Annexe technique — opérations de décodage des diskmags belges

Cette annexe détaille, opération par opération, les techniques mises en œuvre pour extraire le texte des diskmags de la scène démo belge. Le cas central est *Scenial* (numéros 1, 2 et 4), dont les trois formats internes diffèrent ; on renvoie ponctuellement, par contraste, à *Blister* (scène ANSI) et au *Belgian Scene Report*, traités selon les mêmes principes. Chaque technique est présentée par son principe, puis illustrée par un ou deux exemples concrets, avec les octets effectivement rencontrés. Les valeurs sont notées en hexadécimal (préfixe `0x`) ; un « octet » vaut huit bits, soit une valeur de `0x00` à `0xFF`.

Le contexte commun est le suivant : un diskmag est un exécutable MS-DOS qui embarque ses propres données (textes, images, musiques, fontes) et les affiche au moyen d'un moteur écrit sur mesure. Aucune de ces structures n'est documentée ; tout se déduit du fichier.

---

1. Reconnaissance : cartographier le fichier

Principe. Avant toute tentative de décodage, il importe de localiser les zones de texte parmi les graphismes, le code machine, les fontes et la musique. Deux instruments suffisent : le calcul, bloc par bloc (par exemple toutes les 4 096 octets), de la proportion d'octets imprimables et de lettres ; et l'extraction des chaînes de caractères lisibles. Une zone de texte se signale par une forte densité de lettres et d'espaces ; une zone graphique, par des rampes de valeurs voisines ; du code, par une distribution sans structure lexicale.

Exemple A — Scenial 4 (`SCENIAL4.YOU`). Le fichier (2,57 Mo) s'ouvre sur près d'un mégaoctet de données d'image. La densité ne devient « textuelle » qu'à partir de l'offset `0x101000`, où commence le flux des articles. La cartographie indique donc d'emblée où chercher, et où ne pas chercher.

Exemple B — Scenial 1 (`SCENIAL.EXE`). La densité ne signale qu'une seule zone franchement textuelle, autour de `0x66B00`, qui se révèle être le sommaire. Partout ailleurs, la proportion de lettres reste basse : les corps d'articles ne sont donc pas stockés en clair, et un simple « extraire les chaînes » ne donnera rien. Ce constat négatif oriente la suite vers l'hypothèse d'un encodage.

---

2. Éprouver des transformations d'octets réversibles

Principe. Lorsqu'un texte n'apparaît pas en clair, il peut être masqué par une transformation simple et réversible : un décalage additif (`+k`, `-k`), un OU exclusif (`XOR k`), une inversion de bits. On applique chaque transformation candidate, bloc par bloc, et l'on retient celle qui produit non seulement une forte densité de lettres, mais de véritables mots.

Exemple A — Scenial 1, la clé `+0x0A`. Les corps d'articles (ressource `SCENIAL.NNN`) sont décalés. La transformation correcte ajoute `0x0A` à chaque octet (puis retranche `0x20` si le résultat dépasse `0x7F`, point repris en technique 7). Sur les onze premiers octets utiles :

```
brut    : 5D 5E 65 69 6A 6D 68 5F 6A 5B 68
+0x0A   : 67 68 6F 73 74 77 72 69 74 65 72
texte   :  g  h  o  s  t  w  r  i  t  e  r       → « ghostwriter »
```

Cette clé n'a pas été trouvée d'emblée : les premiers tests portaient sur `±0x20` (le décalage qui sépare majuscules et minuscules), choix naturel mais infructueux. L'élargissement de l'éventail des décalages a seul permis d'atteindre `0x0A`.

Exemple B — un faux positif (Scenial 2 et 4). Sur ces fichiers, `+0x20` et `XOR 0x20` produisent des blocs où plus de 80 % des octets tombent dans la plage des lettres ; le test de densité les valide à tort. Le décodage révèle pourtant des suites comme `UTTUX]ceba` ou `,-/-/-/-,` : des rampes de valeurs issues d'images, non des mots. La leçon est méthodologique : la densité de lettres ne suffit pas, il faut exiger des mots attestés.

---

3. Reconnaître le format « copie d'écran » : paires attribut/caractère

Principe. De nombreux diskmags stockent leurs pages comme une copie du tampon de l'écran en mode texte VGA : chaque case occupe deux octets, l'un pour le caractère, l'autre pour l'attribut de couleur. Un octet sur deux est donc du texte ; l'autre, intercalé, abaisse de moitié la densité apparente et brise les chaînes, ce qui explique l'échec des recherches naïves. Il faut isoler la bonne parité, et déterminer l'ordre des deux octets.

Exemple A — Scenial 2, attribut d'abord. Les octets pairs portent l'attribut, les impairs le caractère :

```
octets  : 0E 45 0E 4E 0E 54 0E 45 0E 52
attribut:  0E    0E    0E    0E    0E
caractère:    45    4E    54    45    52
texte   :     E     N     T     E     R          → « ENTER »
```

Exemple B — Scenial 4, caractère d'abord (parité inverse). Le même principe, mais l'ordre des deux octets est permuté :

```
octets  : 47 00 48 00 4F 00 53 00 54 00
caractère: 47    48    4F    53    54
texte   :  G     H     O     S     T           → « GHOST… » (Ghostwriter)
```

La détection de la parité correcte est elle-même une opération : on essaie les deux, et l'on retient celle qui donne des mots.

---

4. Retrouver la largeur de la grille

Principe. Dans ces pages, il n'existe aucun caractère de retour à la ligne : le texte est justifié sur une largeur fixe de colonnes, et les passages à la ligne sont purement positionnels. Pour reconstituer les paragraphes, il faut donc retrouver cette largeur `W`. Le critère le plus fiable consiste à minimiser les coupures en plein milieu d'un mot aux frontières de ligne : à la bonne largeur, un texte justifié ne scinde jamais un mot.

Exemple A — Scenial 2, `W = 40`. À cette largeur, aucune coupure de mot n'apparaît, et le rendu se lit comme un texte continu (« This is the second issue of our diskmag SCENIAL… »). Les largeurs voisines produisent des scissions du type `releasi|ng`, `grap|hics`, signalant une valeur erronée.

Exemple B — Scenial 4, `W = 38`, et le piège de l'indicateur mécanique. Ici, un premier indicateur — la régularité des lignes pleines ou vides — désignait `32`, et un second, faussé par les nombreux centrages de la mise en page, suggérait `75` ou `84`. Seul le critère des coupures de mot, confirmé par une lecture à l'œil, a fixé `38`. La preuve par l'erreur est nette : rendu à `76` (soit `2 × 38`), le texte se dédouble, chaque ligne logique en contenant deux.

---

5. Lire l'annuaire des ressources

Principe. Un diskmag compilé regroupe ses ressources dans une archive interne, décrite par un annuaire — une suite d'entrées de taille fixe contenant un nom de fichier (au format 8.3), un offset de début, un offset de fin et des tailles. L'annuaire se trouve souvent en fin de fichier. Le repérer permet d'isoler chaque ressource sans ambiguïté.

Exemple A — Scenial 2, annuaire à `0x1c58a0`, entrées de 64 octets. Chaque entrée se lit ainsi (valeurs en petit-boutiste) :

```
nom (12) : "SNLSUPRT.DAT"
début (4): 80 2C 05 00   → 0x00052C80
fin   (4): C1 4B 05 00   → 0x00054BC1
taille   : 41 1F 00 00   → 0x1F41   (= fin − début)
```

L'annuaire recense 104 articles `.DAT`, tous non compressés. L'ordre de lecture, les titres, les auteurs et les rubriques proviennent d'une autre ressource, `SCENIAL.MNU` (le sommaire éditorial, cf. technique 8).

Exemple B — Scenial 1, même schéma à `0x113f96`. Le même format d'entrées de 64 octets révèle la structure du numéro : une ressource `SCENIAL.NNN` (le conteneur des articles, encodé), `MENU.SNL` (le sommaire, en clair), `MAG.EXE` (le moteur d'affichage, lui-même stocké comme ressource), ainsi que les musiques. L'annuaire est ici la clé qui sépare le contenant du contenu.

---

6. Identifier un exécutable compressé — et décider s'il faut le défaire

Principe. Un exécutable peut être compressé par un utilitaire d'époque, repérable à sa signature et à la forme de son en-tête. La question n'est pas seulement technique mais décisionnelle : la décompression est-elle sur le chemin critique, ou les données utiles se trouvent-elles ailleurs ?

Exemple A — Scenial 1, LZEXE 0.91 : décision de ne pas décompresser. La signature `LZ91` figure à l'offset `0x1C`. L'en-tête MS-DOS décrit toutefois un chargeur d'environ 4 Ko seulement ; l'essentiel du fichier (plus d'un mégaoctet) est un *overlay* ajouté à la suite. Or le texte réside dans cet overlay, non dans la partie compressée par LZEXE. Décompresser le chargeur n'aurait donc rien apporté : la bonne décision était de l'ignorer et de traiter l'overlay directement.

Exemple B — Blister 3, PKLITE : décision inverse. Dans ce cas voisin, les écrans du magazine sont effectivement enfermés dans l'image compressée de l'exécutable (`BLIST3R.EXE`, comprimé par PKLITE 1.x). Il a fallu, cette fois, décompresser réellement — par émulation du processeur 8086 exécutant le stub de décompression (bibliothèque *unicorn*), puis relecture de la mémoire obtenue. Même famille d'indice, décision opposée : la généralité utile est qu'une signature de compression n'implique pas mécaniquement une décompression.

---

7. Casser un encodage propriétaire et reconstruire la table de ponctuation

Principe. Lorsque les octets sont décalés ou remappés, retrouver les lettres ne suffit pas : il faut reconstituer toute la table de correspondance, classe de caractères par classe de caractères, en alignant la sortie décodée sur de l'anglais lisible. Les caractères qui « ne collent pas » trahissent des codes de mise en page ou de la ponctuation déguisée.

Exemple — Scenial 1 (table complète). La règle de décodage est : `d = (brut + 0x0A) & 0xFF` ; si `d ≥ 0x80`, alors `d -= 0x20` (le bit haut distinguait deux « couleurs » de caractère, que l'on réunit). Le texte est stocké entièrement en minuscules. À partir de là, l'alignement sur des phrases connues a livré la table :

```
0x40  '@'          → espace
0x14 0x17          → retour chariot + saut de ligne (fin de ligne)
0x47  'G'          → apostrophe          (ex. « ainGt » → « ain't »)
0x4C  'L'          → virgule             (ex. « heartL » → « heart, »)
0x4D  'M'          → trait d'union
0x4E  'N'          → point               (ex. « NNN » → « … »)
0x4F  'O'          → barre oblique       (ex. « downloadingOcopying »)
0x50 / 0x51        → codes de début de ligne (normal / centré)
0x7C  '|' + code   → changement de couleur
0xC4 (×n)          → filet horizontal
```

Limite assumée : les grands chiffres (années, prix, caractéristiques techniques) étaient dessinés par une fonte spéciale et ne se laissent pas relire ; ils sont restitués par des crochets, sans valeur inventée. Cette technique est par nature propre au numéro 1 ; un seul exemple suffit à l'illustrer.

---

8. Découper en articles (I) — par l'annuaire et le sommaire éditorial

Principe. Quand l'archive nomme chaque article et qu'un sommaire éditorial en donne l'ordre, le découpage est direct : l'annuaire fournit les octets, le sommaire fournit l'ordre, les titres et les auteurs.

Exemple — Scenial 2. La ressource `SCENIAL.MNU` se lit en clair selon un motif régulier de quatre lignes par article :

```
RUBRIQUE        : "THE GHOSTWRITER"
fichier         : "GHOSTWRI.DAT"
sous-titre      : "Ghostwriter"
auteur\groupe   : "Venior\Beans"
```

Il suffit d'apparier chaque `…DAT` du sommaire à l'entrée homonyme de l'annuaire (technique 5) pour obtenir, dans l'ordre de lecture, les 103 articles avec leurs rubriques (Home Base, Scene Hall, The Lobby…) et leurs auteurs.

---

9. Découper en articles (II) — par la signature de fin d'article

Principe. En l'absence d'annuaire de textes, les articles peuvent se suivre sans séparateur explicite. Une régularité de mise en page sert alors de borne : un motif récurrent, encadré de filets, marque la transition.

Exemple A — Scenial 4. Chaque article s'achève sur une signature prise en sandwich entre deux filets, immédiatement suivie du titre de l'article suivant, lui-même suivi d'un filet :

```
… corps de l'article k …
═══════════════════════  (filet)
Written by VENIOR/BEANS
═══════════════════════  (filet)
JUNKMAIL — YEAH, WE DO READ YOUR MAIL   ← titre de l'article k+1
═══════════════════════  (filet)
… corps de l'article k+1 …
```

La mention « Written by … » fournit donc à la fois la fin d'un article, son auteur, et l'amorce du suivant. Ce repère a permis de segmenter les 122 unités du numéro.

Exemple B — Scenial 1. Le même dispositif de signature « written by … » se retrouve, après décodage (technique 7), dans le premier numéro, pourtant d'un format tout différent. La technique éprouvée sur le numéro 4 s'y est donc transférée telle quelle : c'est un cas concret d'enquête qui en éclaire une autre.

---

10. Découper en articles (III) — par appariement de titres

Principe. Lorsque la segmentation par signature et le sommaire ne coïncident pas exactement — fusions, sous-parties, granularités divergentes —, on réaligne les deux sources par similarité de titre, plutôt que par simple ordre, pour récupérer des intitulés et des rubriques propres.

Exemple — Scenial 1. Les titres reconstitués depuis les corps (« the rumor alley », « bbs advertisements », « things that make this scene suck »…) ont été appariés à ceux du sommaire `MENU.SNL` par mesure de similarité (recouvrement de triplets de lettres). Le procédé corrige la dérive d'ordre : un titre de corps trouve son équivalent canonique dans le sommaire même lorsqu'un article a été fusionné ou déplacé, et hérite de sa rubrique. La couverture passe ainsi d'un appariement par position, partiel, à un appariement par contenu, robuste.

---

11. Reconstituer le texte de lecture (reflow)

Principe. Une fois la grille reconstituée, il reste à la transformer en texte linéaire. Les règles sont simples : une ligne entièrement vide marque un changement de paragraphe ; une ligne de filets (`0xC4`) est une règle horizontale ; les espaces de justification multiples sont compactés. On distingue par ailleurs deux espaces : `0xFF`, espace « mot » appartenant au flux du texte, et `0x20`, espace de remplissage de la mise en page.

Exemple A — Scenial 2, éditorial à `W = 40`. Le rendu reflowé donne un texte propre :

```
ENTER THE WORLD OF SCENIAL!
---------------------------
This is the second issue of our diskmag SCENIAL. First of all I’d
like to thank all the guys who have supported us in releasing this
issue by sending graphics, music or articles …
```

Exemple B — Scenial 4, contrôle par la largeur. Un article en questions-réponses se reflowe correctement à `W = 38` ; rendu à `W = 76`, les colonnes se chevauchent (« more articles then most other » et « diskmags » se collent en « otherdiskmags »), ce qui confirme a contrario la largeur retenue.

---

12. Valider

Principe. Chaque décodage doit être contrôlé : par les comptes, par l’échantillonnage de passages lus intégralement, et par le recoupement avec des sources indépendantes (les fichiers NFO de présentation, le menu interne).

Exemple A — Scenial 2, cohérence des comptes. L’annuaire recense 104 fichiers `.DAT`, le sommaire éditorial 103 articles. L’écart d’une unité s’explique : le fichier surnuméraire est le fond de l’écran-menu, non un article. La cohérence est donc vérifiée, et l’anomalie comprise plutôt qu’ignorée.

Exemple B — intégrité de la traduction. Le nombre d’articles concorde, numéro par numéro, entre la version originale et la traduction française (35, 103 et 122), ce qui garantit qu’aucun article n’a été perdu en route ; les signatures d’auteur, enfin, sont recoupées avec les génériques des fichiers NFO d’époque.

Récapitulatif des formats

Numéro	Conteneur	Encodage du texte	Largeur	Segmentation
Scenial 1 (1994)	overlay LZEXE → `SCENIAL.NNN`	décalage `+0x0A` (puis `-0x20` si ≥ 0x80), minuscules, ponctuation remappée	variable	signature « written by » + appariement au sommaire
Scenial 2 (1996)	archive de ressources, articles `.DAT`	copie d’écran, attribut d’abord	40 colonnes	annuaire + sommaire `SCENIAL.MNU`

Numéro	Conteneur	Encodage du texte	Largeur	Segmentation
Scenial 4 (1997)	`SCENIAL4.YOU`	copie d’écran, caractère d’abord	38 colonnes	signature « Written by … »

Les techniques ne sont pas propres à un magazine : elles forment un répertoire transférable. La reconnaissance (1), l’épreuve des transformations (2) et la validation (12) s’appliquent à tout fichier ; le format en copie d’écran (3-4) se retrouve d’un numéro à l’autre et jusque dans la scène ANSI ; la lecture d’annuaire (5) et la décision face à un exécutable compressé (6) valent pour l’ensemble du corpus. C’est l’accumulation de ce répertoire qui rend chaque nouvelle enquête moins coûteuse que la précédente.