. The mappings and the metadata are then combined in the result writer to produce the final output in the form of a csv file. This process is summarised in Fig. 14.

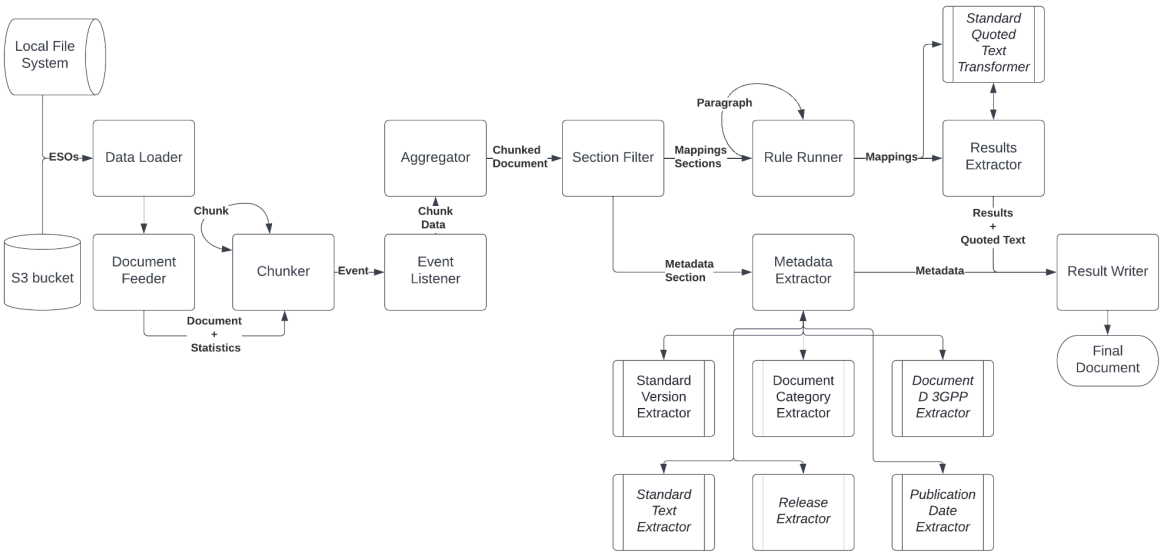


Fig. 14: Full document pipeline

4 Results

The results section includes the report of coverage and statistics extracted from running the implemented approach over the full dataset, performance metrics extracted based on the annotations provided by Septigent, and limitations of our current approach.

4.1 Coverage and Statistics

From 22,903 analysed documents we obtained a total of 187,383 mappings. These are all mappings extracted from the document regardless of them being 3GPP citing or not. It should be highlighted that the functionality of filtering non-3GPP citing documents is in the code and can be triggered at any time. Table 1 contains relevant statistics regarding the coverage of the developed extraction tool for chunking and filtering documents, extracting mappings and metadata sections.

Table 1: Summary of the relevant statistics of the developed extraction tool.

	Raw Number	**% of total**
Total Number of documents	**22,903**	100.0%
Number of chunked documents	22,803	99.6%
Number of filtered documents	20,831	91.0%
Number of docs with mappings	20,628	90.1%
Number of documents with metadata section extracted	19,274	84.2%

Table 2 includes the quoted text statistics, i.e. from the total number of quoted text extracted, the number and percentage of the extracted quoted text that did contain balanced quotation marks.

Table 2: Mappings - quoted text statistics

	Raw Number	**% of total**
Quoted text	**78,087**	100.0%
Mappings with balanced Quotation Marks	50,151	64%

Table 3 includes the total number and percentage of fields extracted for the 22,903 documents when the 3GPP filtering criteria is dismissed.

Table 3: Number and percentage of the extracted fields.

Extracted field	**Raw Number**	**% of total mappings**
patent number	**187,382**	100.0%
claim number	172,426	92.0%
feature text	187,382	97.8%
document passage text	187,382	100.0%
document reference text	187,382	100.0%
document passage type	187,382	100.0%
document passage extracted	187,382	100.0%
quoted text	78,087	41.7%
document number	165,312	88.2%
version	17,926	9.6%
standard text	154,586	82.5%
parsed standard and version (category)	18,101	9.6%
release	15,826	8.4%
publication date	137,545	73.4%

Table 4 compiles the same statistics for 3GPP citing documents, resulting in 56,994. Overall, the final solution for the filtered documents seem to present better extraction capabilities mainly on the fields related to the document metadata section. The version number is the field where the improvement is more noticeable whilst remaining in a fairly low precision score. As previously discussed this can be due to a multitude of reasons, including the fact

that a single version number is extracted per prior art document. In some situations, multiple versions can be found (sometimes relevant for other types of documents), which can be affecting the extraction rule.

Table 4: Number and percentage of the extracted fields for 3GPP documents

Extracted field	**Raw Number**	**% of total mappings**
patent number	**56,994**	100.0%
claim number	52,158	91.6%
feature text	55,553	97.6%
document passage text	56,994	100.0%
document reference text	56,994	100.0%
document passage type	56,994	100.0%
document passage extracted	56,994	100.0%
quoted text	28,392	49.9%
document number	56,994	100%
version	15,972	28.05%
standard text	56,994	100%
parsed standard and version (category)	17,751	31.2%
release	15,499	27.2%
publication date	50,667	89.0%

4.2 Performance

Two different experiments were conducted to measure the performance of our solution (*Manual Annotations vs DI Results*) against the performance of an experiment conducted by GSMA (*NLPClaimMaps vs DI Results*). Since the only source of truth that we currently have are the manual annotations, we will be comparing the results from manual annotations with those extracted by our solution and GSMA's.

The levenshtein distance (or edit distance) was used to measure the similarity between the annotated and extracted fields (Fig. 15). The levenshtein distance measures the number of edits we need to apply to string a so that it converts to string b. This distance was transformed into a similarity measure by normalising it by the maximum length between

both strings and subtracting it from 1. This way, the resulting similarity measure varies between 0 and 1, where 0 means that the strings are not similar and 1 they are similar.

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}\left(\operatorname{tail}(a), \operatorname{tail}(b)\right) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}\left(\operatorname{tail}(a), b\right) \\ \operatorname{lev}\left(a, \operatorname{tail}(b)\right) \\ \operatorname{lev}\left(\operatorname{tail}(a), \operatorname{tail}(b)\right) \end{cases} & \text{otherwise,} \end{cases}$$

Fig. 15: Levenshtein distance formula.

After measuring the similarity between the extracted fields, a threshold is defined for classifying the mapping as true positive (TP) or a false positive (FP) result. The precision metric is then used to evaluate the performance of each field extracted.

The precision measures the ratio between true positives (TP) and all the positives (i.e. TP + FP). The precision values range between 0-100. A precision of 0 means that none of the extracted results matched the annotations (*bad precision*), whereas a precision of 100 means that all extracted results match the annotations (*good precision*).

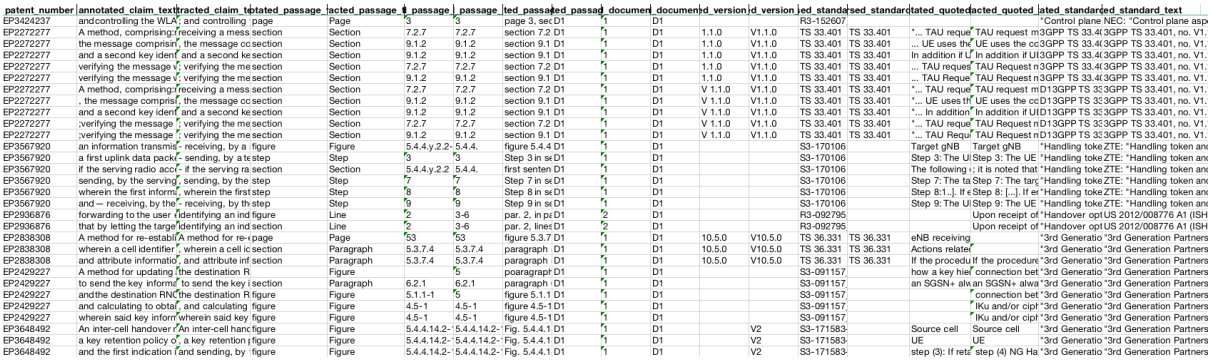
4.2.1 Manual Annotations vs DI Results

The performance of the implemented solution was measured based on the comparison of the results obtained from our solution against manual annotations provided by Septigent. The fields used for this comparison are the claim text, document passage type, document passage extracted, document number, version number, document category, and quoted text.

Although we extracted more fields, we can only extract performance metrics from those that we have a ground truth to compare with.

From 50 annotated documents, we were able to generate results from 34 documents. The main reason why we didn't get results for all 50 documents is related to limitations attached to the claim number extraction from the mappings paragraphs. These limitations are explored with more detail in the Limitations section.

Fig. 16 illustrates the results of the annotated and extracted fields with the highest similarity. The first column is the patent number, and the remaining columns follow the same structure as the provided annotations. This allows us to have a visual perception of the performance of the results compared with annotations.



patent_number	annotated_claim_text	tracted_claim_te	otated_passage	acted_passage_t	passage	passage	ted_passage	ted_passag	document	document_version	d_version	ed_standard	sed_standard	tated_quoted	acted_quoted	ated_standard	ed_standard_text
EP3424237	and controlling the WLA	and controlling	page	Page	3	3	page 3, sec	D1	1	D1			R3-152607			"Control plane	NEC: "Control plane asp
EP2272277	A method, comprising	receiving a mess	section	Section	7.2.7	7.2.7	section 7.2	D1	1	D1	1.1.0	V1.1.0	TS 33.401	TS 33.401	"... TAU reque	TAU request m	3GPP TS 33.4(	3GPP TS 33.401, no. V1.
EP2272277	the message comprises	the message cc	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	1.1.0	V1.1.0	TS 33.401	TS 33.401	... UE uses the	UE uses the cc	3GPP TS 33.4(	3GPP TS 33.401, no. V1.
EP2272277	and a second key ident	and a second ke	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	1.1.0	V1.1.0	TS 33.401	TS 33.401	In addition if U	In addition if UI	3GPP TS 33.4(	3GPP TS 33.401, no. V1.
EP2272277	verifying the message	verifying the me	section	Section	7.2.7	7.2.7	section 7.2	D1	1	D1	1.1.0	V1.1.0	TS 33.401	TS 33.401	... TAU reques	TAU Request n	3GPP TS 33.4(	3GPP TS 33.401, no. V1.
EP2272277	verifying the message	verifying the me	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	1.1.0	V1.1.0	TS 33.401	TS 33.401	... TAU Reque	TAU Request n	3GPP TS 33.4(	3GPP TS 33.401, no. V1.
EP2272277	A method, comprising	receiving a mess	section	Section	7.2.7	7.2.7	section 7.2	D1	1	D1	V 1.1.0	V1.1.0	TS 33.401	TS 33.401	"... TAU reque	TAU request m	D1 3GPP TS 33	3GPP TS 33.401, no. V1.
EP2272277	, the message comprise	the message cc	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	V 1.1.0	V1.1.0	TS 33.401	TS 33.401	"... UE uses th	UE uses the cc	D1 3GPP TS 33	3GPP TS 33.401, no. V1.
EP2272277	and a second key ident	and a second ke	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	V 1.1.0	V1.1.0	TS 33.401	TS 33.401	"... In addition	In addition if UI	D1 3GPP TS 33	3GPP TS 33.401, no. V1.
EP2272277	;verifying the message	verifying the me	section	Section	7.2.7	7.2.7	section 7.2	D1	1	D1	V 1.1.0	V1.1.0	TS 33.401	TS 33.401	"... TAU reque	TAU Request n	D1 3GPP TS 33	3GPP TS 33.401, no. V1.
EP2272277	;verifying the message	verifying the me	section	Section	9.1.2	9.1.2	section 9.1	D1	1	D1	V 1.1.0	V1.1.0	TS 33.401	TS 33.401	"... TAU Requ	TAU Request n	D1 3GPP TS 33	3GPP TS 33.401, no. V1.
EP3567920	an information transmis	receiving, by a	figure	Figure	5.4.4.y.2.2-	5.4.4.	figure 5.4.4	D1	1	D1			S3-170106		Target gNB	Target gNB		"Handling toke	ZTE: "Handling token an
EP3567920	a first uplink data packe	sending, by a te	step	Step	3	3	Step 3 in sc	D1	1	D1			S3-170106		Step 3: The UI	Step 3: The UE		"Handling toke	ZTE: "Handling token an
EP3567920	if the serving radio acc	if the serving ra	section	Section	5.4.4.y.2.2	5.4.4.	first senten	D1	1	D1			S3-170106		The following	it is noted that		"Handling toke	ZTE: "Handling token an
EP3567920	sending, by the serving	sending, by the	step	Step	7	7	Step 7 in sc	D1	1	D1			S3-170106		Step 7: The ta	Step 7: The targ		"Handling toke	ZTE: "Handling token an
EP3567920	wherein the first inform	wherein the first	step	Step	8	8	Step 8 in sc	D1	1	D1			S3-170106		Step 8:1..]. If s	Step 8: [...]. If e		"Handling toke	ZTE: "Handling token an
EP3567920	and – receiving, by the	receiving, by th	step	Step	9	9	Step 9 in sc	D1	1	D1			S3-170106		Step 9: The UI	Step 9: The UE		"Handling toke	ZTE: "Handling token an
EP2936876	forwarding to the user	identifying an ind	figure	Line	2	3-6	par. 2, in pa	D1	2	D1			R3-092795			Upon receipt of		"Handover opt	US 2012/008776 A1 (ISH
EP2936876	that by letting the targe	identifying an ind	section	Line	2	3-6	par. 2, lines	D1	2	D1			R3-092795			Upon receipt of		"Handover opt	US 2012/008776 A1 (ISH
EP2838308	A method for re-establi	A method for re-i	page	Page	53	53	figure 5.3.7	D1	1	D1	10.5.0	V10.5.0	TS 36.331	TS 36.331	eNB receiving			"3rd Generatio	"3rd Generation Partners
EP2838308	wherein a cell identifier	wherein a cell ic	section	Paragraph	5.3.7.4	5.3.7.4	paragraph	D1	1	D1	10.5.0	V10.5.0	TS 36.331	TS 36.331	Actions relate			"3rd Generatio	"3rd Generation Partners
EP2838308	and attribute informatio	and attribute inf	section	Paragraph	5.3.7.4	5.3.7.4	paragraph	D1	1	D1	10.5.0	V10.5.0	TS 36.331	TS 36.331	If the procedu	If the procedure		"3rd Generatio	"3rd Generation Partners
EP2429227	A method for updating	the destination R	figure	Figure		5	poaragraph	D1	1	D1			S3-091157		how a key hier	connection bet		"3rd Generatio	"3rd Generation Partners
EP2429227	to send the key inform	to send the key i	section	Paragraph	6.2.1	6.2.1	paragraph	D1	1	D1			S3-091157		an SGSN+ alw	an SGSN+ alwa		"3rd Generatio	"3rd Generation Partners
EP2429227	and the destination RNC	the destination R	figure	Figure	5.1.1-1	5	figure 5.1.1	D1	1	D1			S3-091157			connection bet		"3rd Generatio	"3rd Generation Partners
EP2429227	and calculating to obtai	and calculating	figure	Figure	4.5-1	4.5-1	figure 4.5-1	D1	1	D1			S3-091157			IKu and/or ciph		"3rd Generatio	"3rd Generation Partners
EP2429227	wherein said key inform	wherein said key	figure	Figure	4.5-1	4.5-1	figure 4.5-1	D1	1	D1			S3-091157			IKu and/or ciph		"3rd Generatio	"3rd Generation Partners
EP3648492	An inter-cell handover r	An inter-cell hanc	figure	Figure	5.4.4.14.2-	5.4.4.14.2-	Fig. 5.4.4.1	D1	1	D1		V2	S3-171583		Source cell	Source cell		"3rd Generatio	"3rd Generation Partners
EP3648492	a key retention policy c	a key retention p	figure	Figure	5.4.4.14.2-	5.4.4.14.2-	Fig. 5.4.4.1	D1	1	D1		V2	S3-171583		UE	UE		"3rd Generatio	"3rd Generation Partners
EP3648492	and the first indication r	and sending, by	figure	Figure	5.4.4.14.2-	5.4.4.14.2-	Fig. 5.4.4.1	D1	1	D1		V2	S3-171583		step (3): If retz	step (4) NG Ha		"3rd Generatio	"3rd Generation Partners

Fig. 16: Overview of the annotated and extracted fields with the highest similarity.

The following table (Table 5) contains the precision values extracted for the 48 analysed documents.

Table 5: Comparison between the annotations and the developed extraction tool.

	claim number	claim text	passage type	passage extracted	document number	document version	document category	quoted text	standard text
nº samples	207	224	202	202	187	50	62	124	184
Precision	97.09	71.3	75.62	84.58	85.03	100	88.52	74.8	82.61

The precision values were obtained by measuring the precision of the mappings that are common in both annotations and extracted results.

The obtained results show great applicability of our approach. The best performing values were obtained for claim number and the lowest precision value was of 71.3% which, for an initial approach, shows great room for improvement.

4.2.1.1 Passage type precision

Some annotated passage types were incorrectly annotated, for example for "*paragraph 5.3.7.4*" the annotated passage type was "*section*" and the extracted passage type was "*paragraph*". In this case, the annotated passage type should have been "*paragraph*". This is just one example impacting the final precision value.

If we consider that every time a paragraph extracted that was annotated as a section is actually a paragraph, we can achieve a 91% precision.

4.2.1.2 Document version precision

Although we achieved a 100% precision for the document versions that were both annotated and extracted, we identified that about 52% of extracted document versions were not annotated. This means that our algorithm is capable of extracting more version numbers compared to those that were manually annotated.

4.2.1.3 Document category precision

For extracting the document category, currently our algorithm only supports the TS/CR/TR/Tdoc/TSG categories. By comparing the existing extracted categories with the annotated ones, we reached a 88% precision. However, most of the categories found in the annotations include R3/S3/F that are not currently covered by our algorithm. This can be further improved by adding the new categories to the document category pattern.

Around 46% of the annotated document categories were empty, while our algorithm was capable of extracting these missing categories correctly (evaluated by manual check).

4.2.1.4 Quoted text precision

As mentioned in the previous sections there are a plethora of cases where the digitisation tool is not able to correctly identify some of the characters in the body of the text. Additionally, some manually induced errors were identified where the writer didn't properly format the text. As this situation mainly affects text with punctuation , the quoted text is the most affected category. The 74.8% precision score obtained in this category was the result of the implementation of a set of rules that try to cover as many edge-cases as possible. The implemented rules take into account variations identified during the manual check of the preliminary results to obtain the quoted text. Most limitations are described in the limitations section of this work.

4.2.2 NLPClaimMaps vs DI Results

The NLPClaimsMaps contain the results from an experiment conducted by GSMA that automatically extracts mappings between claims and prior art documents. We measured the performance of the NLPClaimMaps and DI results against the manual annotations provided by Septigent. This performance evaluation was only performed for the fields that are common in both approaches and annotations. On Table 6 the differences between the number of fields extracted by our solution versus the NLPClaimMaps is exposed.

Table 6: Comparison of the number of extracted fields between the NLPClaimMaps mappings and the developed extraction tool.

Extracted Fields	DI Solution	NLPClaimsMaps
patent number	X	X
claim number	X	X
feature text	X	
document passage text	X	X
document reference text	X	X

document passage type	X	
document passage extracted	X	
quoted text	X	
document number	X	X
document version	X	
standard text	X	X
document category	X	
3GPP citing	X	
document release	X	
publication date	X	
TOTAL	15	6

From the analysis of the existing extracted results for both approaches, the only fields we can compare the annotations with are the claim number, passage text, document number and standard text. A total of 49 annotated documents were analysed and the obtained results are summarised on Table 7.

Table 7: Comparison between the NLPClaimMaps mappings and the developed extraction tool.

	claim number	**passage text**	**document number**	**standard text**
NLPClaimMaps Precision	79.01	62.35	93.21	78.43
DI Results Precision	95.24	80.75	85.19	79.55

DI results show an increased performance compared to NLPClaimMaps except for the document number field. The main reason for this, is that we are not currently capturing all document numbers within the same section. For example, in Figure 17, our tool can only extract the document number 1 while NLPClaimMaps extracts both D1 and D2.

> **Documents D1** and **D2** disclose repectively the details of the authentication and attach procedure and are inter-linked with each other since both cite each other. These documents will be handled as a single document. Document **D1** and **D2** disclose in accordance with the following features of **claim 1** (the references in parentheses applying to these documents):

Fig.17: EP3531654 - Two documents being mentioned in the same mapping section.

By comparing both approaches we can conclude that DI approach not only returns overall better performance results, but also a bigger coverage of extracted fields. NLPClaimMaps

extracts 6 fields, while DI approach extracts almost double the number of fields - 15 different fields.

4.3 Limitations

In this section we will highlight some of the limitations identified in the current approach that can be improved in a longer time project.

4.3.1 Claim and document number extraction

The current approach considers that all the information about claim numbers and documents for a mapping is contained within a single section. However, we noticed a few cases where the claim is mentioned in a section different from the document number. In this case, the current approach is not able to link the claim to the corresponding document number. Fig.18 illustrates a concrete example, where claims 1-12 are mentioned in section 2., while document D1 is mentioned in section 2.2.

2 The present application does not meet the requirements of Article 52(1) EPC because the subject-matter of claims 1-12 is not new within the meaning of Article 54(1) and (2) EPC.

2.1 Documents D2 and D3 are considered as background documents, the contents of which is known to any person skilled in the art reading document D1, because D1 refers explicitly to D2 and D3 (see D1, paragraph 2), and in particular explicitly cites the document D2 for performing the steps 1-2 of Figure 4.1.1 (see D1, Paragraph 4.1).

2.2 Document D1 discloses

a first network unit of a device management, DM, network system (system UE-NAF-BSF of Figure 4.1.1, paragraph 4.1, where according to paragraph A. 2.5, 5th section "Characteristics", the NAF is taken to be the Device Management service, see "Source of the management message must be identifiable i.e. the NAF", together with the 4th section of paragraph of A.2.5. showing the embodiment with the source of the management message being the Device Management service), for enabling protection (4th section of paragraph of A.2.5., "pushed in a secure manner") of a bootstrap message ("secure push", 2nd and 3rd sections of paragraph A.2.5), the first network unit (D1, paragraph A.2.5, 1st section, "Device Management" service)

Fig. 18: EP2394452 - claim numbers mentioned in a section different from the document number.

This can be fixed by creating a rule that checks the section header (in case of a subsection).

4.3.2 Standard Quotation Transformer

The standard quotation transformer extracts the quoted text from the document reference. There are some limitations attached to the current approach:

1. Multiple quoted texts in the same string are not being extracted.
2. Quoted text not extracted if the writer forgets to close the first quotation mark (Fig. 19).

receiving, by the peripheral equipment, a second local interface shared key (ch. 2.1 -Proposed Solution, step 7:" *The BSF sends a request response message to the ME with the following payload: ...,KE,...*; step 6:"*...NAF derives Ks_int_SC from Ks_int_NAF... Ks_int_SC is encrypted as KE...*"; figure 1) which is calculated by

Fig. 19: EP1933498 - Missing quotation marks.

3. Textract is not able to recognise " ..." " - instead identifies it as a single dot " . ", thus it is not possible to extract this quoted text because there are no quotes (Fig. 20).

Proposed Solution, step 3: "...*(ME) sends a "service request" message to ... (NAF)...request may contain the following payload: an identity (B_TID), the terminal identity (IMEI)...*"; figure 1)

↓ **Textract output**

```
Proposed Solution, step 3: .(ME) sends a "service request" message to (NAF) request may contain the following payload: an identity
(B_TID), the terminal identity (IMEI). figure 1
```

Fig. 20: EP1933498 - Quotation marks not recognised by textract.

4. It's not always clear where to split the text and what's the actual quoted text.

Fig. 21 and Fig. 22 present a case where the text on the left column (manual annotated) is more comprehensive than the text on the left (the result of our tool). The precision score for this type of extractions will always be represented by a lower value.

an information transmission method, comprising: — receiving, by a serving radio access device	- receiving, by a serving radio access device	0.4623656

Fig. 21: EP3567920 - Comprehensive vs concise mapping problem in the quoted text.

said second message requesting the second network unit to provide the first network unit with a bootstrap key that is based on the information identifying the subscriber; said receiver is further configured to receive from the second network unit, a third message	- said receiver is further configured to receive from the second network unit, a third message	0.3484848

Fig. 22: EP2394452 - Comprehensive vs concise mapping problem in the quoted text.

5. In Fig. 23 the claim feature and the passage text are split in two different paragraphs, a situation not accounted for in our current implementation.

and sending, by the source cell, a handover request message to the target cell, wherein the handover request message is used to request to hand over the terminal device from the source cell to the target cell, the handover request message comprises first indication information (Fig. 5.4.4.14.2-1: step (4) NG Handover Request including retain-key-cell)***, and the first indication information is used to indicate whether the terminal device and the target cell use the first key to communicate with each other.***

(Fig. 5.4.4.14.2-1, step (3): If retain-key-cell is true, the source cell reuses the old key KgNB;

Fig. 23: EP1933498 - Claim feature and passage text in different paragraphs.

4.3.3 Standard Text Extractor

Fig.24 illustrates an example where the standard text extractor is not capable of extracting the standard text accurately. It is visible in the figure that D2 and D4 are not aligned with the first line of the standard text. Therefore, the first line will not be considered in the standard text.

1 The following document has been cited in the international search report; the numbering will be adhered to in the rest of the procedure.

D1 EP 1 737 192 A1

Reference is made to the following further documents; the numbering will be adhered to in the rest of the procedure:

D2 RESEARCH IN MOTION: "Clarification that GRUU should be used", 3GPP DRAFT; S2-082804, 10 April 2008 (2008-04-10), XP050265064,

D3 WO 2008/047195 A1

D4 SA3: "Reply LS on IMEI checking for SAE", 3GPP DRAFT; S2-072371_S3-070469_LS, 14 June 2007, XP050627420,

Fig. 24: Document numbers not aligned with standard text.

4.3.4 Document category extractor

Currently the document category extractor does not extract multiple document categories from within the same standard text document. This can be handled by adapting the regex to find all matches and adding them to different entries of the output document.

4.3.5 Document D 3GPP Extractor

The document D 3GPP Extractor does not recognise standard documents that do not contain the string 3GPP. Those documents won't be flagged as 3GPP citing.

4.3.6 Publication Date

The limitation attached to the publication date extractor consists in the extraction of false publication dates that are the ones that match one of the patterns of the publication date extractor. Example:

1. In the string “2006/085169” the pattern is matched and returns a date “2006/08”.

This might be solved by adding a word boundary in the regex pattern ‘*b*’.

4.3.7 Document References Extractor

The main limitations of the document references extractor include matching rules that were not considered during the development of this project. Some examples are presented below:

1. 1st and 2nd section ⟶ None (*not capable of extracting Section 1 and 2*)
2. Figure 2.1-Starting a new sentence ⟶ extracts “Figure 2.1-” (*should not extract '-', this can be handled with further post processing*)
3. Fig. 4a, 4b ⟶ None (*is not being extracted because the letter is lowercase, by supporting this case the number of false positive results would increase*)
4. paragraph [0003]; [0008] ⟶ None (*the pattern is not compatible with this type of document reference*)

This can be handled by adapting the regex so that it covers these edge cases.

5 GSMA-ESO Repository

The GSMA-ESO repository contains all the code required to reproduce the results presented in this report. The **bin** folder accommodates all user-facing scripts which can be used to run the entire pipeline (and its variations : local vs s3) , visualise bounding boxes for individual tokens, paragraphs, and sections, given an input file, and compute metrics. The **gsma_eso** folder contains all the supporting code required to run the pipelines. Inside there’s a **chunker** - responsible for chunking the document into meaningful pieces; an **extractor** - which contains all logic required to extract information from the mappings and the metadata section a **results** folder - where the code that prepares the final format of the results is implemented, and a **rule_running** folder - contain rules to capture the mappings text and the rule runner logic. There are two additional folders **tests** - where some validation code is run.

6 Conclusion

The proposed deliverables of this project included the creation of a dataset with digitised ESO documents, the extraction of patent-standards mappings, the rule-based extraction logic, and a report detailing the approach, metrics, limitations and next steps. This project covers all proposed deliverables. The implemented solution includes the creation of a dataset with 22,905 digitised ESO documents, using textract. The creation of a chunker that automatically chunks the digitised document into lines, paragraphs and sections. The development of a filtering technique that filters the sections with relevant mappings within the document. The implementation of a rule-based extraction logic capable of extracting not

only patent-standards mappings but also specific features and metadata from those mappings which are relevant for further processing (e.g. patent number, claim number, feature text, document passage text, document reference text, document passage type, document passage extracted, quoted text, document number, document version, standard text, document category, 3GPP citing, document release, publication date). A total of 187,382 mappings were extracted from the 22,905 documents. By comparing our approach with manual annotations, we obtained precision values of over 80% for most of the extracted fields.

7 Next Steps

In this section, we introduce suggestions for next steps that can follow up the work implemented in this project.

1. In the results section, we highlighted some limitations that were identified during the analysis of the extracted results. Most of the limitations can be fixed by improving the regex patterns so that they cover all edge cases identified. As for the next steps, and for improving the quality of the current extracted dataset, we propose the implementation of the potential solutions we identified on each section.

2. Currently, the extracted mappings are only based on the information contained in the ESO documents. A next step should be to consult the original claims and standard document files to compare the extracted information, and rectify, where needed, the extracted mappings. This is important because legal attorneys might introduce errors (e.g. paraphrasing, misquotes) when citing claims or standard documents. By comparing the extracted information with the original documents, we can deliver a higher quality dataset.

3. Not all mappings contain the quoted text of the standard mapping. In order to increase the standard mappings text, a next step could include the extraction of the standard mapping by using the extracted document types and passages and linking them with the original standard document.

4. The current approach does not filter patent wording such as “comprising” or “consisting”. These words do not belong to the claim text and are usually used by legal attorneys during the writing of the mappings. Thus, they are not relevant data for our dataset. We suggest the removal of these patent wording as a next step.

5. The rule-based extraction logic was implemented for English written ESO documents. Considering that the structure of the mappings will be the same for different languages (e.g. French, German), we can adapt the regex rules to cover ESO mappings from languages.