STUDY CASE – STAGE OCR & DOCUMENT UNDERSTANDING

Conditions Générales d'assurance – Analyse, modélisation et extraction automatisée

1. Contexte

Dans le domaine de l'assurance, les Conditions Générales (CG) sont des documents contractuels centraux.
Elles définissent le cadre du contrat d'assurance : garanties, exclusions, montants, procédures, droits et obligations.

Ces documents sont utilisés dans de nombreux contextes :
• Consultation par les équipes métiers
• Support à la gestion des sinistres
• Alimentation de bases de données
• Indexation pour des moteurs de recherche ou des systèmes RAG

Le document fourni en pièce jointe est une CG d'assurance réelle, utilisée comme cas d'étude principal.

2. Input Documents

Deux documents sont fournis en pièce jointe :
– **Une Condition Générale** complète, utilisée comme document de référence pour analyser la structure et définir le schéma JSON d'extraction ;
– **Un fichier « Échantillon »,** plus court, sur lequel sera réalisée l'extraction automatisée afin d'évaluer les compétences techniques d'OCR, de layout et de structuration des données.

3. Objectif du study case

Votre mission est de concevoir une solution automatisée permettant de transformer un document d'assurance (PDF / screenshots) en données structurées et exploitables, prêtes à être intégrées en aval (base de données, moteur de recherche, système RAG).

Ce study case vise à évaluer :
• Votre compréhension du rôle d'une CG en assurance
• Votre capacité à modéliser des données
• Votre maîtrise des problématiques OCR et layout
• Votre raisonnement technique global

4. Travail attendu

Le travail est structuré en trois étapes, correspondant à un projet réel de document understanding.

Étape 1 – Compréhension métier (recherche préalable)

Avant toute implémentation, vous devez effectuer une courte recherche afin de comprendre:
• Ce que sont les Conditions Générales en assurance
• Dans quel cadre elles sont utilisées
• Quelles sont les grandes parties que l'on retrouve généralement dans les documents de Conditions Générales

À titre indicatif, une CG contient souvent :
• Un sommaire
• Des définitions
• La description du contrat et des biens assurés
• Les garanties
• Les exclusions
• Les montants, plafonds et franchises
• Les procédures en cas de sinistre
• Les règles de vie du contrat
• Les dispositions d'assistance

Cette étape vise à vérifier votre compréhension fonctionnelle, et non à produire un rapport théorique.

Étape 2 – Analyse de la CG fournie et modélisation JSON

En vous concentrant sur **la CG fournie en pièce jointe**, vous devez :
• Identifier les différentes parties réellement présentes dans le document
• Comprendre leur organisation et leurs relations
• Proposer un schéma JSON permettant de représenter cette CG de manière structurée, en vue d'un stockage en base de données ou d'une exploitation future (ex. RAG)

Pour vous guider, une Template JSON de référence est fournie ci-dessous.
Vous pouvez l'adapter ou l'enrichir si vous justifiez clairement vos choix.

Template JSON de référence :

C: > Users > EmnaFazaaTendanz > Downloads > {} exemple.json > [ ] garanties > { } 0

```json
{
  "metadata": {
    "nom_produit": "Assurance Automobile",
    "assureur": "",
    "type_doc": "Dispositions générales",
    "line_of_business": "Auto",
    "date_effet": "07/2023",
    "lang": "fr"
  },

  "preambule": "",

  "garanties": [
    {
      "nomGarantie": "",
      "objet_de_garantie": "",
      "beneficiaire": "",
      "conditions_mise_en_oeuvre": "",
      "exclusions": [],
      "plafond_de_garantie": "",
      "franchise": "",
      "indemnite": ""
    }
  ],

  "formules": [
    {
      "nom": "",
      "garantie": [
        { "nomGarantie": "" }
      ]
    }
  ]
}
```

L'exemple ci-dessus illustre **le type de structuration attendu**, sans imposer une modélisation unique ni exhaustive.
Il sert uniquement à montrer **comment organiser les informations clés** d'une Condition Générale de manière exploitable sans rentrer dans les détails (c'est à vous de le faire).

Hints importants

- Une garantie = un objet.
- Les exclusions doivent être structurées en liste lorsque possible.
- Les informations générales du document sont regroupées dans « metadata ».
- Le préambule est conservé comme un bloc textuel.

Étape 3 – Extraction intelligente et automatisée

À partir du **fichier "Échantillon"** fourni, vous devez implémenter une solution **entièrement automatisée** permettant de remplir, au moins partiellement, le schéma JSON défini à l'étape 2.

Aucune règle manuelle par page ou par document n'est attendue.
La solution doit pouvoir être exécutée de bout en bout via une commande ou un script unique (input → output JSON).

Votre extraction doit prendre en compte :
• La variabilité du layout (titres, paragraphes, multilignes)
• Le mapping champ / valeur (labels variables, distances visuelles)
• La gestion des champs manquants ou ambigus
• Le traitement des images (logos, tampons, signatures, pictogrammes).
Si aucune image n'est détectée, cela doit être indiqué explicitement dans la sortie (ex. champ vide ou liste vide).

Exigence spécifique – Reconstruction des tableaux

Les tableaux doivent être reconstruits de manière structurée, afin de permettre l'extraction des données et l'expression des relations entre les éléments, et non pas sous forme de texte brut.

La solution doit impérativement :
• Détecter les zones de tableau et leurs limites
• Identifier la structure du tableau (en-têtes, colonnes, lignes, totaux, remarques)
• Associer correctement chaque cellule à son en-tête, y compris en cas d'en-têtes multilignes
• Produire une représentation exploitable (obligatoire : headers + rows, ou un objet par ligne avec mapping header → valeur)

La compréhension du layout du tableau est un critère central d'évaluation.
Toute sortie où un tableau est fourni uniquement sous forme de texte brut sera considérée comme insuffisante.

4. Livrables attendus

Le candidat devra fournir :

- ✓ Le code source, sous forme de dossier compressé ou de lien vers un repository GitHub.
- ✓ Un rendu écrit comprenant :
    - o le schéma JSON de structuration défini à l'étape 2 ;
    - o un schéma du pipeline d'extraction mis en œuvre à l'étape 3.
- ✓ Une démonstration rapide, lors de l'entretien, de l'exécution de l'étape 3.
- ✓ Le résultat de l'extraction du fichier « Échantillon », généré automatiquement au format JSON, structuré et cohérent.

La complétude parfaite n'est pas attendue.
La priorité est donnée à la qualité de structuration et à la logique d'extraction.

5. Liberté technique

Vous êtes libre dans le choix :
• Des technologies
• Des bibliothèques
• Des modèles OCR / vision
• De l'architecture globale

L'important est de proposer une solution automatisée, justifiée, et orientée données exploitables.

6. Critères d'évaluation

L'évaluation portera sur :
• La compréhension métier
• La pertinence de la modélisation
• Le niveau d'automatisation
• La gestion du layout et des tableaux
• La clarté du raisonnement technique
• La capacité à généraliser (pas de hardcoding par page)

7. Points importants

• Il n'existe pas de solution unique attendue
• Le raisonnement prime sur la quantité de code
• L'automatisation est un critère central
• L'objectif est de produire des données structurées et exploitables

Bonne préparation et à très bientôt pour l'échange.