



ORPALIS
imaging technologies



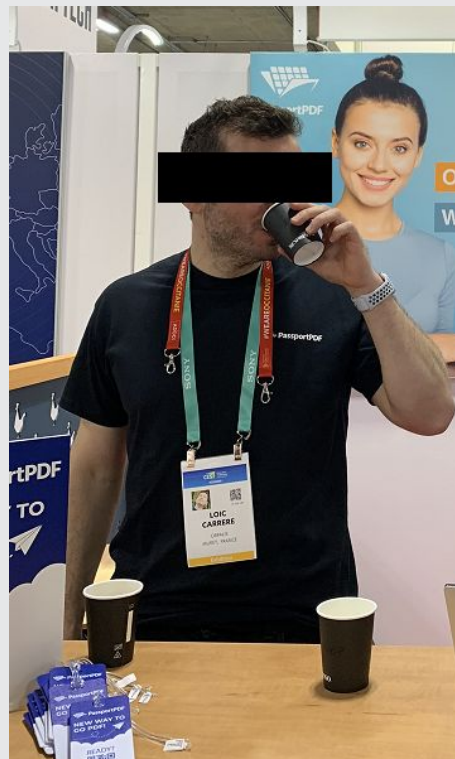
Supprimez les données sensibles et personnelles de vos PDF

Sécurisez vos documents avec la rédaction

OctoberPDFest
ONLINE

- Introduction : un outil critique
- Définitions : “redaction”, “sanitization” et autres appellations
- Processus manuel
- Processus automatique
- Conservation et archivage des documents obtenus
- Conclusion : à garder en tête
- Nos outils de rédaction
- Questions

Loïc CARRERE
CEO, CTO et développeur



Elodie TELLIER
DG et membre du CA de la
PDF Association

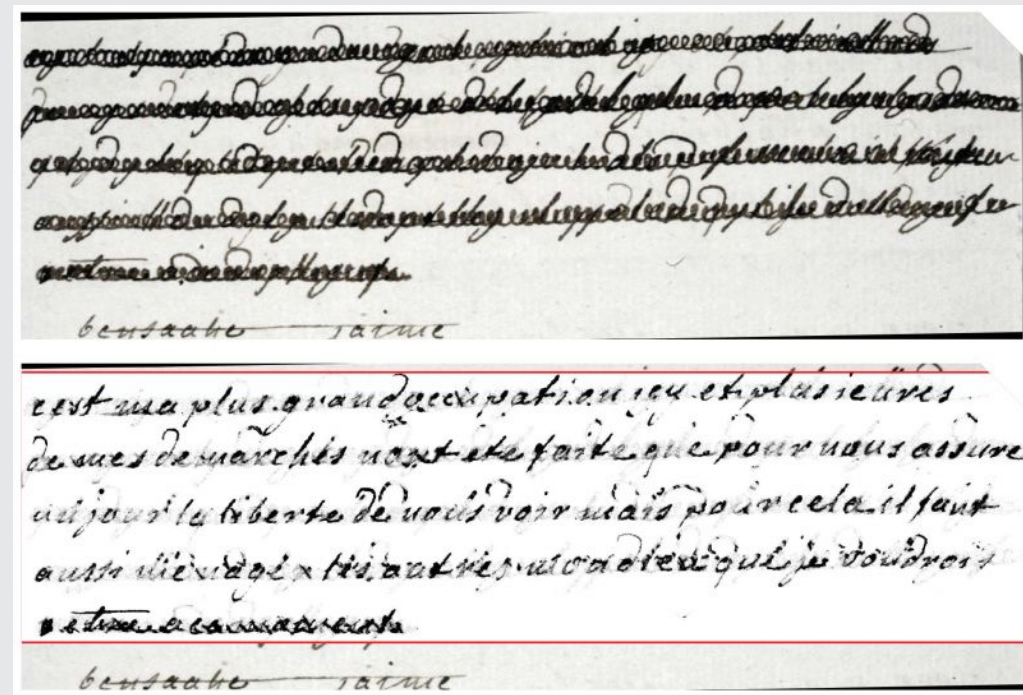


Avec l'aide de Matus PIZUR
développeur sénior et spécialiste
du PDF



Introduction : caviarder, censurer, biffer...

Lettre de Marie-Antoinette au comte de Fersen



<https://inshs.cnrs.fr/fr/cnrsinfo/ecritures-masquees-ecritures-revelees-la-correspondance-de-marie-antoinette-aux-rayons-x>

Le rapport Mueller

II. Russian “Active Measures” Social Media Campaign

22 pages



<https://www.npr.org/2019/04/18/713974980/see-how-much-of-the-mueller-report-is-redacted>

MOTHERBOARD
TECH BY VICE

How a Simple Copy/Paste Revealed Explosive New Detail in Manafort's Case

Redacted court documents keep getting unredacted because of simple, relatively easy to avoid errors. Here's how to better redact PDFs.



By Lorenzo Franceschi-Bicchieri

https://www.vice.com/en_us/article/8xpye3/paul-manafort-russia-case-redaction-fail

https://www.americanbar.org/groups/judicial/publications/judges_journal/2019/spring/embarrassing-redaction-failures/

La **loi** : certains domaines ont l'obligation de supprimer des informations
ex. : justice / médecine / secret défense

Les réglementations sur les **données personnelles**



Un peu d'anglais...

Redaction is the process of permanently removing visible text and graphics from a document.

Pour les **documents électroniques** :

Ce traitement se fait généralement avant la **publication**, **distribution** ou l'**archivage** de documents électroniques.

Pour le PDF, il ne s’agit pas de supprimer uniquement la partie visible du document électronique, il faut aussi pouvoir **supprimer des informations cachées** :

- métadonnées (auteur, titre, date de création, etc.)
- annotations
- commentaires
- couches de texte caché (sous les annotations par ex.)

Les points **sensibles** à vérifier :

- images de formats différents
- les tableurs
- les fichiers joints
- les index



Ceci n'est pas une rédaction :

- annotation
- encryption
- supprimer uniquement les informations cachées



1. ouverture du document
2. création des zones de rédaction :
 - localisation
 - couleur
3. appliquer les rédactions
4. supprimer les informations cachées (pas nécessaire pour les PDF images)

Le « Portable Document Format »

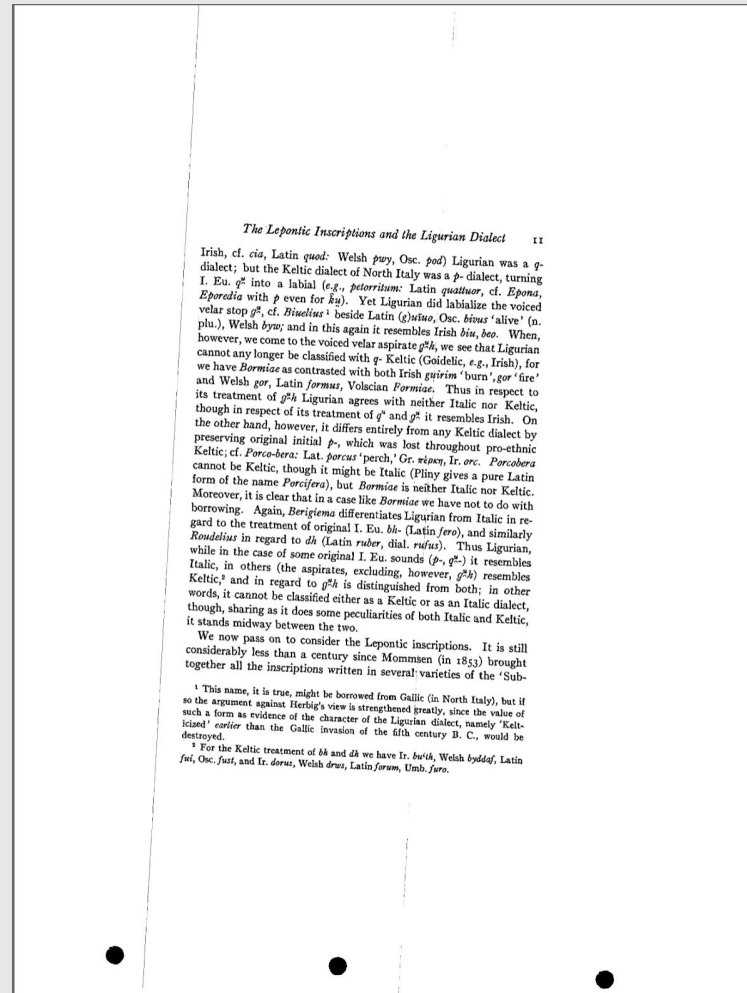
En 1991, [REDACTED] le co-fondateur d'Adobe Systems, annonce Camelot, un système qui est devenu par la suite « Portable Document Format » (PDF). Celui-ci est publié pour la première fois en 1993. La même année, Adobe Systems a rendu publiques les spécifications du format PDF afin que n'importe quel développeur puisse utiliser ce format. En parallèle, Adobe décide d'offrir gratuitement Adobe Reader. Le format PDF est resté propriété d'Adobe Systems jusqu'à sa sortie officielle comme norme internationale et a été publié par l'Organisation Internationale pour la Standardisation comme ISO 32000-1 le 1er juillet 2008.



Les fondateurs d'Adobe [REDACTED] et Chuck Geschke

Phase de **pré-traitement** pour les documents **scannés** :

- nettoyage (redressement, suppression des parasites, contraste, etc.)
- OCR



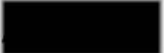
Définition des informations à supprimer : regex

Introduction

ISO 32000 specifies a digital form for representing documents called the Portable Document Format or usually referred to as PDF. PDF was developed and specified by Adobe Systems Incorporated beginning in 1993 and continuing until 2007 when this ISO standard was prepared. The Adobe Systems version PDF 1.7 is the basis for this ISO 32000 edition. The specifications for PDF are backward inclusive, meaning that PDF 1.7 includes all of the functionality previously documented in the Adobe PDF Specifications for versions 1.0 through 1.6. It should be noted that where Adobe removed certain features of PDF from their standard, they too are not contained herein.

Vérification et assainissement

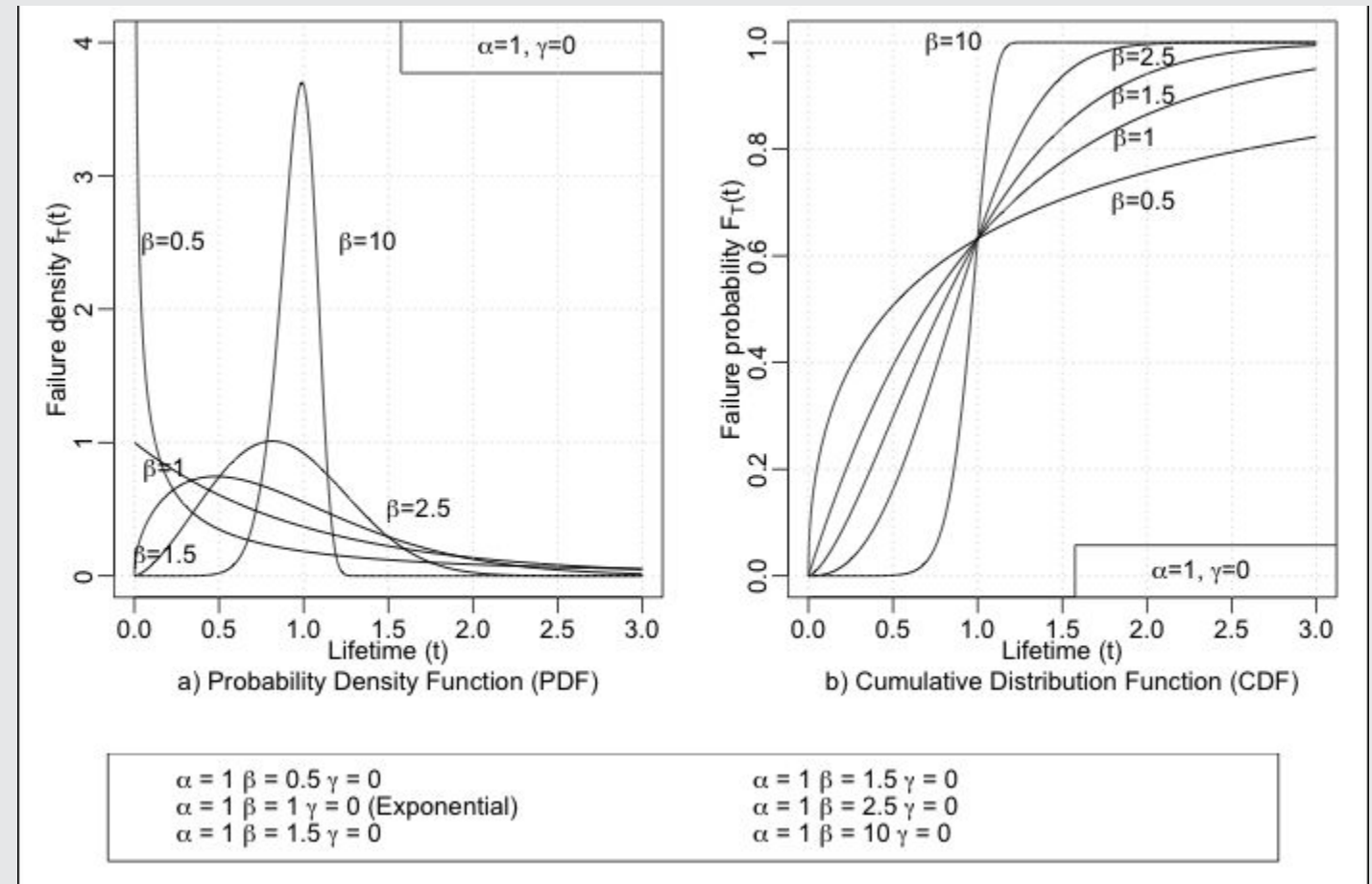
Introduction

ISO 32000 specifies a digital form for representing documents called the Portable Document Format or usually referred to as PDF. PDF was developed and specified by  Systems Incorporated beginning in 1993 and continuing until 2007 when this ISO standard was prepared. The  Systems version PDF 1.7 is the basis for this ISO 32000 edition. The specifications for PDF are backward inclusive, meaning that PDF 1.7 includes all of the functionality previously documented in the  PDF Specifications for versions 1.0 through 1.6. It should be noted that where  removed certain features of PDF from their standard, they too are not contained herein.

Automatiser à 100 % ?

Il est difficile de se passer complètement de l'**humain** :

- erreurs d'**OCR** (dans les tableaux, graphiques, position du texte), notamment sur les textes scannés non nettoyés
- **images** (attention aux différents formats)



Baptista, Márcia & de Medeiros, Ivo & P. Malere, João Pedro & Nascimento Jr, Cairo & Prendinger, Helmut & Henriques, Elsa. (2016). Improved Time-Based Maintenance in Aeronautics with Regressive Support Vector Machines.

NF Z 42-013 : 1999, 2009

ISO 14641-1 : 2012 (norme internationale)

NF 461 - Système d'archivage électronique. 2013

OAIS (ISO 14721) : 2002, 2012

ISO 19005-1 - PDF/A. 2005, 2011, 2012 (et à venir)

<https://www.pdfa.org/presentation/pdf-a-4-pdf-x-6-and-the-other-new-pdf-standards/>



La règle d'or dans le PDF :

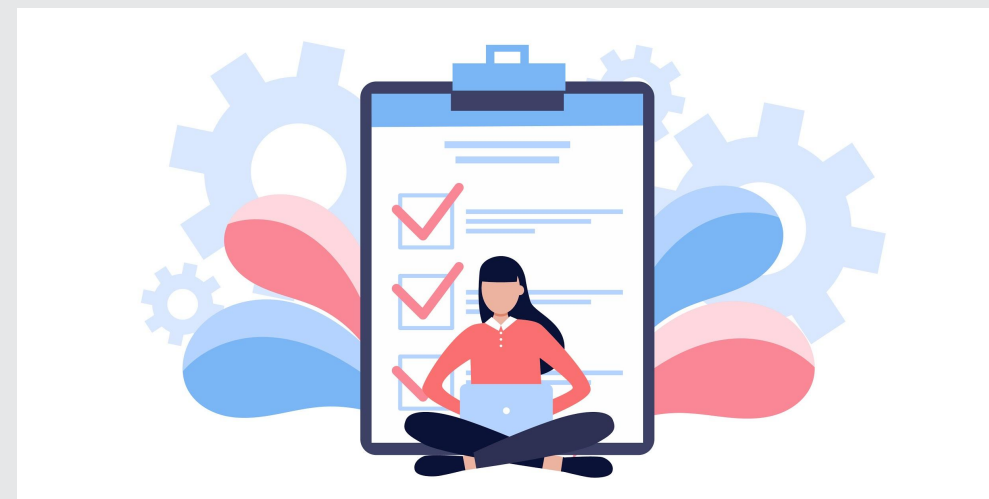
ce n'est pas parce que c'est décrit dans la spécification que l'outil que vous utilisez le fait

La spécification :

- PDF 1.7 : ISO 32000-1, 12.5.6.23 Redaction annotations
- PDF 2.0 : ISO 32000-2:2017, 14.8.2.2.2 (new artifact structure element type)

Conclusion : la rédaction n'est qu'une étape !

1. définir le type d'information à supprimer
2. choisir une méthode selon le type de PDF
3. **tester** différents outils
4. définir le moment de rédaction
5. établir les rôles et permissions des membres de l'équipe
6. gérer la distribution et/ou de l'archivage des documents modifiés, mais aussi des originaux !



Attention au workflow pour les données confidentielles, de l'acquisition à l'archivage !

GdPicture.NET

- outil complet pour la gestion des documents électroniques
- **PDF** SDK
- **OCR** SDK
- **formats** SDK
- **barcode** SDK
- **imaging** SDK

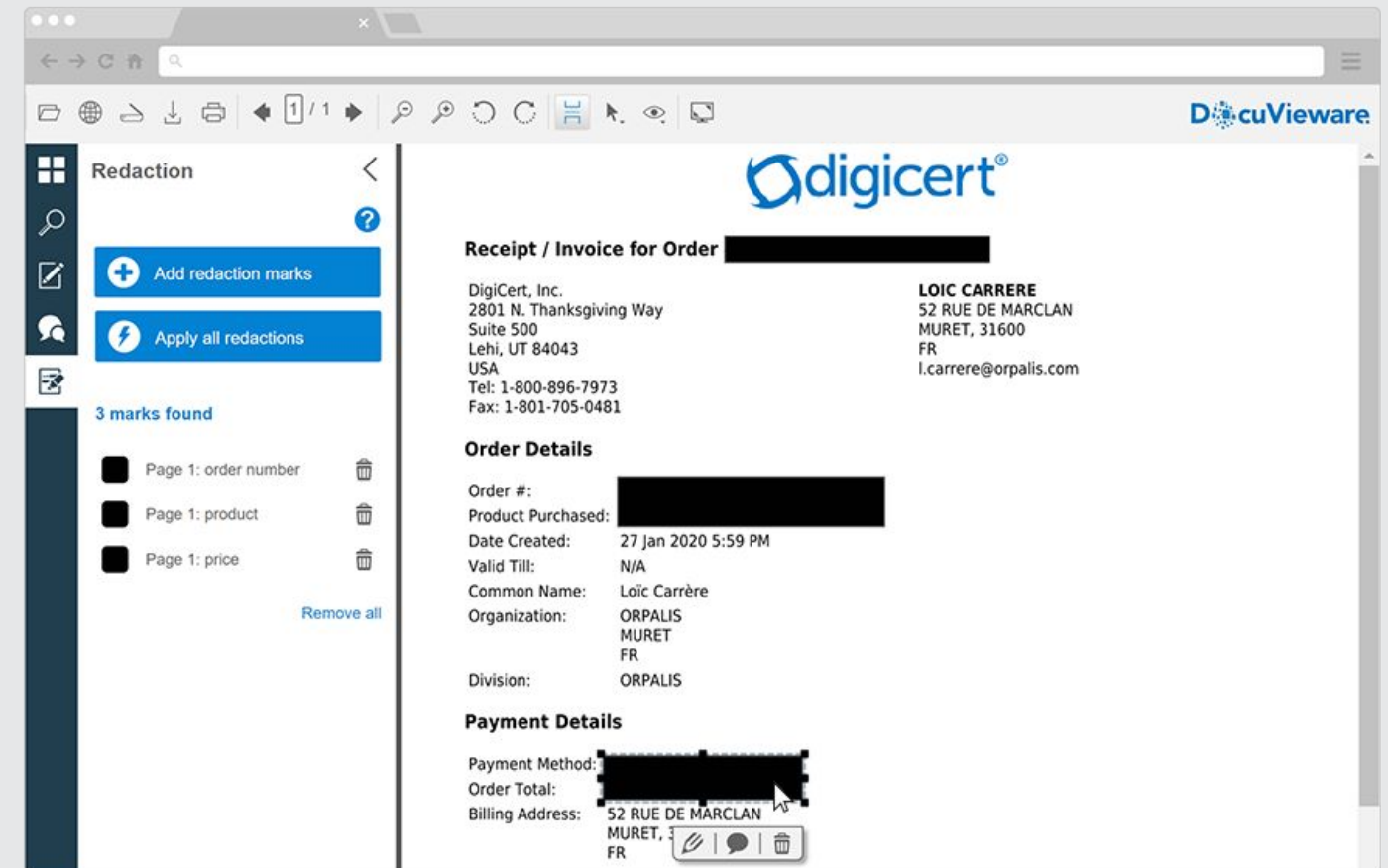
→ Adds redaction region on the selected page of the loaded PDF document.

```
using (GdPicturePDF oGdPicturePDF = new GdPicturePDF())
{
    oGdPicturePDF.LoadFromFile("input.pdf");
    oGdPicturePDF.SetOrigin(PdfOrigin.PdfOriginTopLeft);
    oGdPicturePDF.SelectPage(1);
    oGdPicturePDF.AddRedactionRegion(120, 300, 140, 20);
    oGdPicturePDF.SelectPage(2);
    oGdPicturePDF.AddRedactionRegion(100, 280, 70, 15);
    oGdPicturePDF.AddRedactionRegion(100, 310, 70, 15);
    oGdPicturePDF.ApplyRedaction();
    oGdPicturePDF.SaveToFile("output.pdf");
}
```

<https://www.gdpicture.com/pdf-sdk/pdf-redaction/>



Capturez, visualisez, modifiez,
annotez, biffez, convertissez, traitez,
compressez, gérez et signez tous vos
documents **en ligne**



<https://www.docuviewware.com/demo/>

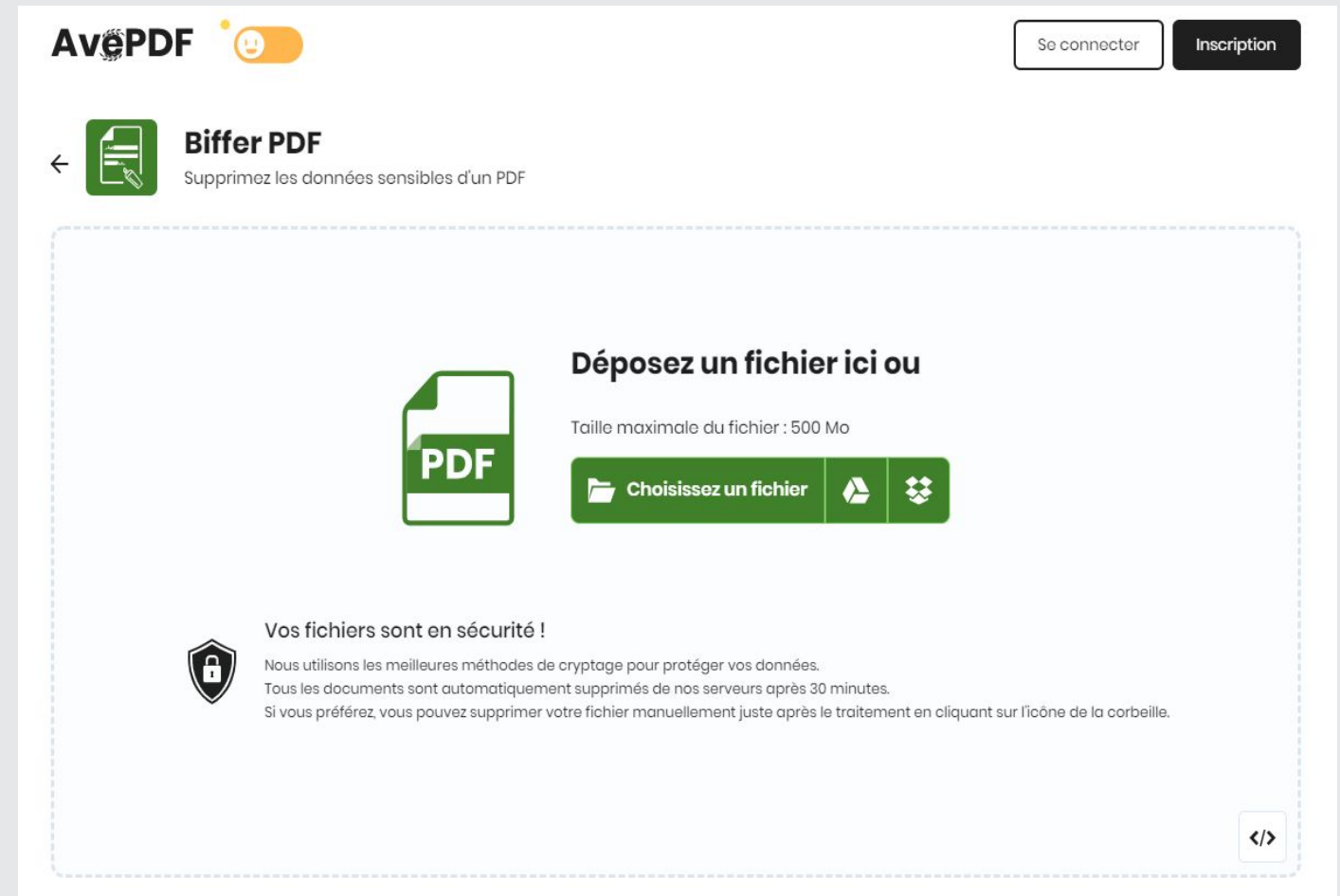


APIs, applications et microservices
pour le traitement des documents
dans le Cloud



Microservices

- outil **gratuit** en ligne pour les PDF et les documents électroniques
- sécurisation des processus
- des fonctionnalités pour les professionnels



<https://avepdf.com/fr/redact-pdf>



QUESTIONS

Restons en contact: e.tellier@orpalis.com / l.carrere@orpalis.com

<https://www.orpalis.com/fr/>