Noctu@edu
COMPRENDRE, ÉCLAIRER, TRANSMETTRE

A1

PRACTICAL
INVESTIGATION SERIES

SÉRIE A

Détecter la fraude dans les documents PDF

Manuel pratique d'investigation documentaire



- Métadonnées
- Texte & polices
- Images & OCR
- Structure PDF
- Signatures
- Cas pratiques

Anthony Vandepeutte

> Le total demandé dépasse de 480 euros la somme des lignes et de la TVA. La facture indique également un IBAN différent de celui du référentiel. Le règlement sur ces coordonnées est mis en attente jusqu'au contrôle du montant et du bénéficiaire auprès du fournisseur par le canal déjà connu.

Cette conclusion a un objet, des faits et une action. Le chapitre consacré à la facture poursuit l'enquête : comparaison des fichiers, réponse de l'émetteur fournie dans le scénario, puis qualification de la pièce. Vous pourrez y vérifier pourquoi les preuves permettent alors d'aller plus loin.

La méthode en cinq gestes

Le même fil guide tout le livre.

1. **Préserver** le fichier reçu, son empreinte et les conditions de sa réception.
2. **Observer** la page, le texte extrait et les éléments internes qui répondent à la question posée.
3. **Tester** l'explication envisagée et celle qui pourrait la contredire.
4. **Corroborer** les faits auprès d'une source indépendante ou d'une référence dont la provenance est connue.
5. **Décider** en nommant ce qui est établi et la suite à donner.

Les grilles et les fiches complètent ces gestes lorsque vous en avez besoin. Vous n'avez pas à les mémoriser avant de commencer.

Choisir votre point de départ

Vous n'avez pas besoin de lire tous les chapitres avant d'examiner une première pièce. Après le socle commun, choisissez ci-dessous le besoin qui ressemble au vôtre. Les chapitres techniques restent disponibles pour expliquer une trace rencontrée en cours d'enquête.

Le socle commun

Commencez par le chapitre 1. Il explique comment conserver le document reçu, poser la question du contrôle et distinguer un fait d'une hypothèse. Suivez ensuite les **quatre étapes du fichier miniature au chapitre 2**, jusqu'à la fin de l'étape 4. Le dossier **Anatomie pas à pas** du kit accompagne cette découverte sans installation.

Ce départ commun suffit pour aborder les enquêtes. La suite du chapitre 2 et les commandes de collecte peuvent attendre le moment où vous en aurez besoin.

Vous devez vérifier un justificatif

Choisissez d'abord le document qui vous concerne :

- **Un bulletin de salaire** : le chapitre 10 montre comment vérifier les rubriques, le net et les cumuls.
- **Une facture** : le chapitre 11 suit le contrôle des montants et du compte bénéficiaire.
- **Un relevé bancaire** : le chapitre 12 explique comment vérifier les opérations, les soldes et une extraction erronée.

Vous pouvez commencer par une seule de ces enquêtes. Dans le kit, ouvrez les pièces correspondantes depuis `COMMENCER-ICI.html`, refaites les calculs et écrivez votre conclusion avant le corrigé. Un navigateur, un lecteur PDF et de quoi noter vos résultats suffisent pour cette première lecture.

Consultez ensuite le **chapitre 15** pour présenter le résultat. Lorsque vous souhaitez vous entraîner seul, passez à l'enquête du **chapitre 17**. Les deux autres cas métier pourront être lus plus tard ; ils ne sont pas des prérequis.

Vous devez examiner le fichier en détail

Poursuivez le **chapitre 2** après la miniature, puis choisissez le chapitre correspondant à la trace que vous cherchez à expliquer :

- **Dates et logiciels déclarés** : chapitre 3.
- **Texte, polices ou montant mal extrait** : chapitre 4.
- **Images, scans ou OCR** : chapitre 5.

CHAPITRE 2

Comprendre la structure d'un fichier PDF

Une page PDF donne l'impression d'être une feuille figée. Pourtant, ce que vous voyez résulte de plusieurs composants : des instructions de dessin, des polices, des images, parfois des annotations et des champs. Le lecteur retrouve ces composants, résout leurs références et compose l'apparence. Pour expliquer un changement de montant, il faut apprendre à suivre ce chemin.

Nous allons le faire sur un fichier suffisamment petit pour être parcouru entièrement. Cette première lecture ne demande ni programmation ni connaissance préalable de la norme. À la fin de la miniature, vous saurez relier une valeur visible à son contenu et distinguer un numéro d'objet de son adresse. L'approfondissement montrera ensuite ce qui change dans les fichiers ordinaires : compression, ressources partagées, autres tables de références et récupération.

Pour suivre : ayez lu le chapitre 1 et ouvrez le dossier **11_ANATOMIE_PAS_A_PAS** du kit. Un navigateur et un lecteur PDF suffisent aux quatre premières étapes. Les commandes de la fin sont facultatives. Si vous reprenez votre parcours de lecture après la miniature, vous conserverez un socle suffisant pour comprendre les cas métier.

2.1 Suivre un montant dans un fichier simple

Ouvrez INDEX.html, puis le PDF et sa transcription octets.html. Le spécimen contient **755 octets**. Il a été fabriqué pour cet apprentissage, avec un contenu non compressé et des caractères ASCII lisibles. Il ne

représente aucune facture réelle. Les positions citées dans ce chapitre appartiennent à ce fichier précis : dans un autre document, même une apparence identique pourrait être produite par une autre organisation.

Notre question est volontairement simple : **d'où vient le montant affiché ?** Nous suivrons d'abord les références qui relient la page à ses instructions. Nous regarderons ensuite les adresses qui permettent au logiciel de trouver ces instructions dans les octets. Ces deux chemins sont complémentaires. Le premier explique le rôle des composants ; le second explique où ils se trouvent.

La transcription est un support de lecture, pas une invitation à modifier la source. Un éditeur de texte peut changer des fins de ligne, un encodage ou des octets binaires. Dans un PDF, ces changements peuvent rendre fausses les adresses de la table. Gardez donc le spécimen fourni comme référence et utilisez la transcription pour observer.

Il n'est pas nécessaire de mémoriser tous les mots-clés. Pour chaque élément, posez une question concrète : « Est-ce la description de la page ? Une instruction ? Une ressource ? Une adresse ? » Le vocabulaire sera plus facile à retenir lorsque chaque nom correspondra à une fonction rencontrée.

2.2 Étape 1 Observer la page



Figure 2.1. – Rendu du fichier miniature.pdf fourni dans le kit. Le montant visible est 1 250,00 EUR.

Le lecteur affiche **1 250,00 EUR**. Notez le libellé, la valeur et son emplacement. Dans un dossier réel, ce premier relevé fixe ce que vous cherchez à expliquer. Il évite de confondre ensuite un autre montant trouvé dans les octets avec celui qui se trouve effectivement sur la page.

Essayez de sélectionner le texte. Dans cette miniature, il s'agit de texte dessiné par une police. La sélection fournit une indication, mais ce geste ne suffit pas pour un autre fichier : une image peut être accompagnée d'une couche textuelle invisible. Le chapitre 4 vous apprendra à reconnaître cette différence.

Dans octets.html, cherchez maintenant **`/Type /Page`**, au singulier. Vous trouverez la description de notre unique page. Ne vous laissez pas guider par le numéro d'objet comme s'il s'agissait d'un numéro de page. Les objets du fichier sont des composants ; certains ne produisent rien de visible.

2.3 Étape 2 Retrouver le contenu et les ressources

Voici l'objet qui décrit la page :

CHAPITRE 8

Contrôler les QR codes et les codes 2D-Doc

Pour entrer dans ce chapitre. Le chapitre 1 suffit pour les premiers contrôles. Le chapitre 7 aide à comprendre la signature et le temps de validation. Nous avancerons du symbole visible vers les données, puis vers la preuve d'origine. Une application qui affiche du texte après un scan n'a pas forcément réalisé toutes ces étapes.

Un justificatif arrive avec un carré noir et blanc. Le téléphone le reconnaît et annonce « code lu ». L'opérateur se rassure : le document semble sécurisé. Mais que vient-il de lire ? Un lien vers une page web ? Des données déclaratives ? Des champs protégés par une signature ? Et ces champs décrivent-ils réellement le justificatif présenté ? Les réponses changent entièrement la portée du contrôle.

Dans ce chapitre, nous allons suivre un code jusqu'à son usage dans la décision. L'objectif est de pouvoir expliquer au lecteur du rapport ce qui a été décodé, ce qui a été authentifié et ce qui concorde avec la page. Un code correctement signé peut rester associé à la mauvaise pièce. Une page sans code peut être parfaitement légitime. Le travail consiste à établir les relations utiles, en conservant les données qui permettront de refaire le contrôle.

8.1 Identifier le type de code présent

Le mot « QR » est souvent employé pour n'importe quel code carré. Il faut distinguer la **symbologie**, c'est-à-dire la manière graphique d'encoder l'information, du **format de données**, qui donne un sens à l'information obtenue. Un QR Code et un DataMatrix sont des symbolo-

CHAPITRE 16

L'IA dans l'analyse et la falsification des PDF

Une réponse peut être parfaitement rédigée et se tromper sur le document qu'elle prétend expliquer. C'est une difficulté particulière de l'analyse assistée par intelligence artificielle : la fluidité de l'explication rend parfois moins visible l'absence de vérification. Devant un PDF, le modèle peut lire correctement un nom, manquer une ligne de tableau et annoncer malgré tout que l'ensemble est cohérent. La première compétence à acquérir est donc de savoir lui confier une tâche dont on pourra contrôler le résultat.

Dans ce chapitre, nous examinerons aussi la situation inverse : un document dont le texte ou les images ont pu être produits ou modifiés à l'aide de l'IA. La question mérite une place dans l'enquête, mais elle ne doit pas absorber toutes les autres. Si un montant est faux, il reste faux que sa modification provienne d'un éditeur classique, d'un script ou d'un outil génératif. Établir la discordance, identifier les informations qui la corroborent et rechercher son origine sont trois opérations différentes.

Vous apprendrez à organiser une analyse assistée sans déléguer la conclusion, à conserver ce que le modèle a réellement reçu, à reconnaître une réponse dépourvue de preuves et à tester les hypothèses de falsification. Le travail pratique TP14 utilise des pièces fictives et des réponses d'IA **simulées**. Aucun modèle n'a été exécuté pour produire ses résultats. Cette distinction permet de travailler sur la méthode sans transformer l'exercice en démonstration de performance d'un produit.