Orbital Witness - engineering coding exercise

Context

At Orbital Witness, we deal with a lot of text content as part of extracting insights from legal documents. One of the biggest challenges we face is the variable (and often poor) data quality in content coming from the original documents. This is especially true for documents that we have to OCR, however it can also be seen in third party data sources that we rely heavily on.

One example of this is for a document type that is provided by the Land Registry, known as Title Registers. Along with the PDFs that our users often review, we are also able to access structured, digitised content through the Land Registry's API. This exercise concentrates on a specific section of these documents known as the "Schedule of notices of leases" which outlines all of the sub-leases associated with the parent Title. The schedule is set out in a tabular format within the PDF document but is returned in a lossy format through the API so in order to make use of the content, we need to reconstitute the tabular structure.

The Data

When set out within the document, the Schedule of notices of leases takes the following form:

Schedule of notices of leases

	Registration date and plan ref.	Property description	Date of lease and term	Lessee's title
1	28.01.2009 tinted blue (part of)	Transformer Chamber (Ground Floor)	23.01.2009 99 years from 23.1.2009	EGL551039
2	09.07.2009 Edged and numbered 2 in blue (part of)	Endeavour House, 47 Cuba Street, London	06.07.2009 125 years from 1.1.2009	EGL557357
3	25.09.2009 Edged and numbered 3 in blue (part of)	Mayflower House, Westferry Road, London	10.09.2009 125 years from 1.1.2009	EGL560877
4	16.12.2009 Edged and numbered 4 in blue (part of)	Flat 1602, Landmark West Tower(sixteenth floor)	12.11.2009 999 years from 1.1.2009	EGL565026
5	17.12.2009 Edged and numbered 4 in blue (part of)	Flat 207, Landmark West Tower (second floor flat)	01.12.2009 999 years from 1.1.2009	EGL565086

In it, we have 4 columns of content set out:

1. **Registration date and plan ref.** - This is the date at which the sub lease was registered with the Land Registry and a qualitative description of how it is represented on the Title Plan (a separate document)
2. **Property description** - Typically the address of the sub lease property
3. **Date of lease and term** - The date the lease was executed and the duration of the lease from that date
4. **Lessee's title** - The unique identifier for the sub lease property

Some entries / rows also have one or more "notes" which run across all of the columns:

```
93      22.02.2010      Flat 2308 Landmark West       03.02.2010      EGL568130
        Edged and       Tower (twenty third floor     999 years from
        numbered 4 in   flat)                         1.1.2009
        blue (part of)
        NOTE: See entry in the Charges Register relating to a Deed of
        Rectification dated 26 January 2018
```

The response back from the API takes the following format (extracted from the wider API response):


```
{
    "leaseschedule": {
        "scheduleType": "SCHEDULE OF NOTICES OF LEASE",
        "scheduleEntry": [
            {
                "entryNumber": "1",
                "entryDate": "",
                "entryType": "Schedule of Notices of Leases",
                "entryText": [
                    "13.11.1996      Retail Warehouse, The         25.07.1996      SY664660   ",
                    "1 in yellow     Causeway and River Park       25 years from              ",
                    "Avenue, Staines              25.3.1995                  ",
                    "NOTE: The Lease comprises also other land"
                ]
            },
            {
                "entryNumber": "2",
                "entryDate": "",
                "entryType": "Schedule of Notices of Leases",
                "entryText": [
                    "13.11.1996      land adjoining The Causeway   25.07.1996      SY664662   ",
                    "2 in yellow     and River Park Avenue,        25 years from              ",
                    "Staines.                     25.3.1995"
                ]
            },
```


As you may see, the entryText is an array of strings which reflect the lines of text on the page rather than the cells in the table. The information that is lost as part of this format is that the cell data is delimited by whitespace, however these delimiters aren't always included for all 4 columns.

The first entry should be structured as so:

1. **Registration date and plan ref.** - 31.10.2016 1 in yellow
2. **Property description** - Retail Warehouse, The Causeway and River Park Avenue, Staines
3. **Date of lease and term** - 25.07.1996 25 years from 25.3.1995
4. **Lessee's title** - SY664660
5. **Note 1** - NOTE: The Lease comprises also other land

The Challenge

Below are links to (1) a JSON document with a number of examples of schedules such as the one above as well as (2) an example Title Register PDF with a particularly long schedule.

1. https://drive.google.com/file/d/1VcerdKYnvbiVSyuelbiTdU8mzPkYwNdu/view
2. https://drive.google.com/file/d/1UTiK9DHmXTxl7qm_ba0_hegK8irBch5M/view

Using whatever approach you see fit, prototype a solution for structuring the Schedule of notices of lease data so that the column data and optional notes can be referenced independently. Your solution should use either the JSON or the PDF provided as input but does not need to support both.

Guidelines

- This exercise is more about approach rather than completeness so, whilst a working example is preferable, comments / documentation of your thinking and considerations going in to the implementation.
- There will be an opportunity to go into more detail in person as part of the technical interview stage so please aim to spend 2-3 hours for this