CONSERVATION PDF HYBRIDE SUR PAPIER

Combiner numérique et analogique pour conserver des documents critiques pendant des siècles dans un contexte de gestion des déchets radioactifs

Vincent Joguin
Eupalia
France
vincent.joguin@eupalia.com
https://orcid.org/0000-0003-0627-8778

Florence Poidevin
Andra
France
Florence.Poidevin@andra.fr
https://orcid.org/0009-0001-7920-4107

Résumé **– De nombreuses organisations conservent leur contenu informationnel au format PDF. L'Andra, l'Agence pour la gestion des déchets radioactifs, documente ses centres de stockage de déchets nucléaires principalement sous ce format, avec l'obligation de conserver une partie de cette documentation sur des échelles de temps de plusieurs siècles au moins. L'impermanence de l'informatique et la complexité du format PDF rendent ce dernier inadapté à de telles échelles de temps, et l'Andra a donc choisi d'imprimer ces documents sur du papier permanent. Cependant, ce processus induit non seulement un nombre élevé de pages, mais aussi une perte d'intégrité numérique qui compromet le traitement automatisé du texte (par exemple la recherche, la traduction, le tri) car il devient dépendant de la disponibilité hypothétique future d'une océrisation parfaite. Pour tenter de conjuguer le meilleur du numérique et de l'analogique, la solution Micr'Olonys d'Eupalia, déjà testée à l'Andra sur une base de données, a été étendue avec un utilitaire de préparation, appelé Sumetar, pour convertir les fichiers PDF en fichiers texte simples encodés en UTF-8, associés à des images au format BMP ou imprimées sous forme analogique. Le format texte brut et le format BMP sont tous deux très simples et en même temps très répandus, et sont donc adaptés à une accessibilité à long terme comme actuelle. Sumetar regroupe plusieurs fichiers en un seul fichier tar non compressé, un autre format très simple. Micr'Olonys transcrit ensuite le fichier numérique en codes-barres bidimensionnels imprimés sur papier pour une conservation inerte. Des tests réalisés fin 2023 montrent que cette stratégie permet généralement de diviser par quatre le nombre de pages d'un document contenant environ une image toutes les deux pages en moyenne.**

Mots-clés **– PDF, conservation inerte, conservation hybride, papier permanent, permacomputing, gestion des déchets radioactifs**

I. INTRODUCTION

Le programme « Mémoire pour les générations futures » lancé par l'Andra en 2010 a trois objectifs principaux : éviter le plus longtemps possible le risque d'intrusion humaine involontaire dans les centres de stockage de déchets nucléaires, faciliter les décisions des générations futures et transmettre un patrimoine technique et culturel.

Ces trois objectifs complémentaires impliquent de comprendre comment la mémoire et la connaissance peuvent être transmises à différentes échelles de temps, à différentes catégories de personnes, et de construire un système robuste pour conserver les données associées.

Pour répondre à ces enjeux, le programme Mémoire s'articule autour de quatre piliers : les archives et la documentation réglementaires, les interactions sociétales, les études et la recherche, et la coopération internationale, dans une approche multimodale et pluridisciplinaire.

iPRES 2024 : 20e Conférence internationale sur la préservation numérique, Gand et Flandre (Belgique).

L'un des défis auxquels l'Andra doit faire face est celui de combiner des matériaux durables, à même de se conserver au moins pendant plusieurs siècles, avec une énorme quantité de données, qui sont censées non seulement être transmises au successeur de l'Andra (ou plus largement aux générations suivantes), mais aussi triées et comprises avec suffisamment de facilité par eux.

Par exemple, conformément à la loi française (plus précisément au « Code de l'environnement »), l'Andra élabore un « Dossier détaillé de mémoire » (DDM) pour chacun de ses centres de stockage de déchets nucléaires. À terme, ces DDM contiendront des dizaines de milliers de documents techniques décrivant les installations, les colis de déchets, leurs propriétés physico-chimiques et radiologiques, etc.

Cependant, les matériaux les plus résistants sont souvent les moins adaptés pour condenser de grandes quantités d'informations.

En l'espèce, les DDM sont actuellement imprimés sur du papier permanent, qui offre plusieurs qualités (dont une durabilité à long terme dans des conditions d'archivage) mais n'est bien sûr pas aussi efficace pour condenser l'information que les supports numériques magnétiques, électroniques et optiques.

C'est pourquoi l'Andra teste, depuis quelques années, d'autres systèmes innovants, dans une démarche prospective, parmi lesquels la solution Micr'Olonys d'Eupalia.

La première phase d'essais, en 2020-2021, a consisté à transcrire une base de données numérique en codes-barres bidimensionnels (appelés « emblèmes ») imprimés sur du papier permanent, produisant un document de 464 pages au lieu du million de pages qu'aurait nécessité l'impression sous forme de texte en clair [1].

La deuxième phase d'essais, en 2022-2023, décrite dans cet article, visait à convertir des fichiers PDF, comprenant dessins, tableaux et graphiques, en une forme plus simple adaptée à la conservation à long terme sous forme de documents Micr'Olonys sur papier permanent.

II. Papier permanent

L'Andra a adopté le papier permanent comme support de référence pour la conservation à long terme des archives (comme le montre la figure 1), après avoir pris acte qu'aucun contrat de maintenance avec un fournisseur de solutions électroniques ne pourrait prétendre s'appliquer sur une échelle de plusieurs siècles. À l'inverse, de nombreux supports traditionnels (par exemple la pierre, la céramique, le papier) peuvent être durables et ne nécessitent pas d'entretien. Le papier a l'avantage d'être peu onéreux et compatible avec les imprimantes de bureau utilisées couramment, mais, selon sa composition, il peut se détériorer assez rapidement ou durer des siècles. C'est pourquoi l'Andra a choisi le « papier permanent ».

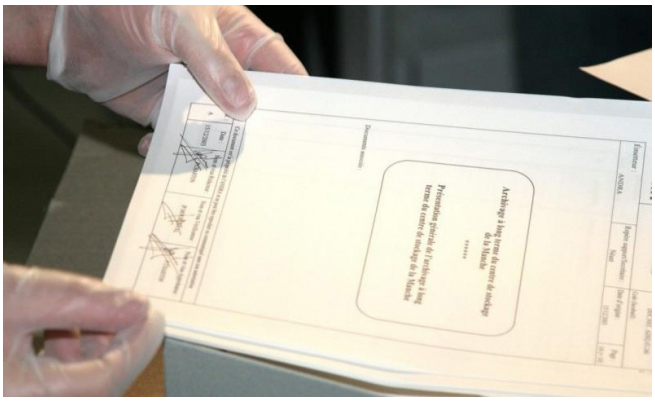

Figure 1 Boîte d'archives et document Andra imprimé sur papier permanent.

Un papier est considéré comme « permanent » s'il répond aux exigences de la norme internationale pour le papier permanent (ISO 9706). Créée en 1994, elle avait pour objectif de répondre à la dégradation très préoccupante des papiers produits au cours du XX^{e} siècle, accélérée par leur composition. « Permanent » signifie qu'il est capable de rester chimiquement et physiquement stable sur une longue période. Les critères suivants sont définis :

- le pH de l'extrait aqueux de la pulpe doit être compris entre 7,5 et 10 ;

- l'indice Kappa de la pâte, qui indique la résistance à l'oxydation (liée à la présence de lignine), doit être inférieur à 5 ; la teneur en substances oxydables, principalement en lignine, est inférieure à 1% ;

- la réserve alcaline doit être supérieure ou égale à 2 % en équivalent carbonate de calcium ;

- la résistance à la déchirure doit être supérieure à 350 mN pour un papier d'un grammage supérieur à 70 g/m^2.

Afin d'établir un critère de fin de vie et d'évaluer la durée de vie des papiers permanents, avec et sans encre, dans le cadre de la conservation archivistique, l'Andra a collaboré à une recherche doctorale de 2020 à 2023. Cette thèse a été menée par Caroline Vibert et encadrée par deux laboratoires très impliqués dans l'étude des phénomènes de dégradation : le PIMM (Procédés et Ingénierie en Mécanique et Matériaux, Paris), qui dispose d'une infrastructure pour l'étude des matériaux polymères, et notamment de leur résistance au vieillissement et aux contraintes, et le CRCC (Centre de Recherche et de Conservation des Collections, Paris), expert des matériaux du patrimoine, dont les papiers anciens [2][3].

Des expériences de vieillissement artificiel à 90°C ont été menées pour accélérer la dégradation du papier dans diverses conditions d'exposition favorisant soit l'hydrolyse, soit l'oxydation de la cellulose. La compréhension des différents mécanismes a conduit au développement d'un modèle cinétique complet de dégradation, visant à extrapoler la dégradation chimique de la cellulose à température ambiante dans diverses conditions d'exposition. La durabilité du papier permanent à 20°C a été estimée à cinq mille ans au moins.

De plus, une étude exploratoire sur une encre toner actuellement utilisée à l'Andra a montré que sa dégradation est plus lente que celle du papier, n'étant ainsi pas un facteur limitant de la durée de vie du papier imprimé.

III. Transcription de données numériques sur papier avec Micr'Olonys

La solution Micr'Olonys d'Eupalia est conçue pour transcrire des données numériques sur papier (ou microfilm) en vue d'une conservation inerte et d'une accessibilité à long terme. Elle étend ainsi le champ d'application du papier permanent aux contenus numériques.

Plus précisément, Micr'Olonys a été conçue en tant que technologie de logiciel-sur-papier pour conserver du contenu numérique sous forme d'emblèmes (codes-barres bidimensionnels) avec des moyens d'accès insensibles à l'obsolescence. Elle imprime une amorce autonome simple en conjonction avec des emblèmes contenant un fichier numérique. L'amorce Micr'Olonys fournit aux utilisateurs du futur tous les moyens pour restituer le fichier sans recourir à une quelconque technologie spécifique telle qu'un matériel, un système d'exploitation ou un langage de programmation particuliers. Dans son approche, Micr'Olonys est en adéquation avec de nombreux principes du concept récemment défini de permacomputing [4], qui lui-même recoupe la conservation numérique[1].

Pour plus de détails, en plus de [1], [5] fournit un aperçu global de l'approche de conservation numérique inerte et de Micr'Olonys, tandis que [6] présente plus en détail la technologie Micr'Olonys. [7] décrit intégralement l'amorce Micr'Olonys.

IV. Extraction du contenu de fichiers PDF

Pour permettre l'impression de documents initialement au format PDF sur papier permanent avec la solution Micr'Olonys, celle-ci a été étendue avec un utilitaire de préparation, appelé Sumetar, conçu pour remettre en forme le contenu extrait des fichiers PDF.

L'extraction de texte et d'images à partir de fichiers PDF s'effectue généralement à l'aide d'une bibliothèque logicielle dédiée, car le format PDF n'est pas simple à manipuler directement. Quelques bibliothèques de ce type existent pour le langage de programmation C# dans lequel Micr'Olonys et Sumetar sont développés, notamment iText PDF[2], ByteScout PDF Extractor SDK[3], Snowtide PDFxStream[4] et Bit Miracle Docotic.Pdf[5]. Cette dernière a été sélectionnée car elle offre une bonne extraction du texte, formaté de façon à s'approcher de la mise en page d'origine des pages PDF.

A. Gérer le texte

Docotic.Pdf extrait tout le texte d'une page spécifiée d'un document PDF sous forme de chaîne C# avec un seul appel à la fonction GetTextWithFormatting(). En interne, les chaînes de caractères en C# sont encodées en UTF-16, la variante 16 bits de l'encodage de caractères Unicode. C# permet une conversion facile de cet encodage vers la variante 8 bits, UTF-8, qui est plus largement en usage (par exemple utilisée par 98,3 % de tous les sites Web[6]) et adaptée à la

[1] permacomputing.net/digital_preservation
[2] itextpdf.com
[3] bytescout.com/products/developer/pdfextractorsdk/
[4] www.snowtide.com
[5] bitmiracle.com/bibliotheque-pdf
[6] w3techs.com/technologies/cross/character_encoding/ranking

langue française ainsi qu'aux langues occidentales plus généralement.

Le texte se compresse très bien. Une fois compressé numériquement et emblémisé (c'est-à-dire converti en codes-barres bidimensionnels Micr'Olonys), il occupe typiquement près de cent fois moins de pages que sous forme directement lisible.

B. Gérer les tableaux

Les tableaux n'existent généralement pas en tant que tels dans les fichiers PDF. Il s'agit d'une combinaison de caractères et de lignes sans relation explicite entre eux.

Comme Docotic.Pdf ne tient pas compte des lignes lors de la génération de texte formaté, la lisibilité des tableaux était parfois dégradée une fois ceux-ci convertis en texte brut. Un algorithme spécifique a donc été développé pour tenter de reproduire les lignes horizontales des tableaux sous forme de texte à l'aide de chaînes de caractères « ─ ». Bien que le procédé de reproduction de la mise en page originale du tableau ne soit pas exempt d'erreurs, le résultat recrée généralement suffisamment de structure pour interpréter correctement les limites des cellules, comme l'illustrent les figures 2 et 3.

Tableau 3.1-4 Situations accidentelles pour l'installation souterraine

N°	Local / zone	Scénarios
A14	Gare basse transfert incliné	Collision du transfert incliné en gare basse à petite vitesse
A15	Galeries souterraines	Incendie du chariot impliquant une hotte
A16	Cellule de manutention MA-VL	Chute d'un colis de stockage suite à une défaillance de l'élévateur
A17		Incendie en cellule de manutention
A18	Alvéole de stockage MA-VL	Incendie du pont stockeur impliquant un colis de stockage
A19	Alvéole de stockage HA	Incendie du vérin-pousseur (+ chemisage acier + colis de stockage HA) en alvéole HA

Figure 2 Tableau original dans un document PDF.

```
Tableau 3.1-4          Situations accidentelles pour l'installation souterraine
N°          Local / zone                              Scénarios
A14         Gare basse transfert incliné              Collision du transfert incliné en gare basse à petite vitesse
A15         Galeries souterraines                     Incendie du chariot impliquant une hotte
A16         Cellule de manutention MA-VL              Chute d'un colis de stockage suite à une défaillance de l'élévateur
A17                                                   Incendie en cellule de manutention
A18         Alvéole de stockage MA-VL                 Incendie du pont stockeur impliquant un colis de stockage
A19         Alvéole de stockage HA                    Incendie du vérin-pousseur (+ chemisage acier + colis de stockage
                                                      HA) en alvéole HA
```

Figure 3 Tableau converti en texte brut avec des lignes horizontales insérées.

L'algorithme détecte les lignes noires horizontales continues au-dessus de chaque caractère et, le cas échéant, insère le caractère Unicode « BOX DRAWINGS LIGHT HORIZONTAL » (U+2500) à l'endroit approprié sur une ligne existante ou sur une nouvelle ligne en fonction de la position relative physique de la ligne noire par rapport au texte sur la page.

C. Gérer les images

Dans la plupart des documents Andra, les images (diagrammes, photographies, plans, croquis, tableaux complexes, etc.) représentent la majeure partie de la taille du fichier PDF. Il était donc essentiel de reproduire convenablement les images de façon à ce qu'elles occupent le moins d'espace possible pour réduire le nombre de pages imprimées par document. Les images peuvent être reproduites sous forme de fichiers d'images numériques ou imprimées en clair sous forme analogique. Il était initialement prévu de conserver les images « simples » (comme les diagrammes) sous forme de fichiers numériques, tandis que les images complexes, principalement des photographies en couleur, seraient imprimées en clair.

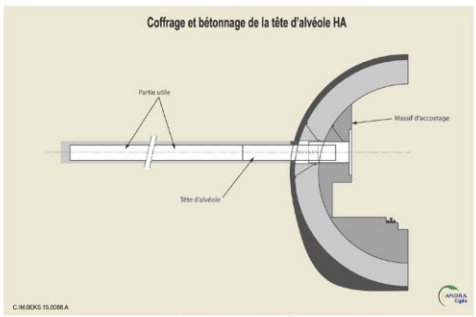

Image n° 166 à la page 243

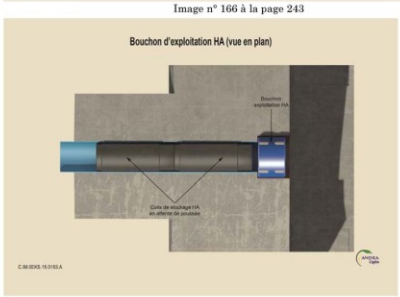

Image n° 167 à la page 243

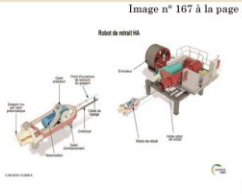

Image n° 168 à la page 244

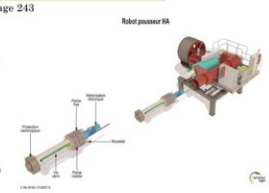

Image n° 169 à la page 244

Figure 4 Images du document original imprimées en clair dans l'addenda.

Le format de fichier d'image BMP non compressé, très répandu, a été choisi pour sa grande simplicité, et les fichiers auraient été compressés de manière transparente par la fonctionnalité de compression intégrée de Micr'Olonys. Cependant, en comparant la taille de la surface physique qu'un fichier BMP compressé occuperait sous forme d'emblèmes avec la taille de la surface physique de l'image analogique, cette dernière était presque toujours nettement plus petite, sauf lorsque l'image n'était essentiellement qu'une seule couleur unie. Il a donc été décidé de regrouper toutes les images sous forme analogique en tant qu'addenda au document numérique Micr'Olonys principal (voir la figure 4 pour un exemple).

D. Gérer plusieurs fichiers

Pour que l'amorce Micr'Olonys reste aussi simple et générique que possible, un document numérique Micr'Olonys ne stocke qu'un seul flux d'octets. Lorsque plusieurs fichiers doivent être stockés conjointement (par exemple lorsqu'un fichier de métadonnées décrit le document d'archives), ils doivent être représentés au sein d'un seul flux d'octets. À cette fin, le format tar original d'Unix non compressé a été retenu. Avec ce format très simple et en même temps très répandu, les noms de fichiers sont limités à 100 octets, ce qui est suffisant pour les noms des documents.

V. IFHM : Interface Futur-Homme-Machine

L'amorce Micr'Olonys est totalement neutre quant à la manière d'interpréter le contenu numérique conservé, restitué à l'utilisateur sous forme d'une série d'octets. Cependant, l'amorce fait référence à une partie liminaire directement lisible qui a vocation à fournir toutes les informations nécessaires pour contextualiser le document numérique. Cette partie liminaire joue le rôle d'une « interface futur-homme-machine » et ne doit évidemment pas présupposer dans sa rédaction que les technologies actuelles, matérielles ou logicielles, seront disponibles au moment de la restitution.

La conversion d'un document PDF se traduit par un fichier texte encodé en UTF-8 et éventuellement en fichiers BMP, tous contenus dans un seul fichier tar. De plus, les images sont référencées dans le fichier texte, quel que soit leur mode de reproduction sous forme numérique et/ou analogique, à l'aide de balises HTML. Par conséquent, la partie liminaire doit inclure une table d'encodage UTF-8, et les spécifications des formats BMP et tar ainsi que des balises HTML. Mais seules les informations utiles à la réinterprétation du contenu numérique conservé sont nécessaires, et toute information superflue ne ferait qu'augmenter le risque de confusion pour l'utilisateur du futur.

La partie liminaire est ainsi générée en ne listant que les caractères effectivement utilisés dans le ou les fichiers numériques et noms de fichier du document, parmi les 149 813 caractères (sans compter les groupes de graphèmes constitués de séquences de caractères) que définit la norme Unicode. La figure 5 illustre les tables de caractères générées avec un exemple de document Andra.

Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.	Sym.	Car.
34	"	43	+	50	2	57	9	66	B	73	I	80	P	87	W	97	a	104	h	111	o	118	v	126	~
37	%	44	,	51	3	58	:	67	C	74	J	81	Q	88	X	98	b	105	i	112	p	119	w		
38	&	45	-	52	4	59	;	68	D	75	K	82	R	89	Y	99	c	106	j	113	q	120	x		
39	'	46	.	53	5	60	<	69	E	76	L	83	S	90	Z	100	d	107	k	114	r	121	y		
40	(	47	/	54	6	61	=	70	F	77	M	84	T	91	[	101	e	108	l	115	s	122	z		
41	)	48	0	55	7	62	>	71	G	78	N	85	U	93	]	102	f	109	m	116	t	123	{		
42	*	49	1	56	8	65	A	72	H	79	O	86	V	95	_	103	g	110	n	117	u	125	}		

Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.
194:167	§	194:177	±	195:128	À	195:155	Û	195:168	è	195:174	î	195:185	ù	206:152	Θ	206:188	μ
194:171	«	194:178	²	195:137	É	195:160	à	195:169	é	195:175	ï	195:187	û	206:177	α	207:131	σ
194:174	®	194:181	µ	195:151	×	195:162	â	195:170	ê	195:180	ô	195:188	ü	206:178	β		
194:176	°	194:187	»	195:152	Ø	195:167	ç	195:171	ë	195:182	ö	197:147	œ	206:179	γ		

Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.	Symbole	Car.
226:128:144	‐	226:128:153	’	226:136:134	∧	239:129:161	↵	239:172:129	fi
226:128:147	–	226:128:162	•	226:137:164	≤	239:130:167	▪		
226:128:152	‘	226:128:166	…	226:148:128	─	239:131:188	✓		

Figure 5 Tableaux répertoriant l'encodage de tous les caractères UTF-8 utilisés dans un exemple de document.

La partie liminaire se poursuit par un paragraphe de 4 lignes qui précise la balise HTML utilisée pour référencer les images dans le fichier texte, à savoir « <img src="Image#.bmp" alt="Image#"> » pour les images numériques et « <img src="" alt="Image#"> » pour les images analogiques. Il mentionne également la balise « <pre> » insérée en tout début du fichier texte pour préserver les espaces et sauts de ligne lors de son affichage dans un navigateur Web.

Si le document converti comprend des fichiers d'images numériques, la spécification du format BMP suit. Il utilise le format d'en-tête le plus simple de 26 octets qui inclut le champ BITMAPCOREHEADER pour définir principalement la largeur et la hauteur de l'image. Les pixels sont des valeurs fixes de 24 bits représentant les niveaux de bleu, vert et rouge, dans cet ordre, chacun encodé sur 8 bits.

La spécification du format tar original d'Unix conclut éventuellement la partie liminaire. Si les spécifications des balises HTML ainsi que des formats BMP et tar sont toutes présentes, elles tiennent globalement sur une seule page A4, augmentant ainsi très faiblement la complexité du processus technique qu'un utilisateur du futur devra appréhender pour accéder aux documents numériques.

VI. Structure des documents d'archives

Un document papier généré à partir de la conversion d'un PDF par Sumetar est généralement structuré en trois parties, chacune avec sa propre pagination : la partie liminaire en chiffres romains minuscules, le document numérique (y compris l'amorce Micr'Olonys) en chiffres arabes et l'addenda en chiffres arabes préfixés par « A- ». Cette structure est destinée à guider l'utilisateur du futur dans la restitution du document sous une forme numérique, analogique ou hybride pleinement intelligible. La partie liminaire fournit, sous forme directement lisible, les informations nécessaires à la réinterprétation correcte du document numérique Micr'Olonys (par exemple l'encodage et les spécifications de formats de fichier, les métadonnées orientées utilisateur, les informations contextuelles plus générales) tandis que l'addenda complète le document numérique avec des éléments analogiques en clair qui peuvent être numérisés au moment de la restitution, en l'espèce les images du fichier PDF d'origine.

Sumetar prend également en charge la conversion par lots de documents considérés comme liés entre eux. Cette fonctionnalité de classeur de documents pagine linéairement tous les documents contenus en sus de leur pagination interne, avec une mention « Folio » suivie du numéro de page global du classeur à partir de 1.

Dans le classeur de documents ainsi créé, l'amorce Micr'Olonys et les sections génériques de la partie liminaire ne sont imprimées qu'une seule fois.

VII. Expérience acquise, travaux futurs et annexes

A. Expérience acquise

En traitant séparément le texte et les images des fichiers PDF, il a été possible de mettre en évidence les propriétés de chaque type d'information relativement à sa reproduction sur papier.

L'impression d'images sous forme analogique présente de nombreux avantages par rapport à l'impression sous forme d'emblèmes Micr'Olonys :

1) Taille : les images occupent presque toujours significativement moins d'espace physique lorsqu'elles sont imprimées en analogique plutôt qu'en tant qu'emblèmes, car chaque bit de donnée numérique occupe 54 pixels d'un emblème (à 600 ppp), ce qui est difficile à compenser, même avec une compression numérique avec perte.

2) Accès : immédiatement, les images imprimées sont discernables et donnent une idée du contenu du document.

3) Simplicité : l'abandon total des images numériques réduit la complexité technique.

L'impression de texte sous forme d'emblèmes présente de nombreux avantages par rapport à l'impression en clair :

1) Taille : le texte compressé et imprimé sous forme d'emblèmes occupe généralement environ 100 fois moins de pages que le texte en clair. La taille du texte numérique est tellement négligeable par rapport à celle du texte en clair que l'adjonction des quelques pages supplémentaires d'une impression numérique à l'impression en clair, lorsque cette alternative est retenue, présente un intérêt incontestable pour bénéficier des autres avantages du numérique et renforcer la robustesse de la conservation en fournissant deux moyens de représentation distincts.

2) Automatisation : la forme numérique préserve les possibilités de recherche et d'autres traitements automatisés du texte avec une intégrité garantie.

3) Robustesse : les mécanismes de correction d'erreurs intégrés à Micr'Olonys lui confèrent une certaine tolérance à la dégradation des documents au cours du temps, alors que la corruption d'un seul caractère imprimé peut induire la perte irrémédiable d'informations précieuses.

Une approche hybride est donc clairement bénéfique pour conserver sur papier des fichiers PDF destinés à l'archivage et qui comprennent de nombreuses images. Des tests réalisés fin 2023 montrent que cette approche permet généralement de diviser par quatre le nombre de pages d'un document contenant en moyenne une image toutes les deux pages.

B. Travaux annexes

Sumetar prend également en charge les fichiers texte brut. Pour ces fichiers, le texte est d'abord normalisé en UTF-8 s'il n'utilise pas déjà cet encodage. Le logiciel prend actuellement en charge la normalisation à partir de l'UTF-16 et de l'UTF-32 (big et little endian), ainsi que de l'UTF-7, de l'encodage système par défaut (par exemple Windows-1252 en Europe occidentale et en Amérique) ou d'un encodage spécifié dans le fichier texte au moyen d'une instruction « encoding= » ou « charset= » couramment employée

en HTML et XML. Cette fonctionnalité a été utilisée pour ajouter la table UTF-8 spécifique au document de base de données Andra produit en 2021 [1] afin de le rendre encore plus autonome.

Parallèlement aux travaux décrits précédemment, la version microfilm 16 mm de Micr'Olonys a été améliorée pour rendre la solution plus robuste lors de l'utilisation d'une gamme plus large de scanners de microfilms. Le microfilm offre une densité plus élevée que le papier, mais la numérisation nécessite un zoom optique approprié pour atteindre la résolution requise, et les résultats peuvent varier considérablement d'un scanner à l'autre. Dans un premier temps, les tests avec le scanner de film numérique Konica Minolta / Covergold SL1000 ont donné des résultats satisfaisants.

Début 2023, des tests ont été réalisés avec un scanner de microfilms Canon MS800II. Ces tests ont montré qu'une diminution de la résolution des emblèmes sur le film améliorait considérablement la facilité de numérisation, avec une tolérance beaucoup plus grande à un éclairage ou à une mise au point non uniformes et aux résolutions de numérisation plus faibles. Par conséquent, une option de « densité modérée », microfilmant les emblèmes en mode duplex classique et désormais fortement recommandée, est venue compléter la « haute densité » d'origine sur trois canaux qui peut toujours être utilisée pour des informations moins critiques ou plus détaillées. Que l'équipe de Progeima[7] qui a fourni tout le soutien nécessaire au succès de ces tests soit ici remerciée.

Au second semestre 2022 et en juillet 2023, des tests ont également été menés à la Bibliothèque nationale de France (BnF) avec un scanner de microfilm nextScan Eclipse équipé d'une caméra 12K pixels. Ces tests ont confirmé la robustesse de la nouvelle densité modérée d'emblème avec ce scanner, mais ont aussi montré qu'une distorsion horizontale résultant d'une numérisation imparfaite pouvait compromettre la détection de l'emblème par le logiciel intégré Micr'Olonys. La forme du cadre d'emblème Micr'Olonys a été modifiée en conséquence, tant pour la version microfilm que pour la version papier, afin de résoudre le problème. Nous tenons à remercier ici MM. Claude Da Costa et Patrick Bramoullé du département « Images et prestations numériques » de la BnF pour leur soutien dans la conduite de ces tests.

C. Travaux futurs

Certains documents Andra contiennent des images au format A3, et ainsi une fonctionnalité d'impression d'addenda mixte A3/A4 sera ajoutée dans Micr'Olonys et Sumetar.

La conversion de documents PDF en texte brut, même avec un certain niveau de mise en forme, dégrade inévitablement la mise en page d'origine et, par conséquent, la lisibilité ou même le sens du texte, notamment lorsque les caractères sont mis en indice ou en exposant, comme illustré par les figures 6 et 7. Par conséquent, il serait souhaitable de conserver les fichiers PDF et leurs mises en page afin d'extraire le contenu désiré au moment de l'accès dans le futur. Cette stratégie est parfaitement en adéquation avec la technologie de logiciel-sur-papier qui caractérise Micr'Olonys. Les images seraient retirées des fichiers PDF (et imprimées séparément dans l'addenda avec référencement approprié) pour limiter le nombre de pages, et le logiciel et les fonctionnalités nécessaires seraient intégrés à l'amorce Micr'Olonys pour extraire le texte UTF-8 brut ou les pages restituées sous forme d'images afin de conserver la mise en page d'origine. À cet égard, la grande simplicité de la spécification du format BMP le rend bien adapté pour restituer les images générées par le logiciel intégré dans l'environnement Olonys. De plus, le fichier PDF pourrait être restitué tel quel à l'utilisateur pour permettre un traitement spécifique. Sumetar pourrait également continuer à extraire le texte au moment de la transcription et le stocker conjointement au fichier PDF d'origine. La mise en page étant préservée dans le fichier PDF, l'extraction pourrait se concentrer sur le texte linéaire et non formaté, idéalement à partir d'un PDF balisé (par exemple un PDF conforme à la norme PDF/A) ou du document d'origine produit par le traitement de texte.

[7] progeima.com

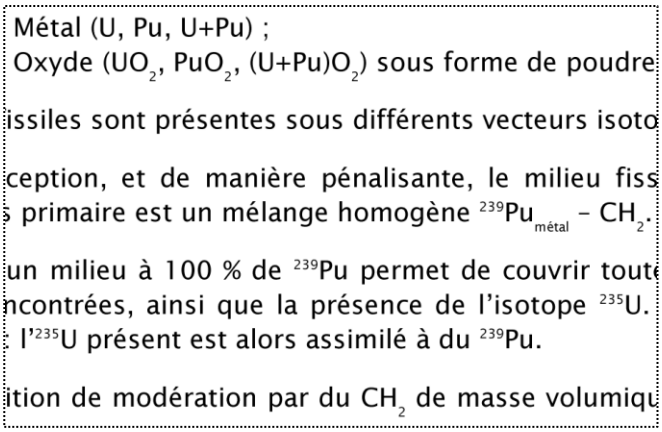

Figure 6 Mise en page originale dans un document PDF faisant un usage intensif d'indices et d'exposants.

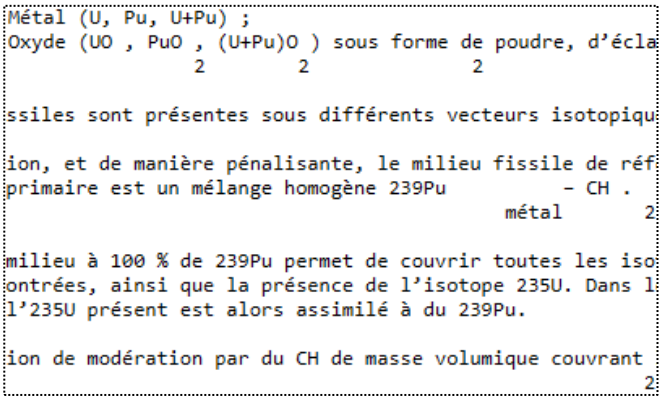

Figure 7 Conversion en texte brut montrant des caractères en indice mal placés et des attributs de caractères en exposant perdus.

Enfin, Eupalia améliore actuellement la vitesse d'exécution de la machine virtuelle Olonys qui est au cœur de l'amorce Micr'Olonys. Ces travaux s'inscrivent dans le cadre du projet HIPPOKAMPE (Hierarchical Implementation of Processor Performance Optimizations for Key naval defense Applications, with Market analysis and Patent Extension), soutenu par le projet Euroclusters LEVIATAD[8]. HIPPOKAMPE vise à développer des solutions de gestion de l'obsolescence logicielle ainsi qu'un processeur matériel robuste conçu pour fonctionner en environnement hostile, s'appuyant sur l'architecture de processeurs imbriqués Olonys.

VIII. Conclusion

Dans cet article, nous avons présenté un travail mené conjointement par l'Andra et Eupalia, reposant sur une stratégie hybride innovante qui combine numérique et analogique pour reproduire des documents PDF sur papier permanent en vue d'une conservation à long terme. Le contenu textuel, y compris une approximation du cadre des tableaux, est stocké sous forme numérique à l'aide de la solution Micr'Olonys, tandis que les images, l'encodage des caractères et les spécifications sont reproduits sous forme analogique en clair.

Tirant parti du meilleur des deux formes en termes d'espace occupé, d'accessibilité et de robustesse, cette stratégie fait de plus apparaître l'intérêt de mettre en œuvre la conservation inerte sous plusieurs formes ou sur plusieurs supports, avec une interface futur-homme-machine sous forme visuelle. Elle est particulièrement pertinente pour le stockage de données sur ADN, dont l'accès nécessite des technologies de pointe et des algorithmes complexes, pour conserver des informations et des spécifications critiques, mais aussi pour la conservation sur support magnétique si le format des contenus numériques est trop complexe (impliquant par exemple des logiciels) ou en l'absence de maintenance active pendant de longues périodes.

[8] leviatad.navigotoscana.it

Références

[1] V. Joguin et J.-N. Dumont, « Passive Digital Preservation on Paper in Practice », *in Actes de la 18e Conférence internationale sur la préservation numérique, iPRES 2022*, septembre 2022, Glasgow, Écosse, pp. 271-276. http://doi.org/10.7207/ipres2022-proceedings

[2] J. Tétreault, P. Bégin, S. Paris-Lacombe et A.-L. Dupont, « Modelling considerations for the degradation of cellulosic paper », *Cellulose*, Springer Verlag, 2019, 26 (3), pp. 2013-2033.

[3] A.-L. Dupont et G. Mortha, « Chimie des processus de vieillissement des papiers et celluloses », 27 (2016).

[4] A. Mansoux, B. Howell, D. Barok et V.-M. Heikkilä, « Permacomputing Aesthetics: Potential and Limits of Constraints in Computational Art, Design and Culture », *in LIMITS '23: Workshop on Computing within Limits*, 14-15 juin 2023. https://doi.org/10.21428/bf6fb269.6690fc2e

[5] V. Joguin, « Passive Digital Preservation Now & Later: Microfilm, Micr'Olonys and DNA », *iPRES 2019*, septembre 2019, Amsterdam, Pays-Bas. https://ipres2019.org/static/pdf/iPres2019_paper_139.pdf

[6] R. Appuswamy et V. Joguin, « Universal Layout Emulation for Long-Term Database Archival », *CIDR 2021*, janvier 2021, Chaminade, États-Unis. http://cidrdb.org/cidr2021/papers/cidr2021_paper30.pdf

[7] V. Joguin, « Support optiquement discernable par un utilisateur, figurant des données numériques et le moyen de les décoder », Brevet WO/2023/001659, 26 janvier 2023. https://patentscope.wipo.int/search/fr/detail.jsf?docId=WO2023001659