PDF contenant des polices de caractères défectueuses : Impossible de rechercher et d'extraire du texte

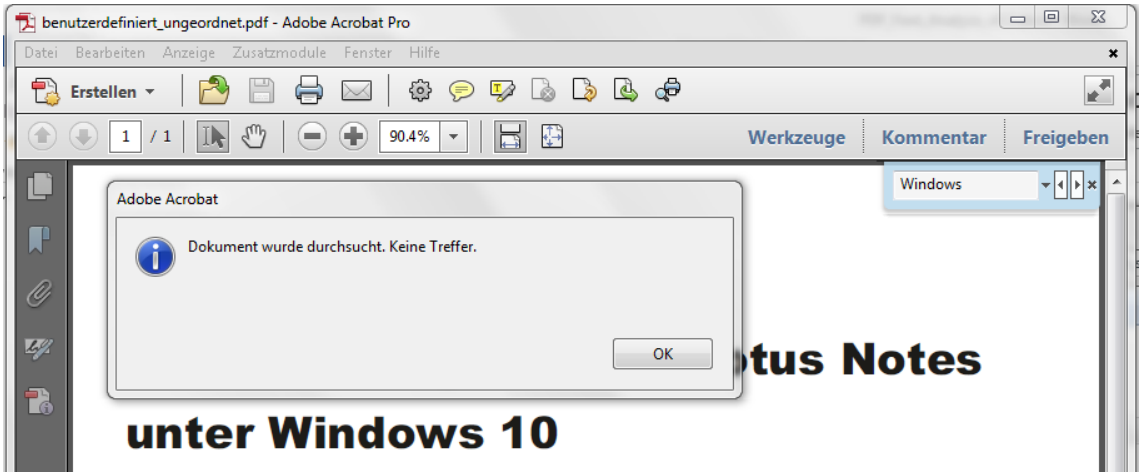
Groupe d'experts Preservation planning : rapport & recommandations

1 Incident 2
1.1 Description 2
1.2 Répercussions 2
2 Analyse 3
2.1 Document affecté 3
2.2 Évaluation et Cfa 4
2.2.1 Résultat de l'évaluation 4
3 Recommandations 6
3.1 Pour éviter le défaut 6
3.2 Pour identifier le défaut 6
3.3 Pour réparer le défaut 6
3.3.1 Réparation du défaut : les modules D et K sont signalés comme défectueux 7
3.3.2 Réparation du défaut : Le module K est signalé comme défectueux, mais pas le module D 7
3.3.3 Contrôle de la réparation des défauts 9
3.3.4 Problèmes lors de la réparation des défauts 9
3.4 Recommandations : résumé 10

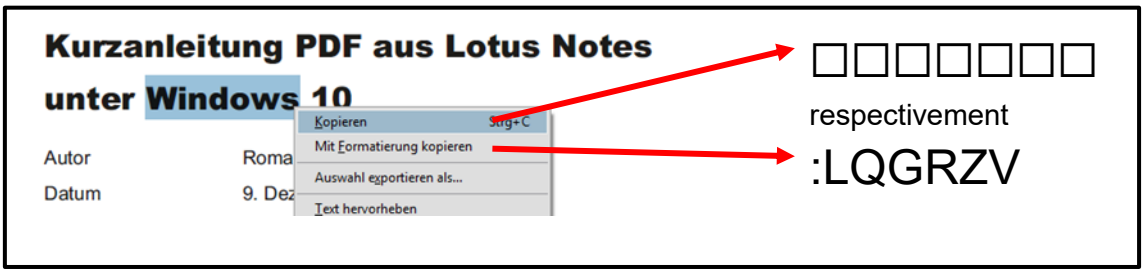
1 Incident

1.1 Description

Suite à un changement de système, les Archives de l'État de Zurich ont constaté qu'il n'était plus possible d'effectuer une recherche dans le document PDF/A nouvellement créé.



La copie de passages du texte ne restitue que des « caractères inconnus » ou ne restitue pas les bons caractères.



1.2 Répercussions

Il est certes encore possible de lire et d'imprimer les fichiers PDF, mais il est très probable qu'à l'avenir il ne sera plus possible de les trouver vu que la recherche en plein texte ne fonctionne pas au travers de ces textes cryptés. Un autre problème réside dans le fait que l'erreur n'apparaît pas au premier coup d'œil.

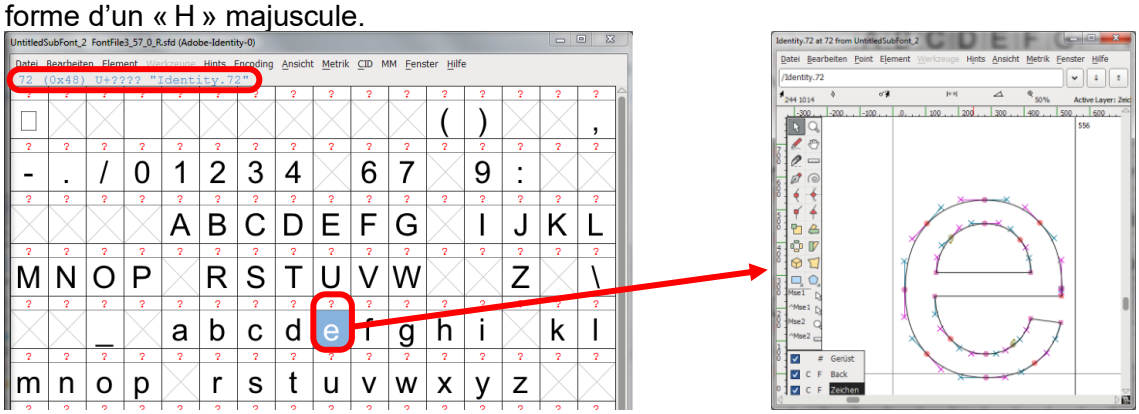
2 Analyse

2.1 Document affecté

Le document présenté ci-après est un PDF/A-2b que non seulement Adobe-Preflight, mais également les validateurs de PDF Tools, Callas et PDFTron, considèrent comme valide. Aucun lecteur testé n'a pu copier du texte à partir du document ou y effectuer une recherche correctement.

Le document utilise uniquement la police de caractère Arial qui est incorporée en sous-groupes.

L'analyse a révélé que les polices incorporées ne garantissaient pas l'option de recherche et d'extraction. L'ensemble du PDF a été exporté vers Word à l'aide d'« Adobe Acrobat Pro X 116 ». Cette opération a montré que chacune des polices Arial incorporées était concernée. La lettre « e » est systématiquement extraite sous la forme d'un « H » majuscule.



Ce phénomène est également visible lors de l'ouverture de la police incorporée avec FontForge. On peut ensuite voir que l'attribution au caractère Unicode n'est pas définie dans la police. Les contours étant toutefois définis, il n'y a pas de problème pour représenter à l'écran ni pour imprimer. La conséquence est que le processus permet la reproduction visuelle et qu'il ne se produit aucune erreur dans la validation du PDF/A-2b. Comme l'attribution aux signes Unicode au moyen d'une CMap « ToUnicode » n'est pas disponible à l'intérieur du fichier PDF, lors de la « copie avec mise en forme » la valeur standard 72 (U+00048 = H) est adoptée et restituée pour le contour du « e ». L'option de recherche et d'extraction n'est possible que pour des PDF/A des niveaux de conformité A et U valides et pas pour le niveau de conformité B. Ce dernier garantit une restitution visuelle impeccable et indépendante de tout système ainsi que l'intégrité des pages. La possibilité d'extraire du texte est en option pour la conformité B.

- La police incorporée Arial ne répond pas aux exigences archivistiques.
- La police ne contrevient pas à l'exigence du PDF/A avec le niveau de conformité B.

2.2 Évaluation et Cfa

Dans son Catalogue des formats de données d'archivage (Cfa)[1], le CECO recommande le PDF/A-1 et le PDF/A-2 en tant que formats de fichiers adaptés pour l'archivage à long terme de documents textuels. Le PDF/A est l'un des formats les plus utilisés dans les archives. Une évaluation du groupe d'experts Preservation planning a eu pour objectif d'enquêter sur la fréquence de l'erreur incriminée dans les fichiers PDF/A-1b et PDF/A-2b au sein des différentes archives.

Comme une telle évaluation était impossible avec les outils existants, le CECO a mandaté l'entreprise PDF Tools pour développer un module additionnel au « 3-Heights™ PDF Validator ». Ce validateur est également intégré à KOST-Val[2].

2.2.1 Résultat de l'évaluation

L'évaluation réalisée avec l'aide de KOST-Val a révélé qu'on trouve de tels documents PDF/A dans toutes les archives[3] et que l'erreur ne se limite pas à certains types de polices ou certaines années de production. Les documents PDF/A de fabricants de logiciels de meilleure qualité sont en moyenne moins touchés que les produits gratuits par exemple. L'erreur a en outre été trouvée dans tous les types de PDF : dans des fichiers PDF/A non valides des niveaux de conformité A ou U, dans des fichiers PDF avec polices intégrées et surtout dans des fichiers PDF/A du niveau de conformité B.

L'outil « 3-Heights™ PDF Validators »[4] réglé avec la configuration « stricte » signale tous les caractères inconnus et non définis qui sont utilisés[5]. Les caractères non définis sont ceux qui contiennent la valeur Unicode U+FFFD (= inconnu). En moyenne, environ 20% de tous les fichiers PDF/A-1b et PDF/A-2b dans les archives membres du CECO qui ont participé à l'évaluation sont concernés.

Ce résultat a eu pour conséquence que le CECO a programmé une nouvelle évaluation dans KOST-Val permettant d'ignorer les documents contenant des défauts moins importants pour la recherche en plein texte et trouver plus rapidement les documents réellement problématiques.

Cette seconde évaluation réglée sur la configuration « tolérante » a été effectuée par KOST-Val. Les défauts suivants des polices ont été ignorés :

---

[1] Catalogue des formats de données d'archivage : https://kost-ceco.ch/cms/formats-de-donnees.html
[2] KOST-Val: https://kost-ceco.ch/cms/kost-val-2.html
[3] Pour cette évaluation, nous avons utilisé, en plus des fichiers du bureau du CECO (collection en provenance de différentes sources et archives), également des fichiers PDF/A archivés par les Archives de l'État de SG, BS et UR ainsi que par les Archives fédérales.
[4] 3-Heights™ PDF Validators: https://www.pdf-tools.com/pdf20/fr/produits/pdf-converter-validation/pdf-validator/
[5] Il est toutefois possible qu'un caractère utilisé ne soit pas visible sur la page si par exemple il est recouvert ou s'il n'y a pas de contraste en raison de la transparence.

1. Tous les défauts dans les polices, lorsque 5 caractères au maximum étaient incorporés.

Von den 4 im Dokument enthaltenen Zeichen sind 4 (100%) unbekannt und 0 (0%) undefiniert.

Font: LuciduxSans-Oblique Full name: LuciduxSans-Oblique (object no 12) Type: Type1 Font file: Type1

P	h	s	u

2. Les défauts dans les polices contenant dans leur nom « Webdings », « Wingdings », « Symbol » ou « Math »

Font: Symbol Full name: ABCDEE+Symbol (object no 12125) Type: Type0 (CIDFontType2) Font file: TrueType

• ∏ ∑ ⎧ ⎨ ⎩ ⎪ ⎫ ⎭

Font: Cambria Math Full name: ABCEEE+Cambria Math (object no 12203) Type: Type0 (CIDFontType2) Font file: TrueType

3. Tous les caractères non définis, dans la mesure où ceux-ci ne dépassent pas le seuil de 20% de tous les caractères

Von den 1436 im Dokument enthaltenen Zeichen sind 0 (0%) unbekannt und 23 (1.60167%) undefiniert.

4. Les caractères connus du CECO qui représentent les espaces, les signes d'énumération et les points de suspension (...)[6]

Font: TT1A0o00 Full name: CRTGOZ+TT1A0o00 (object no 34) Type: Type1 Font file: CFF

–	

5. Lorsqu'à la fin seuls 4 caractères au maximum seraient encore signalés comme erronés[7]

Von den 441 im Dokument enthaltenen Zeichen sind 4 (0.907%) unbekannt.

Font: Calibri Full name: PBMAIO+Calibri (object no 23) Type: Type0 (CIDFontType2) Font file: TrueType

b	c	m	n

Avec cette configuration « tolérante », KOST-Val signale encore en moyenne environ 1% de tous les fichiers PDF/A-1b et PDF/A-2b dans les archives membres du CECO ayant participé à l'évaluation. C'est beaucoup moins qu'avec la configuration « stricte » et ce procédé n'identifie ainsi que les fichiers PDF/A problématiques.

Conclusion :

- ➪ Tous les fichiers PDF avec polices incorporées peuvent présenter ce défaut.

---

[6] En cours d'évaluation, il a été constaté que dans de très nombreux documents seules ces catégories de caractères étaient signalées. Il a été renoncé aux autres caractères étant donné que ces derniers devaient être définis de manière manuelle ou visuelle, ce qui implique des efforts considérables.

[7] Après avoir terminé les étapes 1 à 4 de l'évaluation, il a été constaté que dans un quart de tous les défauts restants seuls quatre caractères ou moins sont signalés. Dans le pire des cas, il s'agit de lettres ou de nombres. Il est toutefois très improbable qu'une recherche en plein texte échoue en raison de ces quatre caractères. Il s'agit la plupart du temps de caractères qui pourraient être attribués aux catégories 2 et 4.

- Si elles ne prennent aucune mesure, les archives doivent s'attendre à ce que 1% de tous les fichiers PDF/A du niveau de conformité B ne puissent être trouvés au moyen d'une recherche en plein texte.

3 Recommandations

3.1 Pour éviter le défaut

Si une institution d'archives souhaite à l'avenir ne posséder que des fichiers PDF/A avec option de recherche et d'extraction, elle doit, pour éviter le défaut incriminé, exiger au moins le niveau de conformité U lors de la création de nouveaux fichiers PDF/A. C'est ce qu'indique déjà le Cfa dans sa conclusion.

- Éviter : convertir au moyen de logiciels de qualité élevée et exiger au moins le PDF/A-2u.

3.2 Pour identifier le défaut

Selon l'évaluation, chaque institution d'archivage peut identifier les fichiers PDF et PDF/A existants dans ses fonds qui ne disposent pas d'option complète de recherche et d'extraction.

Une mesure de conservation minimale consiste à signaler ces fichiers par une notice dans l'instrument de recherche[8].

- Identifier : signaler dans l'instrument de recherche tous les fichiers qui n'ont pas passé la validation dans le module K avec la configuration « stricte ».

3.3 Pour réparer le défaut

Les fichiers PDF et PDF/A existants dans les archives qui n'ont pas passé la validation dans le module K avec la configuration « tolérante » risquent d'être introuvables lors d'une recherche en plein texte.

Ces fichiers devraient non seulement être identifiés, mais également faire l'objet d'une réparation du défaut qui consiste à compléter chaque caractère manquant par le caractère Unicode correspondant.

- Réparer : prendre une mesure de conservation sur tous les fichiers qui n'ont pas passé la validation dans le module K avec la configuration « tolérante ».

Le défaut incriminé présente deux types d'erreurs qu'il faut distinguer :

---

[8] Les AFS ne procèdent pas à l'identification ni à la réparation préalable des fichiers PDF existants. En contrepartie, elles convertissent systématiquement lors du versement tous les fichiers PDF commandés en un nouveau PDF/A-2u avec un outil de reconnaissance optique de caractères (OCR-ROC). Il est ainsi possible de rechercher et d'extraire tous les fichiers PDF versés. Si les AFS introduisaient ultérieurement la recherche en plein texte via le magasin numérique, elles risqueraient, si elles ne prenaient pas d'autres mesures, de voir de nombreux fichiers devenir introuvables pour la recherche en plein texte.

a) Les modules D et K sont signalés comme défectueux

D) Schriften	Das Dokument enthaelt Schriften ohne eingebettete Font-Programme oder Codierungsinformation (CMAPs). [PDF Tools: iCategory_8]
K) Schrift-Validierung	Von den 27 im Dokument enthaltenen Zeichen sind 23 (85.1852%) unbekannt und 0 (0%) undefiniert. **Font: Calibri** Full name: Calibri (object no 65) Type: Type0 (CIDFontType2) T a b c d e g h i l m n o p r s t u w . “ ”

b) Le module K est signalé comme défectueux, mais pas le module D

K) Schrift-Validierung	Von den 971 im Dokument enthaltenen Zeichen sind 14 (1.44181%) unbekannt und 0 (0%) undefiniert. **Font: StoneSans-Semibold** Full name: JLPQRJ+StoneSans-Semibold (object no 397) Type: Type0 (CIDFontType0) Font file: CFF R T a b d e l n o r t u x

3.3.1 Réparation du défaut : les modules D et K sont signalés comme défectueux

Dans le cas a), la réparation du défaut s’effectue la plupart du temps de manière automatique si le fichier PDF est converti en PDF/A-2u au moyen d’un outil professionnel. Durant ce processus, les polices incorporées non valides sont incorporées si possible correctement et complétées par leur valeur Unicode. Cela fonctionne de façon très fiable si la police est également installée sur l’ordinateur ou qu’elle est une police standard. Le cas a) se produit dans moins d’un tiers des cas.

3.3.2 Réparation du défaut : Le module K est signalé comme défectueux, mais pas le module D

Dans le cas b), les polices sont certes incorporées, mais seulement leur contour comme c’est le cas dans l’incident décrit au début. L’évaluation a montré que c’était le cas le plus courant. La façon la plus utile de remédier à ce problème est d’évaluer le contour.

Parallèlement à l’extension de son outil 3-Heights™ PDF Validator, l’entreprise PDF Tools a également adapté 3-Heights™ PDF OCR[9]. Les caractères signalés comme défectueux dans la table des caractères sont lus au moyen du moteur OCR ABBYY FineReader et le résultat est indiqué dans la table des caractères comme caractère Unicode à côté le caractère concerné. La conformité du PDF/A est bien entendu conservée. La reconnaissance OCR est principalement spécialisée sur les nombres et lettres ; la reconnaissance des autres caractères est moins précise.

La reconnaissance OCR peut être nettement améliorée en adaptant les options ou les paramètres à l’institution d’archivage et au document.

Le CECO a utilisé la requête suivante pour l’échantillonnage test :

```
pdfocr -v -ocr "service" -ocp "Profile=C:\Dateipfad\abbyy_text.txt"
-ocl "English,German,French,Mathematical" -otm update -ots
"knownSymbolic" -otu "installedFont,fallbackAllPua" input.pdf
output.pdf
```

[9] 3-Heights™ PDF OCR: https://www.pdf-tools.com/pdf20/fr/produits/pdf-converter-validation/pdf-ocr/

Le profil «abbyy_text.txt », contient les informations suivantes permettant de reconnaître des textes orientés différemment. La reconnaissance du texte contenu dans des images est aussi possible :

```
[PagePreprocessingParams]
CorrectOrientation=TRUE

[PageAnalysisParams]
DetectVerticalEuropeanText=TRUE

[ObjectsExtractionParams]
DetectTextOnPictures = TRUE
```

L'anglais, l'allemand, le français et les symboles mathématiques ont été définis comme langue ou type principal (-ocl). C'est utile étant donné que le programme OCR travaille avec des listes de mots de chacune des langues et atteint ainsi une qualité plus élevée et une meilleure performance.

Tous les caractères sans valeur Unicode significative sont traités avec « -otm update ». Si on spécifie en même temps aussi «-ots "knownSymbolic"» la reconnaissance de texte OCR ignorera cependant les polices de symboles classiques telles que « ZapfDingbats » et « Wingdings ».

En utilisant « -otu "installedFont,fallbackAllPua" », les polices installées sur le système et leur attribution Unicode sont utilisées et avec fallbackAllPua, les caractères pour lesquels aucune valeur Unicode appropriée n'a pu être déterminée sont déclarés comme non définis. Cela permet à un plus grand nombre de documents de passer la validation de polices tolérante.

Au cours du test effectué au bureau du CECO, 66% des documents défectueux ont pu être réparés avec 3-Heights™ PDF OCR. Par ailleurs, le nombre de caractères signalés comme défectueux a été réduit de 87%. L'outil n'a pas été en mesure de reconnaître les 13% restants de caractères défectueux qui ont été déclarés comme non définis via l'option « fallbackAllPua ». Cette opération a permis de réparer tous les documents du point de vue de la validation de polices tolérante. Ces chiffres dépendent cependant fortement des documents primaires.

Original Screenshot	Extraction originale	Extraction après réparation
© Photo-montage	□□□□□□□□□□□□□□□□□□	© Photo-montage
$ax^2 + bx + c = 0$	__ + __ + _ = 0	ax2 + bx + c = 0
newsletter	newsle□er	newsletter

Une utilisation efficace de 3-Heights™ PDF OCR permet d'espérer que l'outil va continuer d'être amélioré, ceci à la condition que des documents qui n'ont pas pu être réparés de manière satisfaisante aient pu être mis continuellement à la disposition du fabricant.

3.3.3 Contrôle de la réparation des défauts

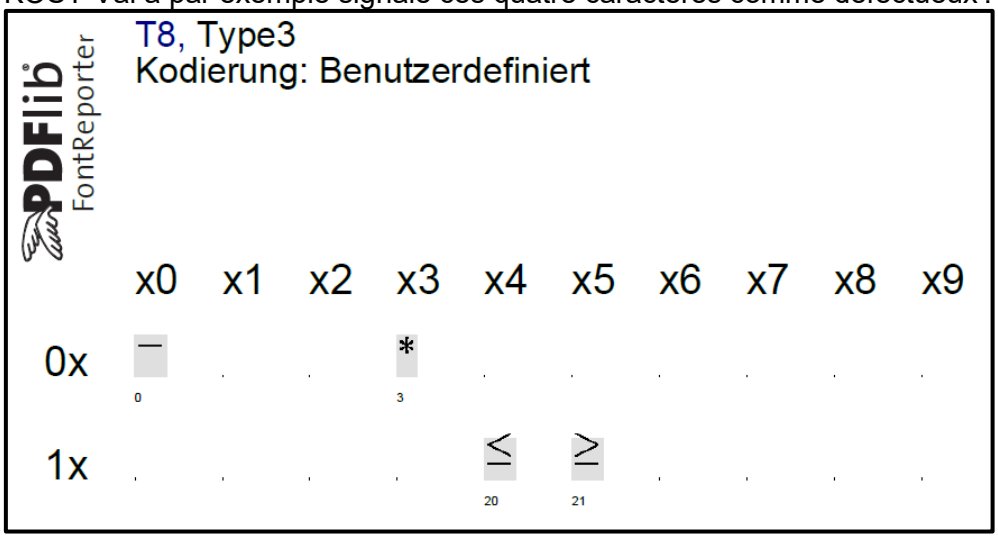
Il faut procéder à une nouvelle validation avec KOST-Val après avoir effectué la réparation. Dans l'idéal, le problème est réparé. La qualité du caractère corrigé ou reconnu ne peut être contrôlée que de manière visuelle. Comme une telle opération nécessite beaucoup d'efforts, il n'est donc pas possible de contrôler beaucoup de documents. Le bureau du CECO a utilisé pour ce contrôle PDFlib FontReporter[10] en plus de KOST-Val. PDFlib FontReporter est un plug-in gratuit pour Adope Acrobate destiné à analyser les polices dans les documents PDF. Sur un simple clic, PDFlib FontReporter génère un rapport détaillé sur un fichier PDF ouvert dans Acrobat. PDFlib FontReporter enregistre des données générales, des informations sur les polices ainsi que des tables glyphes sur toutes les polices disponibles dans un PDF et crée un PDF séparé accompagné d'un rapport.

Pour le contrôle visuel de la qualité, les valeurs Unicode issues du rapport ont été lues et contrôlées chez les caractères signalés défectueux par KOST-Val. On peut alors consulter la valeur Unicode dans un tableau[11] ou elles sont mises dans Word et converties au moyen de la combinaison de touches [Alt] [C].

3.3.4 Problèmes lors de la réparation des défauts

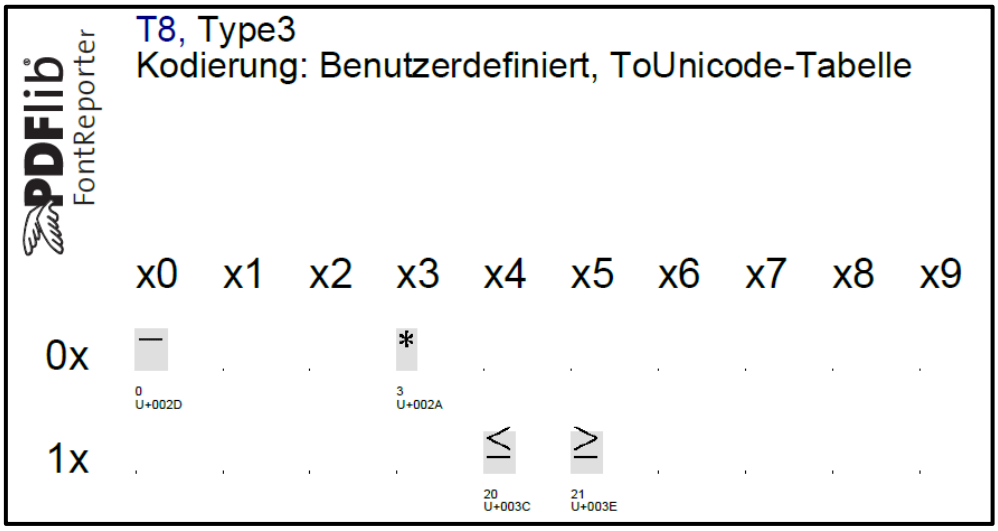
Le contrôle de la qualité a révélé en particulier des problèmes avec des polices définies par l'utilisateur et des caractères qui ne représentent pas les lettres et chiffres latins.

KOST-Val a par exemple signalé ces quatre caractères comme défectueux :



Après réparation, le nouvel encodage est devenu « tableau ToUnicode défini par l'utilisateur » et les caractères ont été complétés par des valeurs Unicode :

[10] PDFlib FontReporter: https://www.pdflib.com/download/free-software/fontreporter/
[11] Par ex. : https://www.compart.com/fr/unicode/

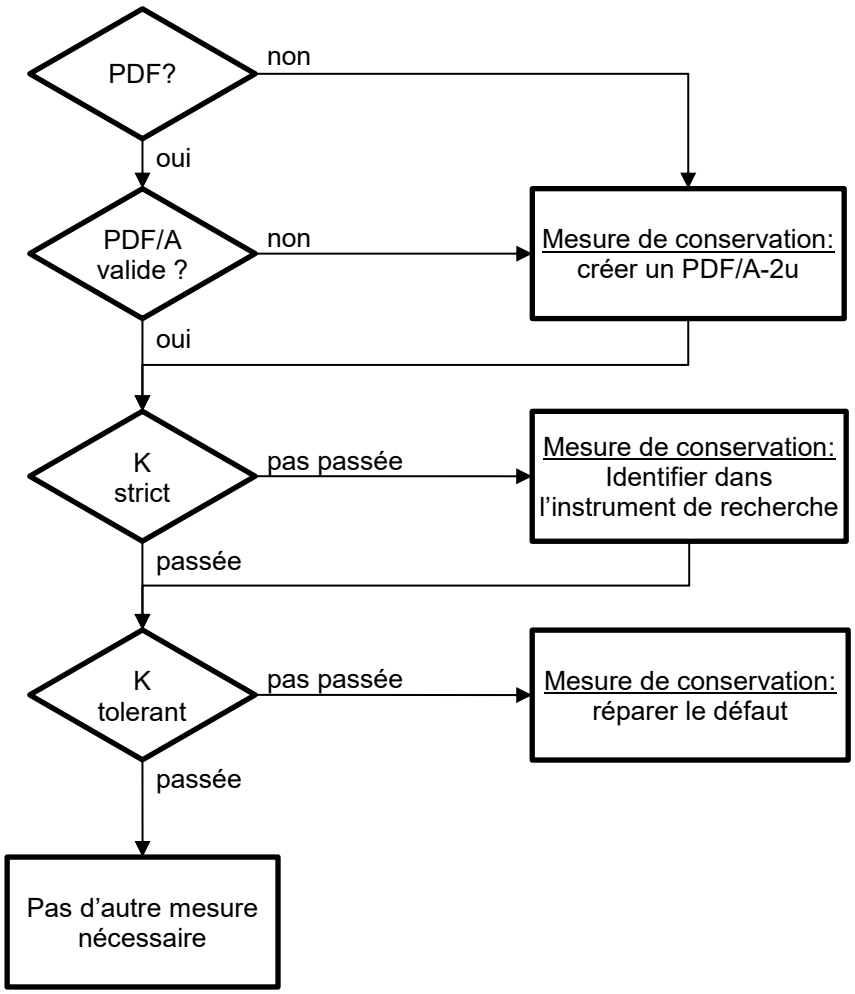


Si maintenant, on contrôle chaque valeur Unicode, il s'avère que la reconnaissance ne marche pas à 100%.

	Problèmes lors de la réparation des défauts			
Représentation KOST-Val	—	*	<	>
Représentation FontReporter	— 0 U+002D	* 3 U+002A	≤ 20 U+003C	≥ 21 U+003E
Passage de texte	range -12	$V * 10^{-N}$	$\leq p \leq n.$	If $N \geq 0,$
Valeur Unicode reconnue	U+002D	U+002A	U+003C	U+003E
Caractère Unicode reconnu	-	*	<	>
Problème lors de la reconnaissance	aucun	L'outil a reconnu une étoile à 5 branches qui a été mise en exposant	Le souligné n'a pas été reconnu par le moteur OCR (ABBYY FineReader), étant donné que, dans cette police, il se trouve en dessous de la ligne de base de l'écriture, chose inhabituelle pour ce caractère	
Valeur Unicode correcte		U+2217 ∗	U+2264 ≤	U+2265 ≥

Recommandations : résumé
Les archives devraient à l'avenir éviter le défaut décrit en exigeant au moins du PDF/A-2u, en identifiant les fichiers existants qui ne passent pas la validation « K stricte » et en corrigeant en plus les fichiers qui n'ont pas passé la validation « K tolérante ».

Flux de tâches : créer des PDF/A avec option de recherche et d'extraction



- Éviter le défaut : convertir avec un logiciel de qualité élevée et exiger au moins le PDF/A-2u.
- Identifier le défaut : identifier dans l'instrument de recherche tous les fichiers qui n'ont pas passé la validation dans le module K avec la configuration stricte.
- Réparer le défaut : prendre des mesures de conservation sur tous les fichiers qui n'ont pas passé la validation dans le module K avec la configuration « tolérante ».