

Les Carnets du  
**Cediscor**

## Les Carnets du Cediscor

Publication du Centre de recherches sur la didacticité  
des discours ordinaires

**8 | 2004**

**Les discours de l'internet**

---

# Les pages personnelles comme terrain d'expérimentation

Valérie Beaudouin, Serge Fleury et Marie Pasquier

---



### Édition électronique

URL : <http://journals.openedition.org/cediscor/710>

ISBN : 2878543149

ISSN : 2108-6605

### Éditeur

Presses Sorbonne Nouvelle

### Édition imprimée

Date de publication : 1 novembre 2004

Pagination : 143-164

ISBN : 2878543149

ISSN : 1242-8345

### Référence électronique

Valérie Beaudouin, Serge Fleury et Marie Pasquier, « Les pages personnelles comme terrain d'expérimentation », *Les Carnets du Cediscor* [En ligne], 8 | 2004, mis en ligne le 01 novembre 2006, consulté le 02 mai 2019. URL : <http://journals.openedition.org/cediscor/710>

---

Ce document a été généré automatiquement le 2 mai 2019.

Les carnets du Cediscor

---

# Les pages personnelles comme terrain d'expérimentation

Valérie Beaudouin, Serge Fleury et Marie Pasquier

---

- 1 Les sites web personnels sont des espaces de publication web offerts par les fournisseurs d'accès ou par des portails à leurs clients ou visiteurs. Ils contribuent fortement à l'audience de l'internet : certains serveurs ont un public essentiellement lié à la fréquentation des pages personnelles qu'ils hébergent. De nombreux travaux considèrent que les sites web personnels constituent un genre de discours autonome, voire un genre spécifique du web, qui n'aurait pas d'équivalent hors de l'espace numérique. Dans cette perspective, ces travaux ont cherché à identifier certains traits de structure et de contenu spécifiques de ces sites. Par-delà les traits propres aux pages personnelles, nous avançons l'hypothèse suivante : les sites web personnels partagent une fonction de *terrain d'expérimentation* et constituent un lieu d'apprentissage. Les sites web personnels ont une fonction de *brouillon*, qui comme tout brouillon a vocation à se transformer. Il s'agit d'un « écrit d'écran » en évolution constante, qui se nourrit de la visite d'autres pages et des réactions des visiteurs. En effet, une des spécificités de ces brouillons est leur confrontation avec un public, qui réagit et commente le site. C'est sans doute le caractère immédiat de la publication, qui fait de ces pages web des écrits révisables à tout moment. Aussi bien à travers le contenu que par la mise en forme utilisée pour formater ce contenu, les auteurs de sites bâtissent de fait par apprentissages progressifs une forme d'identité (Chandler, 1998).
- 2 Pour traiter cette hypothèse du brouillon, nous allons prendre un angle d'attaque spécifique : explorer des pages personnelles qui ont été visitées par un panel d'internautes et voir d'une part comment celles-ci ont évolué en un an, et d'autre part étudier les configurations de traits qui caractérisent des types de pages et qui peuvent être interprétés comme des états différents dans l'évolution des sites web. Pour cette étude, nous avons constitué un corpus des 101 426 pages personnelles visitées par une cohorte d'un millier d'internautes, extraites du panel NetValue<sup>1</sup> 1 entre janvier et juin 2000. Ce corpus de pages web a été soumis à une chaîne de traitement, nommée TypWeb

(Beaudouin *et al.*, 2001). Sur ce corpus, nous avons exploré la répartition de traits de différentes natures ; des traits « textuels » : les pronoms personnels, les déterminants et les adjectifs possessifs ; des traits présentationnels et structurels : la structure des liens hypertextuels internes et externes, la présence d'images, les polices de caractère présentes sur la page, certains éléments structurant les pages HTML<sup>2</sup> et les indications sur l'outil de conception de site utilisé, traits qui peuvent témoigner de l'expertise des concepteurs.

- 3 La difficulté majeure de l'exercice tient au problème de l'identification des traits pertinents pour décrire et différencier les pages web. Nous insisterons sur les difficultés que pose la construction d'indicateurs à partir du langage HTML. Après avoir fait un recensement des travaux portant sur les genres du web et en particulier les pages personnelles, nous décrirons notre corpus et notre méthodologie d'analyse. Ensuite, nous donnerons des indications sur le cycle de vie des pages personnelles avant de présenter quelques types de pages caractérisées par des configurations de traits spécifiques, qui témoignent d'états de maturation différents dans la conception du site en même temps qu'ils marquent des choix de publication spécifiques.

## 1. État de l'art

- 4 Cette section fait un état des lieux des travaux qui, en s'appuyant sur la description des éléments constitutifs des pages, visent à décrire et à comparer des ensembles de sites.

### 1.1. Définition des termes

- 5 En anglais, le terme de *homepage* est particulièrement ambigu, puisqu'il désigne soit la page d'accueil d'un site soit un site complet, tantôt un site personnel, tantôt un site institutionnel. La traduction en français par « page personnelle » redouble l'ambiguïté. Ce terme est utilisé par les fournisseurs d'accès pour désigner les espaces de publications offerts aux utilisateurs. De fait, ces « pages » sont le plus souvent des sites et elles ne sont pas toujours « personnelles », puisque ces espaces peuvent être utilisés par des associations, des PME, etc. Pour éviter ces ambiguïtés, nous appellerons *site web personnel* un espace de publication offert gratuitement par un portail ou un fournisseur d'accès à internet à ses clients ou visiteurs. Il peut être constitué d'une page ou d'un ensemble de pages qui relèvent d'une même *entité éditoriale* dont le statut peut être variable (personne, collectif...).
- 6 Notre corpus de travail est constitué de *pages personnelles visitées*, c'est-à-dire de pages qui appartiennent à des sites web personnels et qui ont été visitées par des internautes. Les pages qui constituent le corpus sont, en raison de ce mode de sélection, hétérogènes en terme de forme et de contenu. Dans quelle mesure peut-on considérer que ces pages personnelles visitées peuvent appartenir à un genre spécifique ? Y a-t-il des configurations de traits spécifiques de telles sous-catégories de pages ? De nombreux travaux ont tenté de montrer que les sites personnels constituent un genre à part entière (Amitay, 1999 ; Rehm 2002), voire le premier genre exclusivement numérique (Dillon et Gushrowski, 2000). Il nous semble au contraire que la catégorie des sites personnels est la plus hétérogène de la Toile et la plus opaque. Hétérogène, car s'y côtoient des sites ayant atteint un haut degré d'élaboration en termes de contenu et de structure avec des pages

embryonnaires ; opaque, car les adresses des sites, qui correspondent le plus souvent au nom de leur concepteur, donnent peu d'indications sur le contenu du site.

- 7 Trois grandes catégories de travaux sur les sites sont pertinents par rapport à notre sujet :
- les travaux portant sur la distribution des traits structurels et textuels dans les sites qui permettent de différencier des sites sophistiqués de sites plus ordinaires ;
  - les travaux portant sur la distribution des traits structurels et textuels qui permettent de différencier des genres au sein du web ;
  - les travaux portant plus spécifiquement sur les pages personnelles.
- 8 L'idée générale est bel et bien de caractériser un site par un ensemble de traits relatifs aux divers éléments constitutifs des pages et des sites web (texte, liens, images, sons, etc.). Il est ensuite possible d'étudier leur répartition à l'aide de méthodes statistiques multidimensionnelles, comme l'ont fait Biber (1995), Kalgreen et Cutting (1994) ou Habert *et al.* (2000) afin d'identifier des types de documents textuels.

## 1.2. Qualification de sites

- 9 En ce qui concerne la distribution des traits permettant de différencier les sites selon leur degré d'aboutissement, il convient de souligner l'intérêt de la démarche d'Ivory et Hearst (2002). Leur objectif est d'identifier les traits structurels et textuels qui sont la marque d'un « bon » site. Pour cela, les auteurs comparent la répartition des traits dans des sites qui ont été reconnus par des experts (*The international Academy of Arts and Sciences, The webby awards 2000 judging criteria*) avec des sites « ordinaires ». Ils proposent une série de traits permettant de caractériser les principaux dispositifs qui affectent la conception et la qualité d'un site web. Ils identifient alors 157 traits, qu'ils regroupent dans plusieurs classes générales : éléments de type texte, de type lien, de type graphique ; formatage des textes, des liens, des graphiques, de la page ; « efficacité » de la page, architecture du site, etc. Ils étudient les différences de répartition significatives entre les sites récompensés et les autres, ainsi que les corrélations de traits.

## 1.3. Identification des genres du web

- 10 Les travaux portant sur l'identification des genres du web sont de plus en plus nombreux. Reprenant et transposant les problématiques de l'identification des genres de discours<sup>3</sup> (Biber, 1995), les travaux portant sur le web cherchent aussi à observer des régularités dans des situations de communication spécifiques et tentent de les caractériser au moyen du triplet <contenu, forme, fonction>. L'internet est un média sur lequel sont repris, partiellement transformés, des genres existants (la lettre personnelle, par exemple), mais sur lequel émergent aussi de nouveaux genres (le *chat*). Il n'est donc pas possible de plaquer un inventaire prédéfini de genres importé d'autres modes de communication et il convient d'aborder la question de l'identification différemment. Sur l'internet, l'enjeu est considérable, puisqu'il s'agit de parvenir à catégoriser les documents dans des genres spécifiques, qui sont des ensembles sémantiquement homogènes. En outre, un découpage en genres est relativement pertinent pour les internautes, qui parviennent à catégoriser les pages web de manière intuitive et implicite. Par exemple, l'identification de genres sur le web peut permettre de pallier les résultats souvent inadéquats des moteurs de recherche en proposant à l'utilisateur de sélectionner d'abord un genre spécifique avant

d'effectuer sa recherche. Ces travaux portent sur des documents relevant de genres différents et tentent de les caractériser les uns par rapport aux autres par des ensembles de traits discriminants (Karlgreen *et al.*, 1994, Beaudouin *et al.*, 2001).

## 1.4. Traits spécifiques pour caractériser les sites personnels

- 11 Un certain nombre d'études se limite à un « genre » spécifique, défini a priori, et le caractérise par un ensemble de traits pertinents (Amitay, 1999 ; Rehm, 2002). Les plus aboutis de ces travaux portent sur les sites personnels. À partir de 1 000 pages personnelles, Amitay (1999 : 6) examine la répartition des hyperliens et la fréquence d'apparition des mots. Elle indique qu'« il n'est pas surprenant que les mots *I (je), my (mon, ma, mes)* et *me (me, moi)* soient très hauts dans la liste. Cependant le mot *you (tu, te, vous)* est également placé très haut, indiquant ainsi une tendance à utiliser un langage direct et informel – de moi (l'auteur) à vous (le lecteur). »
- 12 Dillon et Gushrowski (2000 : 202-205) ont montré que les traits reconnus comme typiques des pages personnelles par les visiteurs sont ceux qui ont la fréquence d'apparition la plus élevée dans les corpus de sites, ce qui montre que la page personnelle s'établit en soi comme un genre numérique unique. Ils ajoutent que « les pages personnelles [...] semblent avoir évolué très rapidement vers une forme standard [...] ».

## 2. Méthodologie

- 13 Nos travaux s'inscrivent donc dans cette mouvance. L'originalité de notre démarche est d'examiner de grands volumes de données provenant d'usages réels du web : le corpus initial, dans lequel nous étudions la répartition de traits textuels, présentationnels et structurels contient 101 426 pages visitées. La section suivante présente plus en détail la méthodologie suivie pour caractériser les pages personnelles.

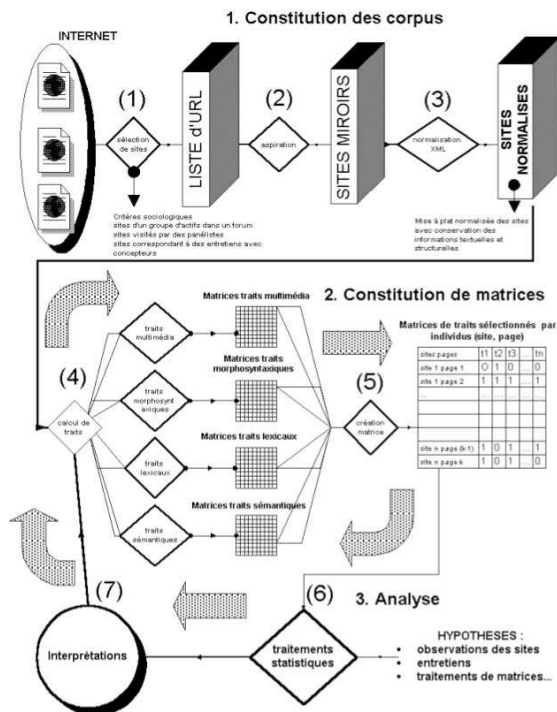
### 2.1. Chaîne de traitement

- 14 Dans le cadre du projet TypWeb, qui trouve ses prolongements dans le projet RNRT SensNet<sup>4</sup>, nous avons mis en place une chaîne de traitement qui permet, à partir de pages ou de sites du web, de constituer un ensemble d'indicateurs ou de traits descriptifs de la page ou du site. Cette chaîne comprend plusieurs étapes<sup>5</sup> :
- L'identification des sites ou pages à aspirer : dans le cas présent, il s'agit d'une liste d'URL visitées par au moins un panéliste entre janvier et juin 2000. Cette liste comprend 101 426 pages.
  - L'aspiration des pages ou des sites consiste à recopier intégralement un site ou une page à une date donnée en les stockant localement sur nos machines.
  - La normalisation des sites web aspirés vise à homogénéiser la description des documents.
  - L'étape de sélection de traits pertinents permet d'extraire à partir des sites ou des pages normalisés des traits spécifiques pour l'analyse. Les traits disponibles dans le corpus normalisé sont à la fois très nombreux et très complexes, ce qui rend indispensable la sélection et la normalisation de ces traits (nous y reviendrons *infra*). Pour cette étude, nous avons exploré la répartition des traits suivants : les marques de la personne (pronoms personnels, possessifs) ; la structure des liens hypertextuels internes et externes ; la

présence d'images ; les polices de caractère présentes sur la page ; certains éléments structurant les pages HTML et les indications sur l'outil de conception de site utilisé.

- La construction de matrices croise les pages et les traits choisis : à l'issue de la phase de normalisation, nous disposons d'états statistiques sur la composition des pages de ce corpus (comptage de mots, d'éléments HTML<sup>6</sup> : TAG, attributs, valeurs d'attribut...). Nous construisons des matrices de traits, que nous agrégeons, regroupons, réduisons... en fonction de leur distribution statistique. Ces matrices sont ensuite soumises, après reformatage, à des outils d'analyse statistique.

Figure 1. Chaîne de traitement TypWeb



## 2.2. Constituer des traits : sélection, regroupement, interprétation

- 15 Le nombre de traits disponibles à partir de l'analyse fine du code HTML est quasiment inépuisable. Traiter sans discernement une telle quantité de traits dont la signification et la valeur sont très variables n'aurait pas de sens. C'est pourquoi, en fonction d'une question posée, nous sélectionnons pour chaque étude particulière un sous-ensemble dans cette liste de traits. Cette phase de sélection de traits constitue de fait une étape cruciale qui, loin d'être triviale, pose des problèmes de différentes natures.

### 2.2.1. Comment sélectionner parmi le nombre considérable de traits disponibles ceux qui sont pertinents ?

- 16 Le choix des traits retenus est le résultat d'un compromis complexe entre les capacités de traitements des outils, notre capacité à organiser des traits dont les valeurs sont très dispersées, et la question de l'interprétabilité des traits.

- 17 L'adjectif « personnel » dans « page personnelle » sous-entend que l'instance d'énonciation est un individu et que le contenu lui-même renvoie à la personne. Le choix des pronoms personnels semble donc aller de soi. Nous avons complété ce choix de traits textuels par la prise en compte des déterminants et adjectifs possessifs. Notre volonté était de mener en parallèle une confrontation entre les éléments textuels et structurels présents dans une page web. Après cette sélection de traits textuels, notre choix d'attributs complémentaires pour cette analyse s'est orienté vers la sélection de traits structurels. Nous avons donc sélectionné de manière empirique des traits structurels, soit parce qu'ils correspondent à des traits traditionnellement reconnus comme pertinents pour caractériser des pages (c'est le cas des liens hypertextuels étudiés dans beaucoup de travaux), soit parce qu'ils nous semblaient être des éléments facilement identifiables et manipulables par les créateurs de page web pour produire des traces de leurs interventions (c'est le cas des traits permettant de marquer l'utilisation d'une police de caractère, sa couleur ; des traits permettant d'attribuer des propriétés à la page web (couleur de fond d'écran, image de fond d'écran (Turner, 2002), etc.).
- 18 Cette phase de sélection des traits utilisés ici a été conduite de manière empirique après de nombreux examens des données manipulées (les pages elles-mêmes) et des résultats provisoires produits (exploitables ou non). Ce travail mérite évidemment des prolongements que nous envisageons de poursuivre dans le cadre de projets en cours (SensNet, 2002).

### 2.2.2. Les traits comme somme de couples (attribut/valeur)

- 19 En ce qui concerne les traits structurels, il convient de préciser que les éléments HTML correspondants sont généralement accompagnés d'une liste d'attributs associés à des valeurs particulières : par exemple un élément FONT (police) peut être caractérisé par un type de police de caractères donné, par sa taille, sa couleur... Pour le trait FONT, on dispose donc au moins de 3 attributs, et pour certains attributs, les valeurs possibles se comptent par milliers : c'est le cas de l'attribut couleur. La couleur est un élément très fréquent dans la génération de pages web pour donner une couleur aux polices de caractères utilisées (la couleur comme attribut de police) ou pour modifier le fond d'une page (la couleur comme attribut de la page). L'utilisation de la couleur est, de fait, corrélée avec d'autres éléments de la page (police ou page), et son utilisation n'a de sens que par contraste (avec le fond ou avec d'autres polices), or le fond est partiellement insaisissable. Le nombre très élevé de couleurs différentes (plus de 7000 dans notre corpus) a rendu difficile l'exploitation des traits intégrant son utilisation dans leur définition (c'est le cas du trait FONT). En outre, il nous a été impossible de résoudre le problème du regroupement de ces couleurs et de leur catégorisation, ces problèmes débordant largement le cadre de notre étude. Il reste donc à trouver pour chaque attribut d'un TAG HTML les découpages pertinents.

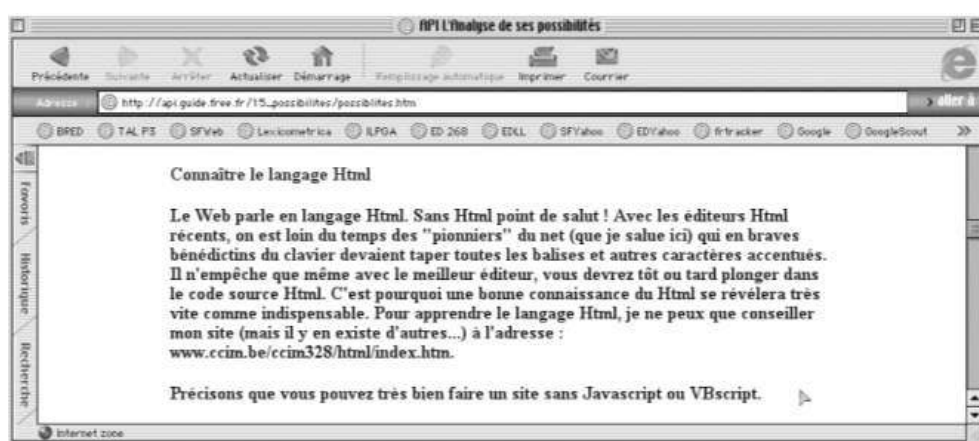
### 2.2.3. Connaître ou ne pas connaître le langage HTML (de la mauvaise écriture des traits)

- 20 La prise en compte de traits structurels passe par la prise en compte d'éléments écrits via le langage de conception des pages web (le langage HTML principalement). Il faut donc avoir une connaissance assez précise des éléments fournis dans ce langage pour réaliser des opérations de formatage ou de structuration des pages web. Cette phase de travail

nous a donc conduits à approfondir notre connaissance de ces éléments, en particulier lorsque nous étions confrontés à des éléments incorrects (mal orthographiés, inexistantes ou incorrects syntaxiquement). En effet, un élément HTML donné possède un nom et une syntaxe particulière, le non-respect de ces contraintes n'est en général pas traité par les navigateurs traditionnels, et, dans ce cas, le résultat attendu n'est pas construit (l'erreur peut rester parfois complètement invisible, même pour son créateur).

- 21 La figure qui suit est un extrait d'une page web pour apprentis créateurs de page web<sup>7</sup>. Le message est clair, « une meilleure connaissance du HTML se révélera très vite indispensable » pour construire des pages web cohérentes, attractives... Cet apprentissage se traduit de fait dans notre corpus par des essais de codage de pages web qui traduisent l'état d'avancement (parfois précoce) de celui-ci : la présence de balises mal orthographiées ou l'écriture de balises de manière syntaxiquement incorrecte.

Figure 2. « Sans HTML point de salut ! »



#### 2.2.4. De l'interprétation des traits

- 22 Si la sélection des traits constitue un problème non trivial, celui de l'interprétation des traits, de leur existence, de leurs utilisations croisées, etc., relève là encore d'une tâche complexe. Quelle signification attribuer au choix d'une police de caractère ? Que signifie l'absence de traits de police dans une page ? Faut-il attribuer cette absence à une mauvaise maîtrise des outils permettant de les utiliser ou bien faut-il l'attribuer à une volonté préméditée de laisser le navigateur décider de la nature de la police à utiliser pour la page visée ?
- 23 Cette difficulté pour un trait spécifique de donner un sens à son existence dans une page nous a conduits à mener des analyses pour lesquelles nous établissons des corrélations de traits : analyse conjointe de plusieurs traits présents/absents sur les pages traitées. Comme le souligne Ivory *et alii* (2000), « sur les pages de bonne qualité, la corrélation entre la couleur et le nombre de couleurs d'affichage suggère que ces pages emploient un arrangement à plusieurs niveaux de titre, les titres à chaque niveau ayant des couleurs différentes. Il y a également une corrélation entre le texte soigné et les bonnes combinaisons de couleurs par zone indiquent que ces pages utilisent simultanément des zones colorées et du texte coloré (par exemple, dans des barres de navigation). Les bonnes pages emploient également des tableaux pour contrôler le formatage des liens textuels et des images. »



### 3. Cycle de vie des pages personnelles

24 Pour valider notre hypothèse de la page personnelle comme terrain d'expérimentation et jouant le rôle de brouillon, nous avons étudié comment évoluaient les pages personnelles de notre corpus avec le temps. Pour ce faire, nous ne disposons pas d'un corpus chronologique au sens strict : nous ne possédons pas plusieurs états de toutes les pages à des dates différentes ; toutefois, la stratégie de constitution du corpus nous permet d'examiner la proportion de pages disparues (mort) et redirigées (évolution). En effet, les pages constitutives du corpus ont été visitées, au moins une fois, par une cohorte d'internautes entre janvier et juin 2000, tandis que la création du corpus n'a été réalisée qu'en mars 2001, soit un an après. Il s'est ainsi écoulé un laps de temps d'un an entre la consultation effective de la page et son archivage, ce qui nous permet de calculer un taux de disparition et un taux de redirection des pages personnelles et d'opposer deux types de trajectoires des pages qui confirment alors partiellement notre hypothèse.

#### 3.1. Taux de disparition des pages personnelles

25 Le taux de disparition révèle que 40 % des pages personnelles ont disparu du réseau un an après leur visite<sup>8</sup>. Effectivement, partant d'une liste initiale d'URL de 101 426 pages personnelles visitées, seules 69 579 pages ont été correctement aspirées et normalisées (bonne syntaxe HTML et sans message d'erreur) : 60 % de pages sont archivées et exploitables, ce qui est finalement relativement peu. En fonction des serveurs d'hébergement, la disparition de pages personnelles varie considérablement, comme l'indique le tableau 1. Les politiques commerciales et marketing des uns et des autres, ainsi que les problèmes liés à l'aspiration de certaines pages, peuvent en partie expliquer certaines disparités. On peut également noter que lorsqu'un serveur d'hébergement n'existe plus, toutes les pages disparaissent avec lui. C'est le cas de citeweb.net, disparu entre 2000 et 2001.

Tableau 1. Disparition des pages en fonction du serveur d'hébergement

Serveurs d'hébergement	Nombre de pages visitées	% de pages du corpus	% de pages disparues
www.multimania.com	20 864	21	27
free.fr	19 192	19	51
perso.wanadoo.fr	16 791	17	23
www.chez.com	11 082	11	32
www.geocities.com	9 068	9	30
ifrance.com	5 351	5	58
perso.club-internet.fr	4 785	5	34

members.aol.com	2 071	2	31
perso.infonie.fr	1 975	2	95
www.citeweb.net	1 638	2	100
autres	8 609	7	
TOTAL	101 426	100	40


### 3.2. Taux de redirection des pages personnelles

- 26 Les pages redirigées correspondent explicitement à une évolution de la page personnelle : il peut s'agir d'une migration vers un autre service d'hébergement, dont le style convient mieux aux attentes du concepteur, ou d'un désir d'autonomisation qui se concrétise par l'acquisition d'un nom de domaine. Pour parvenir à quantifier le nombre de pages personnelles redirigées dans le corpus, nous avons examiné les pages qui ne contenaient aucune marque de personne et un seul lien externe. Sur les 69 579 pages du corpus, 12 635 pages ont ainsi été identifiées, parmi lesquelles 28 % correspondent réellement à une redirection, 67 % sont des erreurs<sup>9</sup> et les 5 % restant pointent vers des compteurs, des adresses électroniques, des sites de mesure d'audience, etc.
- 27 L'analyse minutieuse des pages redirigées révèle que, dans la moitié des cas, la redirection correspond à l'acquisition d'un nom de domaine, tandis que les autres migrations se font chez un fournisseur d'accès concurrent.
- 28 Le cas de Citeweb.net, disparu entre 2000 et 2001, est particulièrement intéressant pour illustrer le phénomène de redirection des pages personnelles. Le service d'hébergement a conservé une page sur le réseau (reproduite en figure 3), sur laquelle il explique les raisons de sa disparition, et propose aux concepteurs de faire héberger leurs pages ailleurs : soit chez un hébergeur professionnel, auquel cas l'hébergement est payant, soit chez un service d'hébergement gratuit concurrent (FortuneCity.fr).

Figure 3. « Les portes de Citeweb sont fermées »

Les portes de Citeweb sont fermées. Le site n'est plus disponible.

Citeweb ne peut pas se maintenir car de nombreux pirates ont abusé de notre hébergement, saturant notre bande passante et nos ressources. Malgré de gros efforts, nous ne pouvons donc plus assurer le service fiable qu'a été Citeweb.



Si vous ne l'avez pas encore fait, il pourrait être temps pour vous d'obtenir un hébergement sûr chez Ampira, qui fournit un hébergement professionnel à tous ceux:

- qui sont fier de leur site et veulent leur nom de domaine propre,
- qui en dépendent pour le commerce et requièrent un service et un support technique ultra-fiables, et
- qui en ont tout simplement marre de la pub.

Si vous êtes intéressés par un tel hébergement, venez chez Ampira et choisissez une de ses offres d'hébergement.

Pour tout ancien membre de Citeweb, les frais d'inscription de \$29,99 vous seront offerts avec l'achat d'un hébergement chez Ampira.

Toutefois, si vous tenez à un hébergement gratuit à service limité, vous pourrez encore le trouver chez [FortuneCity.fr](http://FortuneCity.fr).

Merci de votre fidélité et bonne création,  
L'équipe de Citeweb

- 29 Grâce à ces premiers résultats, il est possible d'opposer deux trajectoires dans le cycle de vie de la page personnelle :
- soit la page est abandonnée : elle se maintient sur le réseau mais perd ses visiteurs et semble condamnée à disparaître ;
  - soit la page évolue, mûrit, et le plus souvent le développement de l'objet s'accompagne d'une migration vers d'autres lieux d'hébergement (acquisition de nom de domaine ou changement de serveur d'hébergement).
- 30 Par conséquent, l'analyse de l'évolution des pages nous confirme que la page personnelle est un objet qui évolue relativement rapidement dans le temps, ce qui valide partiellement l'hypothèse du « brouillon ». En outre, dans la masse de sites présents sur le web, tous n'ont pas la même longévité : les sites marchands ou institutionnels semblent ainsi avoir une existence plus durable que les pages personnelles. De fait, la brève existence de la page personnelle se révèle être une de ses spécificités. En effet, au cours de la constitution de corpus de sites personnels et marchands (voir Beaudouin *et al.*, 2001), nous nous sommes aperçus que les sites marchands « vivaient » beaucoup plus longtemps que les pages personnelles. Pour illustrer ce point, nous nous appuyons sur l'observation d'une quinzaine de sites marchands aspirés et archivés tous les ans depuis 1999, soit quatre états différents des mêmes sites.
- 31 Du côté marchand, les rares changements observés proviennent soit de la fusion de deux sites (exemple des sites de voyage *lastminute* et *degriftour*), soit de l'intégration d'un site dans un autre (exemple de l'intégration du site de voyage *expedia* dans le site de la *sncf*), soit de sa disparition (par exemple, le site de vente de biens culturels *bol*). Dans les deux premiers cas, les deux URL coexistent sur le web et l'utilisateur est automatiquement redirigé vers le site principal : l'opération de redirection est alors transparente pour le

visiteur. Dans le dernier cas, en revanche, l'URL disparaît : le visiteur est alors confronté, au mieux, à une page qui l'informe de la disparition du site ou, au pire, à une page d'erreur.

- 32 Afin de conserver des traces de ces évolutions et changements, de nombreuses actions sont mises en place pour archiver le web. Citons à titre d'exemple le projet [archive.org](http://archive.org)<sup>10</sup> qui collecte, enregistre et indexe des millions de pages depuis 1996 afin de les rendre accessibles sur son site, ou encore, dans un domaine plus restreint, les cimetières de journaux intimes en ligne (Lejeune, 2000).

## 4. Quelques trajectoires de pages personnelles

- 33 Nous venons de voir qu'une grande partie des pages visitées avait disparu quelques mois après avoir été visitées, ce qui montre que l'espérance de vie est assez faible pour un grand nombre de pages, ou qu'elles évoluent sans laisser de trace. Il a paru intéressant de reconstituer l'historique de quelques-uns des sites dont une page a été visitée et de comparer leur état au moment de la visite à leur état actuel. Nous nous appuyons à la fois sur des aspirations de sites réalisées à différentes dates et sur le site d'archivage des pages web, qui enregistre les états successifs de sites (<http://web.archive.org>).
- 34 Nous avons avancé que la page personnelle était un lieu d'expérimentation et de transformation de la production et nous pensons que la transformation des pages personnelles se fait dans une triple direction. Nous avons reconstitué la trajectoire de quelques-uns de ces sites choisis pour leur caractère archétypique que nous illustrerons par deux exemples. Premièrement, l'évolution de la page personnelle s'accompagne d'une disparition ou d'une mise en sourdine du moi. Le titre de la page de la figure 4 comportait dans son premier état le nom de son auteur : « La page de Frédéric Grillot » ; un an plus tard, le titre est devenu : « La buticulamicrophilie ou la passion d'un collectionneur ».
- 35 Le second mouvement de transformation est une autonomisation du site, qui passe par l'acquisition d'un nom de domaine. Ce mouvement permet de s'affranchir de l'image du serveur d'hébergement. Ainsi, le site de François Bon a d'abord été hébergé chez Wanadoo, puis chez Free avant d'acquérir son propre nom de domaine (Remue.net). Au fil de cette trajectoire, l'auteur s'est peu à peu effacé au bénéfice du thème de son site (littérature contemporaine et ateliers d'écriture). La figure 5 montre le même processus d'effacement de l'auteur. « François Bon » apparaissait dans la barre d'adresse (l'adresse du site étant [perso.wanadoo.fr/f.bon](http://perso.wanadoo.fr/f.bon)) et sur la page d'accueil. L'acquisition d'un nom de domaine s'accompagne d'un effort pour rendre moins visible l'auteur et rendre davantage visible le thème principal du site.
- 36 Enfin, les sites tendent à se centrer sur un seul sujet. La page d'accueil du site de la figure 4 est symptomatique de ce mouvement. Dans son premier état, elle présente trois centres d'intérêt : la ville de Carcassonne, le canal du Midi et la collection de petites bouteilles. Un an plus tard, le site est uniquement centré sur la collection de petites bouteilles. Dans leurs premiers états, les pages personnelles sont très centrées sur la personne et peuvent rendre compte des différentes identités ou centres d'intérêt de l'auteur. Dans un état ultérieur, les sites présentent de plus en plus rarement plusieurs axes thématiques, peu cohérents les uns avec les autres, et tendent à se centrer sur un seul thème. La cohérence thématique du site devient prioritaire sur la présentation de l'ensemble des centres d'intérêt de l'auteur. Là encore nous observons un glissement du sujet vers l'objet.

- 37 Il est ainsi intéressant de trouver, à la racine d'un des sites du leader, une page qui recense tous les sites qu'il a fabriqués chez différents hébergeurs sur des sujets très hétérogènes. En dehors de cette page, qui n'est référencée par aucun annuaire du web, il n'y a aucun lien hypertexte entre ces différents espaces. Le moi se trouve ainsi éclaté dans ses différentes projections. Là encore nous observons un glissement du sujet vers le thème, contrainte sans doute imposée par la logique de l'audience et l'accès par les moteurs de recherche qui privilégient le thème du site sur son concepteur. Pour avoir un public, il faut que le site soit reconnu comme un site d'expert, l'expertise étant peu compatible avec la diversification.

Figure 4. « La page de Frédéric Grillot » ou « La passion d'un collectionneur »

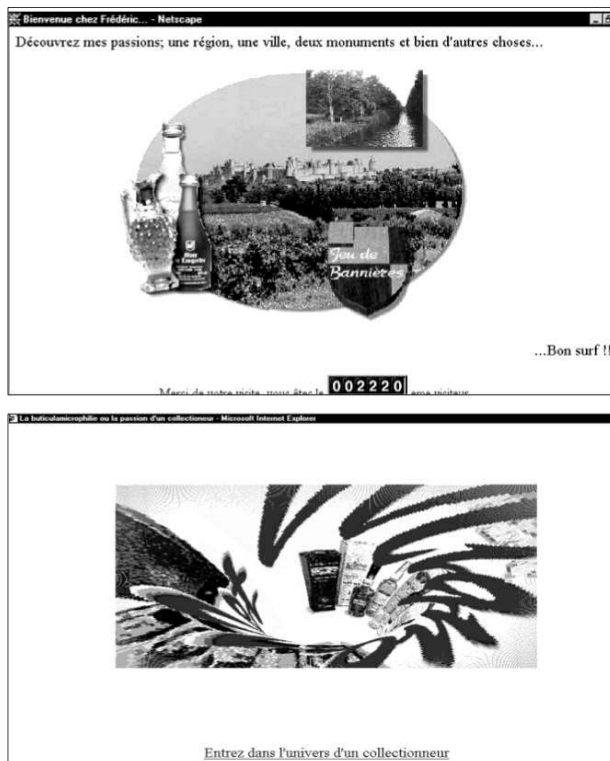


Figure 5. Page d'accueil d'un écrivain



## 5. Corrélation des traits : complexification des pages

- 38 À quels types de documents sont confrontés les visiteurs de sites personnels, quels types de pages s'affichent dans leur navigateur ?<sup>11</sup> Les sciences cognitives distinguent la mémoire de la méta-mémoire qui recouvre l'ensemble des procédures pour retrouver l'information. Dans un contexte de mémoire extériorisée sur le support numérique, la méta-mémoire correspond à toutes les ressources et les services qui peuvent être mobilisés pour accéder au contenu, les moteurs de recherche et annuaires étant les exemples les plus significatifs. Cette distinction nous paraît également pertinente pour les pages personnelles : nous pouvons en effet identifier dans notre corpus une ligne de partage entre les pages à contenu et les pages qui facilitent la navigation à l'intérieur du site. Dans l'ensemble de ces pages d'orientation (qui correspondent à 15 % des pages visitées), peuvent être distinguées : les pages de redirection qui pointent vers la nouvelle localisation du site (nous avons vu que cette pratique est loin d'être négligeable) ; les pages de menu qui donnent accès aux différentes rubriques du site (elles peuvent se présenter comme une page autonome ou être inscrites dans une page à contenu) ; les pages de listes qui regroupent des pointeurs vers d'autres pages du site. Ainsi une page d'un site de musique présente la liste des albums avec un lien vers chaque album ou bien, sur les sites pornographiques, les photo-vignettes présentées sous forme de mosaïque renvoient à des photos en grande taille.
- 39 Dans une perspective d'analyse des contenus sur le web, il est utile de pouvoir différencier ces pages de navigation des autres : elles se définissent davantage par leur fonction que par leur contenu, il est donc peu pertinent d'en analyser le contenu.

- 40 Du côté des documents ou pages à contenu, nous observons une assez grande différenciation dans la forme des documents, qui traduit des degrés d'élaboration très différenciés : 44 % des documents visités peuvent être considérés comme des pages élaborées, parce qu'elles spécifient le cadre de l'écriture hypertextuelle (présence de cadre, définition du fond de la page, choix de polices de caractère, de couleurs), présentent un ensemble de liens hypertextuels et articulent le texte et les images. À l'inverse, 28 % des documents présentent une syntaxe très simplifiée, avec peu de liens, peu d'images, peu de spécifications sur la mise en forme : ceux-ci peuvent être interprétés comme des brouillons peu travaillés. Cette distinction est liée à l'expertise et à l'engagement du concepteur dans l'animation de son site. Les experts exploitent toutes les possibilités de l'écriture hypertextuelle, les contenus textuels y sont plus riches et la relation au visiteur est souvent mise en scène. Les figures 6 et 7 ci-dessous donnent un exemple de pages très peu élaborées à côté d'une page qui s'apparente à un site professionnel pour la figure 8.
- 41 Le premier document montre une page avec peu de liens, peu d'images et un texte sans mise en forme particulière si ce n'est la taille du titre. La situation dans le site n'est pas très claire. À l'inverse, la seconde page organise l'accès à des documents en en présentant un résumé et un lien associé à une image qui pointe vers le sous-document. La mise en forme de la page est travaillée et l'utilisateur peut s'orienter dans le site grâce à la ligne des rubriques.

Figure 6. Page de contenu rudimentaire d'un site personnel

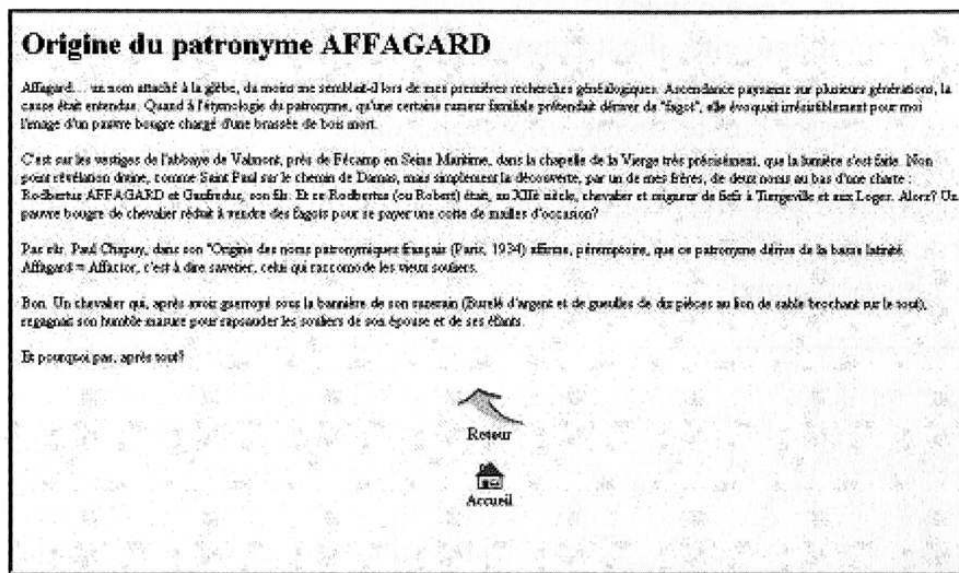
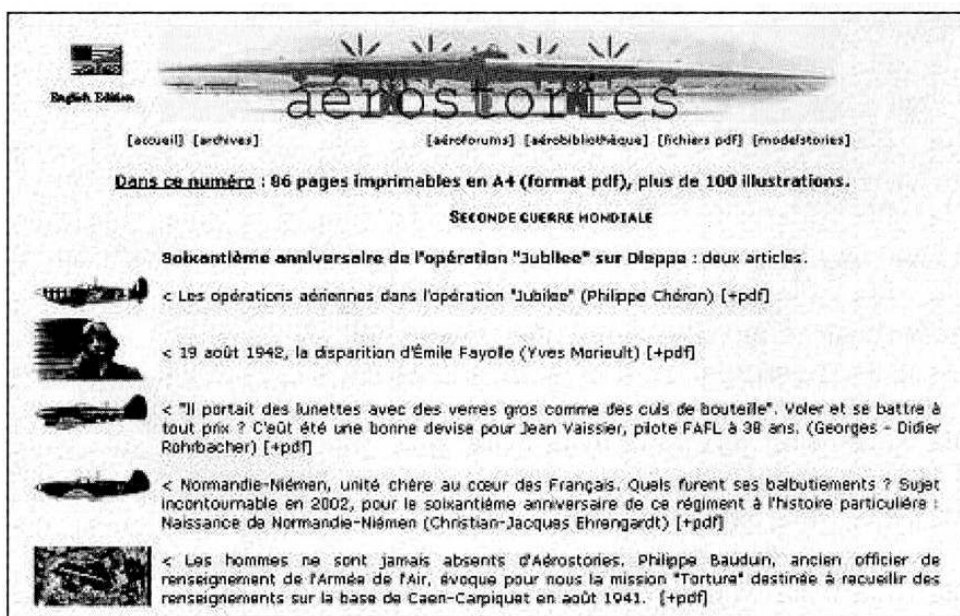


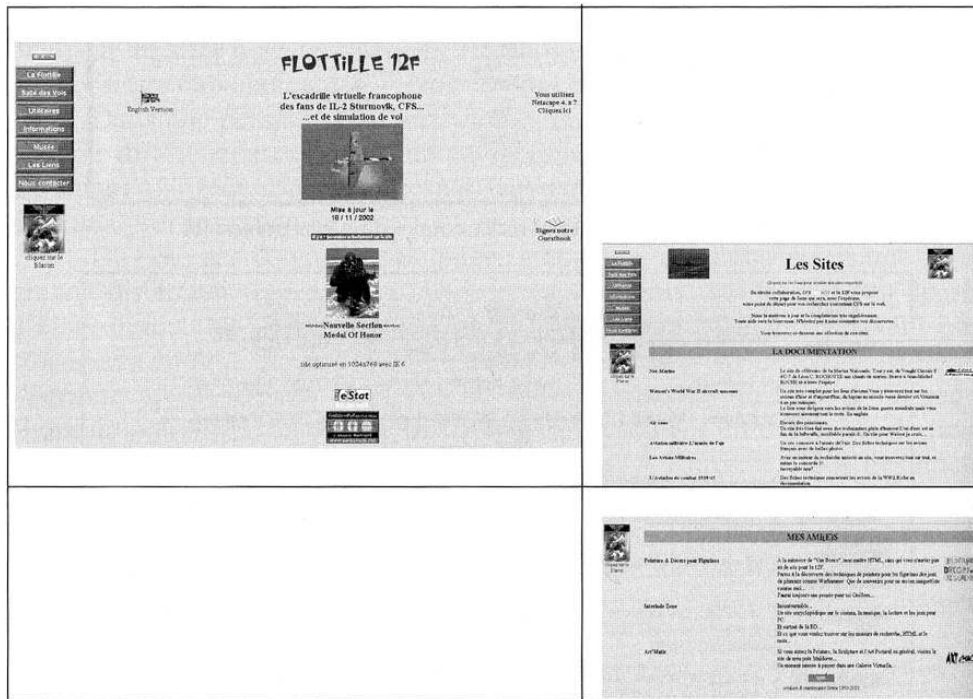
Figure 7. Page de contenu sophistiquée d'un site personnel



- 42 Au sein des pages les plus élaborées (44 % des pages visitées), deux postures éditoriales peuvent à nouveau être distinguées : sur certains documents, il y a une présence forte de l'émetteur et du récepteur (28 % des pages), la page se centre sur la relation en rendant présent le moi (fonction émotive) et en s'adressant directement au visiteur (fonction conative) ; tandis que dans d'autres documents (16 % des pages), les pronoms sont absents et le discours est centré sur le thème de la page. Quand la relation émetteur-récepteur n'est pas mise en scène, il y a au contraire une exploitation approfondie des propriétés de l'écriture hypertextuelle : le texte est moins bavard, mais il s'articule plus savamment aux images et aux autres documents du site (liens hypertextuels très développés).
- 43 Dans un même site, il est courant que la posture d'auteur change selon le type de page. Ainsi dans le site ci-dessous, la page d'accueil adopte un ton impersonnel centré sur le thème, alors que dans la page de liens il y a des commentaires personnels, voire intimes, sur les sites recommandés : ainsi, en bas du document de liens, trouve-t-on une rubrique « ami(e)s ». La vitrine du site garde une forme d'anonymat tandis que l'intérieur du site offre des espaces plus personnels.



Figure 8. Page d'accueil et page de liens d'un site personnel



- 44 La diversité des documents que l'on trouve sur la Toile, du moins dans l'espace des pages personnelles, tient à plusieurs éléments : aux spécificités de l'écriture hypertextuelle, qui rend nécessaire la présence de pages de navigation, au degré d'expertise des concepteurs de sites, qui fabriquent des documents plus ou moins élaborés, et enfin aux choix de représentation de la relation au lecteur sur le site, qui vont d'une posture anonyme, centrée sur le thème, à une mise en scène sophistiquée de la relation au visiteur, en passant par des représentations très narcissiques. Sur ce dernier point, ce qu'on observe sur internet n'est pas fondamentalement différent de ce que montre la littérature qui parcourt tout l'éventail de l'objectivité du récit (absence de référence au narrateur) jusqu'à la subjectivité du discours (présence marquée du narrateur) (Genette, 1969 : 61-69).
- 45 L'analyse des productions du web nous a conduits à mettre au point une chaîne de traitement des documents web, qui pourra être ré-exploitée dans d'autres contextes (pour d'autres types de sites et pages visitées, pour des documents web liés à des intranets...), et à identifier les types de documents que les internautes rencontrent dans leurs visites de pages personnelles. La spécificité de notre approche tient au fait que nous avons un regard sur la production du point de vue de la réception : nous étudions les documents web qui ont été visités.
- 46 Nous avons montré ici que les sites web personnels partagent une fonction de *terrain d'expérimentation* et constituent un lieu d'apprentissage. Les sites web personnels ont une fonction de *brouillon*, ils sont constitués d'écrits révisables à tout moment. Cet aspect particulier des écrits que l'on retrouve dans les sites personnels participe d'ailleurs d'un mouvement beaucoup plus général de mutation des actes de lecture/écriture liés à l'émergence des textes électroniques. La publication quasi instantanée de l'écrit qu'autorise la mise en ligne sur le web entraîne des modifications importantes dans la

configuration des actes de lecture et d'écriture, qui se voient étroitement entrelacés. Chartier (2000) souligne ici certains effets de cette mutation :

Dans le monde des textes électroniques ou, plus exactement, de la représentation électronique des textes, deux contraintes, tenues jusqu'ici comme impératives, peuvent être levées. Première contrainte : celle qui limite étroitement les possibles interventions du lecteur dans le livre imprimé [...] Si le lecteur entend, néanmoins, inscrire sa présence dans l'objet, il ne peut le faire qu'en occupant, subrepticement, clandestinement, les lieux du livre délaissés par l'écriture : intérieurs de la reliure, feuillets laissés en blanc, marges du texte, etc. Avec le texte électronique, il n'en va plus de même. Non seulement le lecteur peut soumettre le texte à de multiples opérations (il peut l'indexer, l'annoter, le copier, le démembrer, le recomposer, le déplacer, etc.), mais, plus encore, il peut en devenir le co-auteur. La distinction, fortement visible dans le livre imprimé, entre l'écriture et la lecture, entre l'auteur du texte et le lecteur du livre, s'efface au profit d'une réalité autre : celle où le lecteur devient un des acteurs d'une écriture à plusieurs voix ou, à tout le moins, se trouve en position de constituer un texte nouveau à partir de fragments librement découpés et assemblés [...] Le lecteur de l'âge électronique peut construire à sa guise des ensembles textuels originaux dont l'existence et l'organisation ne dépendent que de lui. Mais, de plus, il peut à tout moment intervenir sur les textes, les modifier, les récrire, les faire siens.

- 47 Enfin, nous avons passé ici sous silence la lourdeur des dispositifs techniques mis en place pour aboutir aux résultats présentés et les innombrables questionnements qui ont accompagné ce travail. L'objet – les pages et les sites de la Toile – est nouveau, leurs formes sont en transformation (surtout dans les pages personnelles), et il y a peu de modèles dominants. Tous les outils sont également à construire. Bien d'autres pistes restent à explorer, comme celle ouverte par Ivory et Hearst (2002), qui proposent une batterie de 157 traits formels et structurels pour différencier les mauvais, moyens et bons sites en prenant comme référence les sites récompensés par des prix. L'objectif est différent du nôtre puisque leurs travaux visent à aider les concepteurs à améliorer leur site, en reprenant les critères dominants, tandis que les nôtres cherchent à rendre compte des caractéristiques des objets qui sont effectivement visités par les internautes. Cependant, la démarche est très similaire puisqu'elle s'appuie sur des traits formels et structurels, plus nombreux et plus sophistiqués que les nôtres.

---

## NOTES

1. NetValue est une société de mesure d'audience sur internet : les données du panel ont été mises à notre disposition dans le cadre d'un partenariat entre France Télécom R&D et NetValue.
2. Le HTML, Hyper-Text Markup Language, est un langage pour coder les documents électroniques de type hypertexte (avec des liens) utilisés sur le web. Il permet de définir l'habillage d'un document, c'est-à-dire la façon dont il doit s'afficher à l'écran d'un navigateur. C'est un langage de balisage ou de marquage : un document HTML contient du texte brut et une série de marques ou balises (en anglais, *tags*) qui sont utilisées pour mettre en évidence la structure et le format du document : sa forme, sa taille, sa couleur ; le langage HTML permet également d'inclure des images, du son ou des animations dans le document.

3. Les travaux de Biber tentent d'identifier des types de textes en s'appuyant sur la répartition de traits morpho-syntaxiques très fins : les typologies obtenues ne coïncident pas forcément avec des genres, qui sont des formes de cristallisation de pratiques sociales. Il ne s'agit pas de retrouver des styles prédéfinis (narratif, descriptif, explicatif, argumentatif, poétique, etc.) mais de regrouper des documents (ou des portions de documents) en fonction de l'emploi qu'ils font de l'outillage grammatical (pronoms, temps et modes...) et de certains marqueurs lexicaux spécifiques (par exemple, types sémantiques d'adverbes : négation, possibilité, temps et espace... ). La classification des documents se fait donc sur la base de traits linguistiques fins articulant étiquetage grammatical et projection de dictionnaires spécifiques (classes sémantiques d'adverbes ou de conjonctions de subordination, par exemple).

4. Hypertoile : [http://www.telecom.gouv.fr/rnrt/projets/res\\_01\\_39.htm](http://www.telecom.gouv.fr/rnrt/projets/res_01_39.htm) (lien vérifié le 25/10/04).

5. Pour une présentation plus détaillée de la chaîne de traitement, voir Beaudouin *et al.*, 2001. Hypertoile : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb.htm>

6. Une balise HTML a en général l'allure suivante : `<TAG [ATTRIBUT = "VALEUR"]* >`, les crochets droits [] indiquent la présence potentielle d'un attribut pour le TAG en question, et le caractère \* indique la présence potentielle de plusieurs attributs : par exemple, `<FONT FACE = "Arial" SIZE = "2">` permet de définir « un marquage » de type police de caractères (FONT) dont un attribut (FACE) définit le nom de la police choisie (Arial) et un autre (SIZE) sa taille (2).

7. « Pour créer votre site internet », Hypertoile : [http://api.guide.free.fr/15\\_possibilites/possibilites.htm](http://api.guide.free.fr/15_possibilites/possibilites.htm)

8. Ce taux donne la différence entre les pages visitées par les panélistes et les pages non aspirées, les pages renvoyant un message d'erreur, les pages syntaxiquement mal formées et les pages hébergées chez Altern.org.

9. Pour la suite des traitements, nous avons éliminé les pages d'erreurs (8 363 pages) ainsi que les pages ne contenant aucune marque de personne et aucun lien, qui sont des pages vides ou des *frames* (3 725 pages). Le corpus de travail ne contient plus que 57 491 pages, soit une perte de 57 % par rapport à la liste initiale.

10. Accessible à : <http://www.archive.org/> (visité en juin 2002).

11. Nous avons construit une typologie de pages en utilisant comme traits de description : le nombre de liens internes et externes, de liens vers la boîte aux lettres, le nombre d'images, le nombre d'occurrences de mots sur la page, le nombre de pronoms de chaque personne, la présence d'une image en fond d'écran ou la définition d'une couleur de fond, la présence d'indications sur la police. Huit types de pages ont ainsi été identifiés. Les frontières entre catégories sont particulièrement floues et chaque classe doit plutôt être interprétée en terme de modèle ou d'idéal-type.

## RÉSUMÉS

Les pages personnelles sont des espaces de publication Web offerts par les fournisseurs d'accès ou par des portails à leurs clients ou visiteurs. L'adjectif « personnel » sous-entend que l'instance d'énonciation est un individu et que le contenu lui-même renvoie à la personne. Pour autant, peut-on considérer que les pages personnelles constituent un genre spécifique ? Nous avons montré, lors de travaux précédents, que les sites personnels se distinguaient des sites marchands par l'emploi des pronoms personnels et par la structure des liens hypertextuels. Par-delà ces

deux catégories de traits, y a-t-il d'autres éléments qui assurent l'autonomie du genre, outre le nom qui les désigne ? Pour le savoir, nous avons constitué un corpus de 100 000 pages personnelles visitées par une cohorte d'un millier d'internautes extraites d'un panel entre janvier et juin 2000. Nous montrons que les pages personnelles partagent une fonction de terrain d'expérimentation (elles ont une fonction de brouillon) et constituent un lieu d'apprentissage de l'écriture hypertextuelle qui est amené à évoluer.

## INDEX

**Mots-clés** : page personnelle, corpus Web, sites Web visités, présentation de soi, apprentissage, traits hypertextuels, genre

## AUTEURS

### VALÉRIE BEAUDOUIN

**Valérie Beaudouin** est chercheur dans le laboratoire de sciences sociales du centre de recherche de France Télécom au Département des Interactions Humaines/Usages Créativité Ergonomie (DIH/CE). Ses travaux portent sur les usages de l'internet et montrent l'articulation entre production et réception sur ce média. Les parcours sur l'internet, reposant sur des cohortes de plusieurs milliers de personnes, sont croisés avec la structure et les contenus des sites visités. L'approche quantitative est enrichie par des entretiens et observations auprès de concepteurs de sites, qui permettent de comprendre à quel point la conception-crédation de sites est guidée par les usages.

### SERGE FLEURY

**Serge Fleury** est maître de conférences en sciences du langage à Paris 3 et chercheur au Centre de lexicométrie et d'analyse automatique des textes (SYLED-CLAT, Paris 3). Ses travaux portent sur le traitement automatique du langage naturel avec comme champ d'étude : l'analyse automatique, la représentation des connaissances, l'acquisition de connaissances à partir de corpus, la programmation à prototypes, la réflexivité, les métaconnaissances, la normalisation des documents électroniques. Il participe également au développement du logiciel de lexicométrie « Lexico3 ».

### MARIE PASQUIER

**Marie Pasquier** est actuellement doctorante dans le laboratoire de sciences sociales du centre de recherche de France Télécom au Département des Interactions Humaines/Usages Créativité Ergonomie (DIH/CE).