Traitement mixte image/texte de documents anciens

Jacques André(1), Jean-Daniel Fekete(2) et Hélène Richy(1)

9 mai 1995

(1) Irisa, Campus de Beaulieu, F-35042 Rennes cedex
jandre@irisa.fr, richy@irisa.fr
(2) LRI, Orsay ...
jdf@lri.fr

Résumé

Nous nous intéressons aux « textes à intérêt graphique » qui sont des documents (par exemple les manuscrits anciens) que l'on étudie tant pour le fond que pour la forme. Les mettre sous forme électronique nécessite de les manipuler à la fois comme du texte et comme des images. La notion de « zone sensible » dans des images (en HTML par exemple) permet une première approche mais est insuffisante pour manipuler électroniquement des documents anciens de façon professionnelle.

1 Introduction

La dualité saussurienne de signifiant et de signifié s'applique non seulement au sens des textes, mais aussi à leur forme, Ceci se retrouve à plusieurs niveaux.

– À un niveau global : un document peut être vu sous son aspect physique (une ligne avec un retrait de 1 cadratin, en Times-Italique corps 12, en drapeau à droite) ou sous son aspect structurel hiérachique (un titre de sous-section). C'est ce qui distingue notamment les éditeurs de textes graphiques (comme Word, FrameMaker, etc.) des éditeurs de documents structurés (comme LaTeX, Grif, etc. et, par abus de langage, SGML et HTML). Voir [3][4]à ce sujet.

– À un niveau beaucoup plus bas : les typographes distinguent depuis longtemps la notion de caractère de la trace encrée produite par impression de ce caractère qui s'appelle « œil » en typographie. Cette notion correspond à ce que l'on appelle maintenant « glyphe » dans les normes de codage de caractères [2]. À un caractère donné, par exemple « lettre A majuscule » peut correspondre plusieurs glyphes (par exemple un « A » en Times-Romain, un « A » en Courier, etc.) et réciproquement le même glyphe peut correspondre à des caractères différents, par exemple « A » peut correspondre à « lettre A majuscule » ou à « lettre grecque ALPHA majuscule », tandisque certains glyphes (comme la ligature « fi » ne correspondent pas à des entrées des tables de codage d'échange de caractères. Il faut alors des outils informatiques pour passer de l'un à l'autre.

– À un niveau intermédiaire : on peut étendre cette notion de glyphe à un ensemble de caractères, par exemple à un mot ou à un ensemble de mots, voire à un paragraphe ou à un texte entier. Cette dualité sens/image est différente de celle structure/graphique que l'on retrouve dans les document (structurés ou non) de notre premier niveau. Des chercheurs parlent alors de concept semio-linguistique d' « espace graphique » , notamment Jacques Anis [8] . Ceci permet de voir la page d'un mansucrit de Proust soit comme une image graphique, soit comme du texte. Avec comme corollaire la nécessité de disposer d'outils permettant, par exemple, d'afficher sur un écran l'image (notion graphique) d'un brouillon et d'y retrouver le nom propre (concept linguistique) « Verneuil ».

Cette dernière notion s'adresse à plusieurs classes de personnes, notamment aux chercheurs en sciences humaines et plus particulièrement à ceux pour qui le texte est justement soit une image soit un contenu. Nous appellons *texte à intérêt graphique* des documents qui sont parfois consultés pour leur forme et non exclusivement pour leur contenu textuel. Par exemple, un exemplaire de la *Bible à 42 lignes* de Gutenberg sera consulté, aujourd'hui, plus pour sa forme que pour son contenu. Le manuscrit d'un texte de loi dicté par Charlemagne sera aussi utile aux historiens, qui en liront le texte, qu'aux paléographes, qui en étudieront la forme de l'écriture.

L'objet de cette note est de montrer quelques problèmes liés à la manipulation de tels textes à intérêt graphique. Nous nous plaçons ici dans le cadre d'un éditeur de texte ou dans le cadre d'un « poste d'écriture » [1, 22]. Notre propos est par ailleurs plus de poser quelques problèmes que d'y apporter dès à présent des solutions.

Pour analyser ces besoins spécifiques des chercheurs travaillant sur des textes, nous nous basons sur quelques de cas qui ont en commun d'être des travaux

faits par des historiens, critiques littéraires ou pédagogues avec une méthodologie traditionnelle mais dont une version « électronique » expérimentale est en cours.

2 Le *Cartulaire de Saint-Laurent*

Le document en question est un cartulaire privé du XIII^{e} siècle qui a été découvert « par hasard » parmi des papiers de l'abbaye de Saint-Magloire de Paris par Marc Bloch, qui a été étudié par Anne Terroine et dont l'édition finale a été assurée par Lucie Fossier[23] à la mort de celle-là. Il s'agit d'un ensemble de 170 actes rédigés en français, couvrant une période allant de 1264 à 1277, composés sur l'ordre de Geoffroy de Saint-Laurent, un bourgeois parisien du village du même nom (aujourd'hui il s'agit d'un quartier de Paris, proche de la rue Saint-Martin et du boulevard de Sébastopol). La publication de ce cartulaire est très importante car c'est « une tentative réussie de restitution d'un milieu parisien du XIII^{e} siècle » [23].

Cette édition comprend les résumés des actes (les textes complets n'auraient pas présenté d'intérêt suffisant), de très nombreuses notes en bas de page et diverses études. Nous en avons fait une version électronique dont la principale finalité était de tester un système d'indexation [19][20] mis au point dans le cadre de l'éditeur Grif [17]. Mais il nous a paru intéressant de montrer que l'on pouvait aussi faire apparaître une image de la ou des pages correspondantes de l'acte manuscrit (figure 1).

Ainsi que nous l'avons montré [5][6], ce système d'indexation basé sur des concepts hypertextuels, permet de construire et utiliser de façon dynamique des tables d'index. Ainsi, suffit-il de cliquer sur une référence d'un élément de la table des noms de personne, par exemple sur la première référence à *Eudes le Charpentier* (voir figure 2) pour faire apparaître le résumé de l'acte correspondant avec, en surlignage, l'occurrence de ce nom. Signalons que ce qui est référencé peut être non seulement un mot, mais un bout de texte, un paragraphe, une section, etc., plus généralement une zone comprise entre deux marqueurs ou ancres. On a déjà là une sorte d'adressage physique d'une partie de document.

Puisque l'on sait montrer les images du manuscrit original, on a envie, lorsque l'on clique sur une référence à Eudes le Charpentier, de faire apparaître non seulement le résumé de l'acte correspondant comme en figure2, mais aussi la page du manuscrit de cet acte, voire d'y entourer l'occurrence de ce nom. La figure 3 montre ce que l'on aimerait voir sur un écran, mais elle a été faite en partie manuellement : des problèmes de plusieurs ordre sont en effet encore à résoudre avant de faire automatiquement une telle figure, dont les suivants.

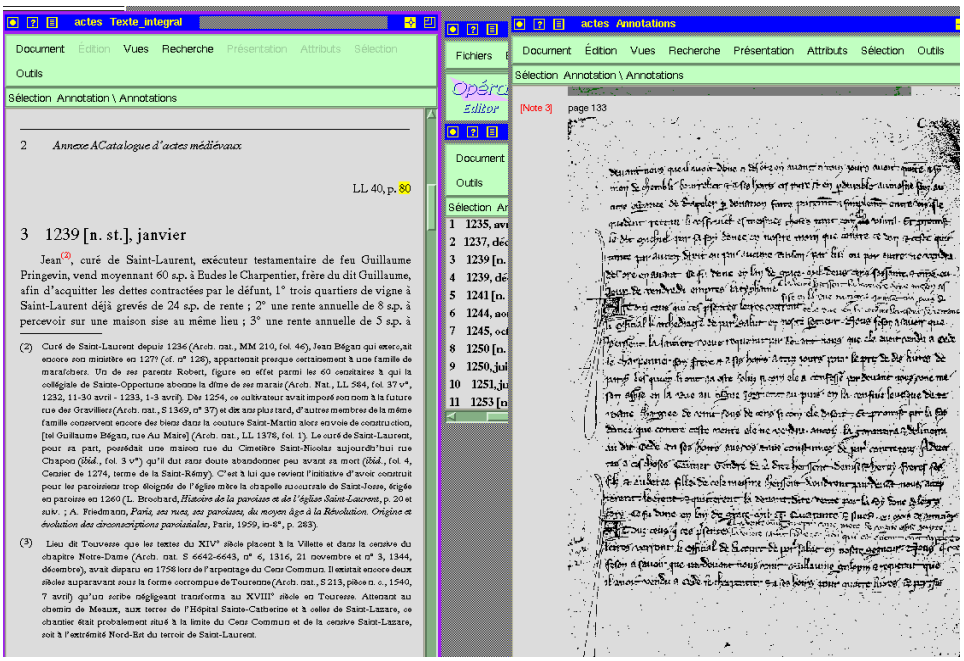

FIG. 1 - *Le cartulaire de St Laurent vu par Grif*

1. La reconnaissance optique des caractères manuscrits anciens n'est pas aujourd'hui complètement au point (même si diverses études montrent que ce sera pour bientôt, voir par exemple [9][10]). Dans notre exemple, nous avons donc « reconnu » nous-mêmes ce nom.

2. Il faut pouvoir « montrer » une partie de manuscrit, ici en l'entourant. Ceci est possible en Grif[1] en intégrant dans un même document structuré des images et des objets graphiques. Le contour d'un objet graphique peut avoir la forme d'un cercle ou d'un rectangle; il peut également suivre le tracé d'une ligne brisée quelconque tracée par l'utilisateur de Grif directement sur l'image affichée à l'écran, et cela à l'aide du clavier et de la souris.

3. Enfin, il faut pouvoir le référencer. Ceci est possible en utilisant les liens (ou références): un lien peut avoir pour origine un graphique et pour destination une légende, ou réciproquement, partant d'un texte ou d'une portion de texte, un lien permet de désigner le graphique correspondant dans une image.

1. Cette nouvelle possibilité de Grif, développée par Irène Vatton à Grenoble, n'est pas encore documentée !

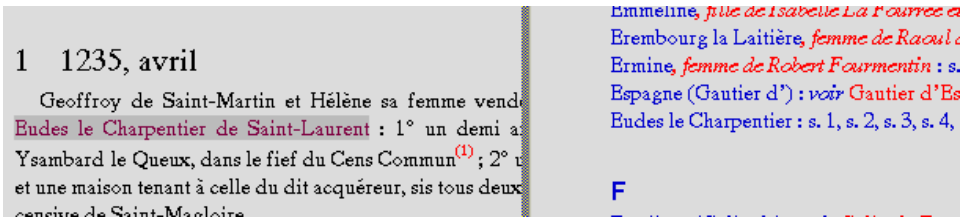

FIG. 2 - *Apparition (fenêtre de gauche) d'un acte après avoir cliqué dans l'index des noms de personne (fenêtre de droite).*

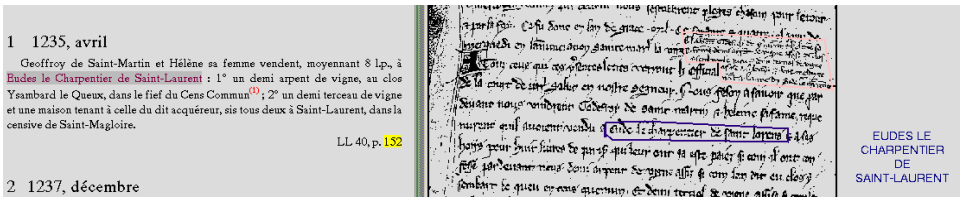

FIG. 3 - *Ce qu'on aimerait : après avoir cliqué (fenètre de gauche) sur un nom, voir ce nom localisé dans le manuscrit (fenêtre de droite).*

La combinaison des points 2 et 3 permet de réaliser simplement des «zones sensibles», de forme quelconque dans des images.

3 Le *Manuel de typographie* de Fournier

Ce *Manuel* du XVIII^e^ siècle est l'un des très grands classiques mondiaux en matière de typographie. Malheureusement, il est pratiquement introuvable tant en librairie que dans les bibliothèques publiques [2]. Il nous a donc semblé important de rééditer cet ouvrage et, tant qu'à faire, de le mettre dans le domaine public par le biais d'Internet [7]. Bien que, contrairement au *Cartulaire de Saint Laurent*, ce n'est pas le manuscrit de Fournier mais bien le livre imprimé qui nous intéresse, nous en avons saisi à la fois une version « moderne » (sans les ligatures ou les espaces de l'époque et sans les terminaisons en « -oit » pour l'imparfait par exemple) et scanné la version princeps du livre. En effet ce dernier comprend la moitié de planches (d'outils ou de caractères), et par ailleurs sa propre mise en page offre un intérêt évident.

2. Toutefois, une édition *facsimile* est sous presse, à Darmstadt (Allemagne), sous la drirection de James Mosley (*Sant James Library*, Londres) avec la traduction anglaise de M. Carter qui date des années 1930.

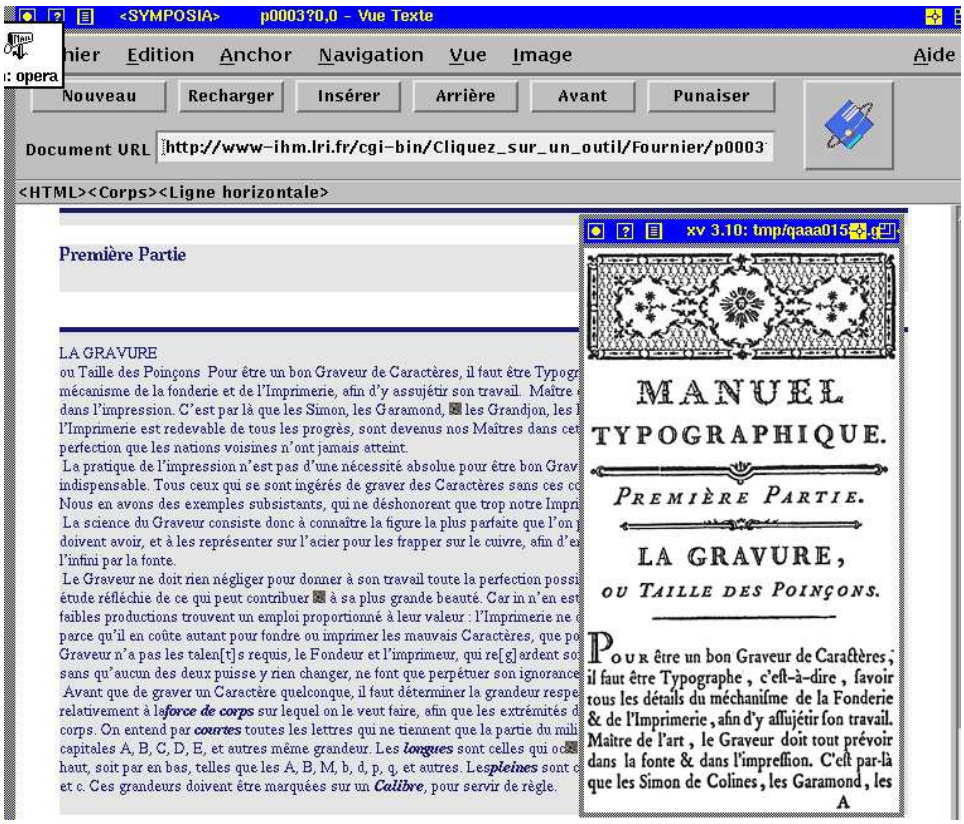

FIG. 4 - Manuel typographqiue *de Fournier dans sa version HTML. Les icones noires dans le texte moderne indiquent les fins de page du texte imprimé au* XVIII^{e} *siècle.*

FIG. 5 - *En cliquant sur l'index (par exemple « beuveau », on aimerait faire apparaître les pages scannées correspondantes.*

Nous avons commencé une version expérimentale pour montrer ce que pourrait être cette édition électronique d'un livre ancien. Une première ébauche est disponible[3]. On peut y voir (figure 4) à la fois la version moderne et la version imprimée du XVIII^e siècle.

3.1 Image physique et texte structuré

Un premier problème apparaît : la version moderne suit un schéma classique de document structuré qui ne connait pas *a priori* la structuration physique en page du livre. Nous avons donc été obligé d'utiliser des « artifices » (icones notamment) pour passer d'une forme à l'autre. Il en est de même pour appeler des gloses de commentateurs modernes qu'il est difficile d'appeler naturellement depuis le texte imprimé.

3.2 Indexation par images

Le second problème est beaucoup plus important ici. On aimerait considérer, comme pour le *Cartulaire de Saint-Laurent*, que les images scannées du *Manuel* soient non seulement des pixels, mais des hypertextes. Par exemple on aimerait (figure 5) cliquer sur une entrée (par exemple « beuveau ») de l'index de l'auteur et faire apparaître à la fois le passage où on en parle, la planche représentant cet objet et l'explication (la légende commentée) de la dite planche.

C'est un problème de même nature que l'indexation du cartulaire (section 2). Mais la table d'index monte le travail qu'il y aurait à prévoir manuellement toutes les zones sensibles (à raison d'une par numéro de page dans l'index !). Il faudra donc trouver un moyen pour rendre actives ces paginations, mais aussi pour les faire pointer automatiquement sur les pages scannées correspondantes !

3.3 Titre ?

Reprenons l'exemple du « beuveau » de la section précédente. Nous aimerions que cette édition moderne du *Manuel* de Fournier ait un côté pédagogique fort

3. `http://www.lri.fr/ jdf/Fournier/Fournier.html`; cette version a été partiellement faite « à la main », partiellement à l'aide de Grif, notamment dans sa version HTML [18] : il en existe une version « domaine publique » connue sous le nom *Symposia* : `http:?`.

et que, par exemple, on utilise les possibilités d'images vidéo. Insérer une image vidéo dans un « texte » HTML pose trois types de problèmes :

1. Afficher des images vidéo ! disons en gros qu'on sait faire. < JDF : peux tu dire en 2 mots qq chose, ou juste donner une ref biblio ou 2 et dire qu'ici ce n'est pas notre probleme majeur; mais que ca pose qd meme des problemes de lenteur pour le passage de films quand ils sont de bonne def etc.>

2. Gérer la synchronisation, le temps, d'une séquence vidéo avec autre chose. Là aussi pas notre prtoblème .< faire ref à Lahil ? >

3. Ancrer ces appels de séquence vidéo dans le texte.

C'est ce troisième problème qui nous intéresse ici ! Reprenons l'exemple du « beuveau » de la section précédente. Nous avons fait tourner[4] à l'Imprimerie nationale[5] quelques séquences sur les techniques de gravures de poinçons. Si maintenant on clique sur la planche où apparaît le beuveau on fait apparaître le nom de cette fausse équerre, et en cliquant sur « voir la video », le film est lancé. On peut consulter en même temps, par exemple, la description de Fournier (figure 6).

La zone sensible où se trouve le beuveau de la planche a été, ici, construit comme suit : < Merci JDF >.

La même technique pourrait être appliquée à de nombreux autres outils ou aux caractères du tome 2. Ces derniers pourraient aussi faire l'objet de commentaires sonores. Mais le problème n'est plus vraiment celui d'utiliser des séquences vidéo ou sonores, mais de les ancrer dans un document ancien de façon ergonomique !

4 Autre exemple ?

< JDF: as tu le temps d'ecrire 10 lignes sur ce que tu as mis dans ton dossier de CR2 sur les pb de gravure, etc. Si non, on laisse tomber ! >

5 Critique génétique

Dire qq chose? <JA>

4. Par le SEDIS de l'INRIA.
5. Au cabinet des poinçons, sous la direction de M. Papu.

FIG. 6 - *En cliquant sur l'image du beuveau (qui est une zone sensible), on fait défiler la séquence vidéo correspondante à cet outil.*

6 Conclusion

< Dire zones sensibles importantes, mais concept pas encore assez mur. Ça manque de typage, et ça peut être très difficile à manipuler « à la main ».

Mais rend déjà de bons services ! >

Références

[1] Jacques André, « Vers un poste de travail sur l'écrit », *Actes de la conférence « Hypertextes et Hypermédias »,* Alain Lelu et Imad Saleh, ed., pp. ?, Hermès éditions, Paris, mai 1995.

[2] Jacques André et Michel Goossens, « Codage des caractères et multilinguisme : d'Ascii à Unicode », *Cahiers GUTenberg,* n° 20, 5–60, avril 1995.

[3] Jacques André, Chrystelle Hérault et Hélène Richy, « Langages de description de feuilles de style », *Cahiers GUTenberg,* vol. 21, juin 1995 (ce numéro).

[4] Jacques André et Vincent Quint, « Documents structurés », *Le document éelctronique,* M. Bornes, ed., 1–53, ADBS et Inria, Rocquencourt, 1991.

[5] Jacques André et Hélène Richy, *Utilisation des index d'un éditeur structuré dans le cadre d'actes médiévaux,* Publication interne Irisa n° 841, Rennes, juin 1994.

[6] Jacques André et Hélène Richy, « Gestion électronique d'index et chartes médiévales », *Colloque Histoire et Informatique,* M. Cocaud, ed., Presses Universitaires de Rennes, Rennes, mai 1995 (à paraître).

[7] Jacques André, Hélène Richy et Jean-Daniel Fekete, « Editing Tools for WWWing Ancient Texts », *W4G Workshop on WWW Authorinh & Integration Tools,* P. Duval, ed., Inria, 8–10 février 1995.

[8] Jacques Anis, « Thought, language and hanwritting: what can we guess from literary drafts? », *Advances in handwriting & drawing: a multidisciplinary approach,* Claudie Faure *et al.*, ed., 531–545, Europia, Paris, 1994.

[9] Andrea Bozzi and Antonio Sapuppo, « Computer-aided Preservation and Transcription of Ancient Manuscripts », *Ercim News,* vol. 19, pp. 27-28, 1994.

[10] Claudie Faure, Paul Keuss, Guy Lorette et Annie Vinter (eds.), *Advances in handwriting & drawing: a multidisciplinary approach,* Europia, Paris, 1994.

[11] Pierre-Simon Fournier, *Manuel typographique utile aux gens de lettres,* Paris, 1764 (2 tomes).

[12] Almuth Grésillon, *Éléments de critique génétique. Lire les manuscrits modernes,* Presses Universitaires de France, Paris, 1994.

[13] Roger Laufer, « L'écriture hypertextuelle : pratique et théorie à partir d'une recherche sur *Rigodon* de Céline », *Littérature,* vol. 96, 106–121, décembre 1994.

[14] Roger Laufer, « Hypertexte : visualisation comparative et explicitation », *Hypermédias, éducation et formation,* E. Bruillard, B. de la Passardière et G.L. Aron, ed., 55–73, Laboratoire MASI, Université de Paris VI, 1994.

[15] Jean-Louis Lebrave, « Déchiffrer, transcrire, éditer la genèse », *Proust à la lettre. Les intermittences de l'écriture,* A. Grésillon, J.L. Lebrave et C. Viollet, ed., 143–162, Tusson, du Lérot, 1990.

[16] O. Mazhoud, E. Pacual et J. Virbel, « Représentation et gestion d'annotations », *3ème conférence << Hypertextes et hypermedias >>,* Imad Saleh et Alain Lelu, ed., Hermès ed., Paris, Mai 1995 (ces actes).

[17] V. Quint, I. Vatton, J. André et H. Richy, « Grif et l'édition de documents structurés : nouveaux développement », *Cahiers GUTenberg,* num. 9, 49–65, juillet 1991.

[18] Vincent Quint et Irène Vatton, « L'édition structurée et le World-Wide Web », *Cahiers GUTenberg,* num. 19, 85–97, janvier 1995.

[19] Hélène Richy, *Grif et les index électroniques,* n° 609, Irisa, Rennes, 1991.

[20] Hélène Richy, « A Hypertext Electronic Index Based on the Grif Structured Document editor », *Electronic Publishing - Origination, Dissemination and Design,* vol. 7, n° 1, 21–34, March 1994.

[21] C. Roisin, I. Vatton, « Merging Logical and Physical Structures in Documents », *Electronic Publishing – Origination, Dissemination and Design, EP94,* vol. 6, n° 4, 327–337, December 1993.

[22] Bernard Stiegler, « Lecture et édition savante assistées par ordinateur : l'hypertraitement de texte », *Actes du congrès Afcet 1993 (tome 4 : Bureautique, document, groupware et multimedia),* Gérard Dupoirier, ed., 37–45, Afcet, Versailles, juin 1993.

[23] Anne Terroine et Lucie Fossier, *Un bourgeois parisien du XIII*e *siècle : Geoffroy de Saint-Laurent, 1245?-1290,* CNRS Editions, Paris, 1992.

Table des matières

1 Introduction **1**

2 Le *Cartulaire de Saint-Laurent* **3**

3 Le *Manuel de typographie* de Fournier **5**
3.1 Image physique et texte structuré 7
3.2 Indexation par images 7
3.3 Titre ? 7

4 Autre exemple ? **8**

5 Critique génétique **8**

6 Conclusion **10**