Études de communication
langages, information, médiations

36 | 2011
Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus

Approche extensive des métadonnées pour un site web : principes d'élaboration et applications d'une taxonomie

An Extensive Approach to Website Metadata: Elaborating Principles and Uses of a Taxonomy

Nathalie Pinède et David Reymond

Édition électronique
URL : http://journals.openedition.org/edc/2645
DOI : 10.4000/edc.2645
ISSN : 2101-0366

Éditeur
Université Lille-3

Édition imprimée
Date de publication : 1 juin 2011
Pagination : 87-108
ISBN : 978-2-917562-05-5
ISSN : 1270-6841

Référence électronique
Nathalie Pinède et David Reymond, « Approche extensive des métadonnées pour un site web : principes d'élaboration et applications d'une taxonomie », *Études de communication* [En ligne], 36 | 2011, mis en ligne le 01 juin 2013, consulté le 01 mai 2019. URL : http://journals.openedition.org/edc/2645 ; DOI : 10.4000/edc.2645

Approche extensive des métadonnées pour un site web : principes d'élaboration et applications d'une taxonomie

An Extensive Approach to Website Metadata: Elaborating Principles and Uses of a Taxonomy

Nathalie Pinède et David Reymond

Introduction

1 La masse considérable de données informationnelles présentes sur le web rend toujours plus ardue la production de réponses discriminantes et pertinentes, dans une logique d'extraction facilitée de sens. Nous proposons ici une approche expérimentale appuyée par une méthodologie hybride pour caractériser les contenus des interfaces web et en produire une vision synthétique. Il s'agit de prendre en compte la dimension sémantique portée par les hyperliens de pages d'accueil de sites web, en postulant que les unités lexicales associées à ces hyperliens clefs (ou « textes passeurs » – d'après Jeanneret, 2000) portent une sémantique significative par rapport au site web concerné. Ces unités lexicales d'hypertexte (ULH), en tant que termes génériques d'accès à des contenus sous-jacents constituent de fait pour les usagers des métadonnées informatives vis-à-vis des contenus sous-jacents.

2 À partir de la collecte automatique d'ULH dans un corpus de sites web d'un même domaine (que nous appellerons « sites web organisationnels »), il est dès lors possible de générer une taxonomie où les UHL sont regroupées par classes en fonction de leur dominante thématique. Cette catégorisation peut-être réalisée en fonction des spécificités de l'activité organisationnelle, des caractéristiques de publication web ou encore des publics visés. La taxonomie ainsi constituée permet de mettre en évidence les dominantes

informationnelles et communicationnelles de la page d'accueil, sa signature, et *in fine* du site web lui-même (sur la base de cette représentation générique du contenu intégral du site à partir des ULH de la page d'accueil), et de produire des profils synthétiques à partir des différentes dimensions associées à la taxonomie (Reymond et Pinède, 2010a). Une autre utilisation possible de cette taxonomie s'inscrit dans le champ du classement automatique de sites web par comparaison de sa signature individuelle avec une signature de référence calculée de façon moyenne sur un corpus de sites web organisationnels (Reymond et Pinède, 2010b).

3 Dans cet article, après avoir développé notre postulat de départ ainsi que la démarche méthodologique globale d'élaboration de la taxonomie appliquée à un corpus de sites web d'universités, nous centrerons notre discussion sur la réalisation d'un outil semi-collaboratif en ligne d'enrichissement de notre taxonomie de référence et de génération dynamique de profils de sites web. Tout d'abord, il offre la possibilité d'une contribution collaborative pour enrichir la taxonomie et par voie de conséquence, son efficacité applicative. Ce dispositif offre aussi la possibilité de générer des profils dynamiques (sous forme de nuages de tags ou de radars) de sites web à partir des signatures : la projection de la taxonomie sur la page d'accueil du site web du domaine considéré. Enfin, il est également possible de mettre en évidence un degré de marginalité calculé à partir des poids relatifs des ULH. Cela permet d'évaluer des effets de normalisation terminologique ou au contraire, de marginalité dans le choix des ULH, pouvant nuire à la lisibilité et à l'accessibilité des contenus sous-jacents du site.

4 En conclusion, nous dégagerons les intérêts de cette caractérisation informationnelle synthétique qui offre un élément de compréhension complémentaire au plan de la sémantique du web grâce à cette agrégation terminologique.

Problématisation de notre démarche

Postulats initiaux

5 Le postulat initial sur lequel nous nous appuyons est que la page d'accueil d'un site représente une grande partie de l'information gérée par l'organisation ou la structure concernée mais traduit aussi des choix sélectifs et stratégiques, valorisant certaines catégories informationnelles plutôt que d'autres. L'hypothèse sous-jacente à notre démarche est que, s'appuyant sur les unités lexicales porteuses de liens structurants[1] sur les pages d'accueil d'un corpus étendu de sites Web pour un type d'organisation, nous pourrons en extraire les similarités et lister l'ensemble des termes et expressions plausibles pour la représentation de sous-contenus similaires.

6 Les auteurs de référence sur l'ergonomie des sites web (Nielsen et Tahir, 2002), définissent le rôle principal de la page d'accueil d'une organisation comme permettant de « véhiculer l'identité de l'entreprise, de mettre en évidence la valeur ajoutée du site par rapport à la concurrence et au monde réel et de présenter les produits et/ou services que fournit l'entreprise. Cette page serait en même temps comparable à la couverture d'un magazine, à une œuvre d'art [...], la table des matières d'un livre et la une d'un journal [...] ». Ce genre de page d'accueil peut-être distingué par l'existence sous-jacente d'une chaîne éditoriale : créée par une organisation, la page d'accueil doit répondre à certaines fonctions et posséder certaines fonctionnalités.

7 La page d'accueil est le point d'orientation du site et a pour mission la mise en valeur de l'information et des fonctions principales du site. Celle-ci se donne aussi à voir comme tout écran, à savoir comme « un espace unique à la surface duquel se cristallisent toutes les fonctionnalités de l'écriture. Confrontés à cet 'unimédia' saturé de signes et de codes, nous disposons de 'signes outils', de 'signes passeurs' qui nous donnent accès aux multiples modalités du texte » (Souchier *et al.*, 2003 : 23). La fonction d'orientation des usagers vers des zones informationnelles spécifiques est donc assurée par ces signes passeurs qui représentent l'ensemble des contenus des niveaux inférieurs (derrière l'hyperlien) et doivent donc être scrupuleusement choisis pour répondre aux contraintes issues de ces interfaces.

8 Dans une approche plus sémiotique de la page d'accueil « on peut distinguer trois éléments : le texte, le contexte (ce qui, en dehors du document affiché à l'écran, détermine les conditions de sa production et de sa réception) et le paratexte, défini provisoirement comme tout ce qui accompagne le texte à l'écran et est, de ce fait, susceptible d'en orienter la réception », avec comme postulat sous-jacent, le fait que « la co-présence d'éléments est génératrice de sens » (Dupuy, 2008 : 26). Même si nous focalisons dans un premier temps notre regard sur du paratexte hypertextuel et à dominante informationnelle, notre démarche globale est bien d'intégrer dans notre modèle final non seulement les éléments textuels présents sur la page d'accueil mais aussi les éléments contextuels (production/réception) afin de mieux comprendre la situation d'énonciation (les liens hypertexte) et donc de mieux qualifier les ressources proposées. Dans le cadre de notre recherche, s'appuyant sur une collecte automatique, nous avons choisi de restreindre[2] la notion large de « signe passeur » à celle de « texte passeur », qui correspond aux unités lexicales présentes sur la page d'accueil et porteuses d'un hyperlien permettant l'accès aux niveaux inférieurs du site.

Terrain d'étude

9 Par rapport à nos objectifs de recherche, nous avons choisi de travailler sur des corpus de sites représentant un certain type d'organisation : nous appellerons ces sites « sites web organisationnels » (SWO). Ce choix de constitution de corpus garantit un facteur de cohérence et homogénéité en termes d'activités, services et publics visés. D'une façon plus spécifique, nous nous sommes intéressés aux sites web des universités françaises.

10 En effet, la notion de site web pour une université reflète une réalité complexe et non unitaire, à l'image de l'organisation qu'elle incarne, découplée et décuplée au plan de ses composantes, services et missions. Parler de site web pour une université représente une commodité de langage mais non une réalité en termes d'objet unidimensionnel. Autour du site web institutionnel, portail d'entrée vers les différentes strates de l'université (en termes de structures/services) et identifié par l'adresse DNS principale correspondant au serveur web de l'organisation, gravite une grappe de sites web que l'on identifiera comme étant rattachés à cette organisation par leur nom de domaine (Reymond *et al.*, 2007)[3].

11 Ainsi, si l'on regarde de plus près les universités de Bordeaux 1 (Sciences) et Bordeaux 2 (Médecine – Sciences sociales) qui appartiennent à notre corpus d'étude, on peut relever[4] respectivement 96 et 121 sites relevant des zones DNS .u-bordeaux1.fr et .u-bordeaux2.fr, c'est-à-dire appartenant au même ensemble organisationnel et dépendant de la même responsabilité éditoriale (celle du président d'université). Si cet ensemble d'éléments juxtaposés, parfois hyper-reliés entre eux (sans que cela soit pour autant systématique),

relève d'une seule instance, on peine toutefois à voir émerger une image forte et cohérente de l'institution, tant les logiques d'édition et de communication peuvent s'avérer singulières et fragmentées. Il est cependant important de dispenser une image institutionnelle cohérente permettant de trouver une véritable audience aux plans individuel et collectif. Et dès lors, le site web d'un établissement universitaire prend une dimension stratégique indéniable (Reymond *et al.*, 2007).

12 Par ailleurs, ainsi que le rappelle Sylvie Lainé-Cruzel, « le site de l'université se doit d'être à jour : renseigner d'une manière efficace, présenter des formations et des équipes existantes, des annuaires exploitables, etc. Sa fonction est d'être utile et de rendre des services. Sa nature est d'être *évolutive*, pour restituer une image aussi fidèle que possible d'un univers en transformation permanente. La qualité du site sera liée à sa capacité à évoluer en même temps que l'univers sur lequel il informe » (Lainé-Cruzel, 2004 : 111). À ce titre, il est caractérisé en tant que *ressource*, s'inscrivant dans une double logique de médiation et d'usage.

13 Nous proposons dès lors d'élaborer un référentiel terminologique, qui s'appuie sur la réalisation d'une taxonomie produite à partir des contenus édités, essentiellement les menus de navigation des pages d'accueil. Ces derniers sont le résultat d'un processus de catégorisation réalisé par les éditeurs des sites, intégrant les contraintes des écrits d'écran (Jeanneret, 2000) et constituent une source d'unités lexicales clés utiles à la description des contenus diffusés par le site. On considérera que celles-ci véhiculent des significations en tant qu'information éditée et que cette information, du fait de son statut de liens, représente un nœud stratégique, tant pour la représentation des fonctions organisationnelles que pour l'accès aux sous-strates du site. Et c'est à ce titre que nous pouvons leur accorder un statut de métadonnées, en tant que meta description des contenus internes au site.

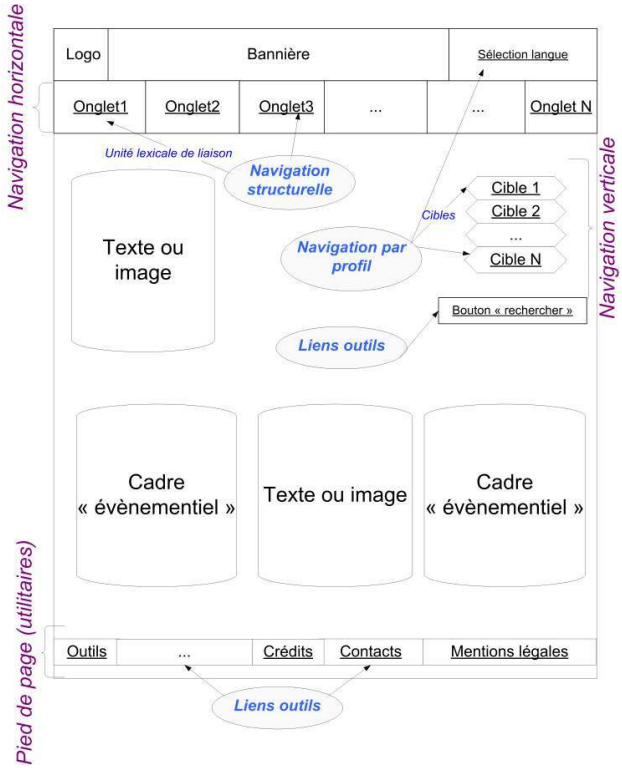
Mise en œuvre de la taxonomie

Principes d'analyse des pages d'accueil

14 Les pages d'accueil de chacun des sites identifiés[5] ont été analysées (« informationellement » déconstruites) selon une méthode d'approche dérivée des principes de construction des pages d'accueil de Nielsen à savoir une décomposition en trois zones de navigation et d'action : la navigation structurelle, la navigation par profil des cibles des sites et les outils d'aide à la navigation. Le modèle abstrait de la figure 1 permet de présenter génériquement les contenus d'une page d'accueil d'un site web.

15 Les outils de collecte automatique actuels ou la version de langage HTML utilisée ne permettant pas encore un repérage d'ordre « sémiotique », spatialisé, des termes ou hyperliens, il était nécessaire dans une première phase d'en passer par une répartition manuelle des différentes unités lexicales selon les trois classes de navigation établies par Nielsen.

Fig. 1 : Gabarit d'une page d'accueil



16 Nous avons donc travaillé à partir de ces trois classes de navigation[6] :

1. Les ULH donnant un accès structurel aux contenus des zones profondes des sites (en navigation horizontale dans le schéma),
2. Les ULH représentant la navigation par profil, soit une recomposition de la vue du site pour les cibles de ces derniers (en navigation verticale dans le schéma),
3. Les ULH correspondant aux liens de type outils qu'ils soient aides à la navigation, à l'appropriation des interfaces, à des procédures de connexion, etc. (en général, dans les en-têtes et pied de page).

17 Par la suite, les hyperliens des pages d'accueil des universités de Bordeaux 2 et Bordeaux 3 ont été collectés automatiquement, et les ULH associées non encore répertoriées sont venues compléter les listes terminologiques des différentes classes déterminées sur la base de l'analyse des sites de Bordeaux 1.

Principes de catégorisation des unités lexicales hypertextuelles (ULH)

18 Au final, après un premier travail de segmentation par classes de navigation, définition des catégories et recensement distribué des ULH, on obtient une catégorisation thématisée. Le tableau suivant reprend les classes choisies ainsi que le nombre d'ULH classées et recensées pour chaque type de navigation.

Tableau 1 : Liste des classes d'ULH recensées par classes de navigation

Navigation Structurelle	Navigation Outils	Navigation Profils
Accueil – Présentation (23)	Accès géographique (15)	Profil anglais (7)
Actualités (51)	Accès Web (8)	Profil français (5)
Culture – Loisirs (42)	Accès contenu (24)	Profil chercheur (2)
Formation (139)	Aide (5)	Profil entreprises (3)
International (12)	Annuaire (4)	Profil étudiants (6)
Logistique – Équipement (18)	Authentification (22)	Profil lycéens (4)
Partenariats/Transfert-Valo (25)	Contacts (14)	Profil utilisateurs autres (14)
Recrutement (12)	Technologies (32)	
Recherche (121)	Liens (5)	
Ressources documentaires (79)	Mentions légales (7)	
Services à authentification (10)	Outils Pull-Push (17)	
Structures transversales (38)	Fonction rechercher (8)	
Vie étudiante (7)		

19 Le choix de classes a fait l'objet d'ajustements successifs. Il est évident que toute tentative de classification représente, quelle que soit la volonté d'objectivation initiale, un choix sélectif permettant de regrouper des termes et expressions selon un plan d'affinité sémantique cohérent. Même si s'imposent spontanément des regroupements, de nombreuses questions émergent en permanence pour affecter les ULH, souvent ambivalentes, à telle ou telle classe. En l'occurrence, le contexte d'action (l'organisation universitaire) ainsi que la finalité d'application associée à la taxonomie ont permis de construire le cadre d'opérationnalisation de la taxonomie, en s'appuyant sur les fonctions, services et activités de l'université, ainsi que sur les potentialités/contraintes de publication via le média web.

20 Les unités lexicales utilisables recueillies sur la base des critères précédemment définis recouvrent un répertoire où coexistent standardisation et hétérogénéité : l'on rencontre aussi bien des expressions (« qui sommes-nous ? »), des abréviations familières (« Cult. Sctif »), que du vocabulaire plus normalisé. Toutefois, cette diversité représente une réalité, celle de la communication éditoriale proposée par chaque site, que nous reprenons de façon exhaustive dans notre recensement. C'est bien à partir de l'information telle qu'elle se donne à lire sur les différentes pages d'accueil des sites, à partir de l'information éditée telle qu'elle communique sur la structure ou le service concerné que nous nous appuyons pour bâtir notre taxonomie. Même si nous nous situons au niveau des hyperliens, soit à un niveau de représentation synthétique d'une information visant à donner accès à un ensemble de ressources et/ou productions situées à un niveau N-1 du site web, on se trouve en prise avec des « pratiques langagières » (Aussenac-Gilles *et al.*, 2004 : 77). À ce titre, les ULH peuvent être assimilées à des marqueurs d'usages discursifs des éditeurs/rédacteurs des pages d'accueil et considérées comme des vecteurs supposés de signifiance pour les utilisateurs du site web. Toutefois,

on remarque des récurrences terminologiques, tant par la nature intrinsèque de la page d'accueil que par la cohérence du contexte organisationnel (Rouquette, 2009), qui feront ultérieurement l'objet d'analyses détaillées.

21 La classe relative à la navigation structurelle est celle ayant le plus de poids au plan du nombre d'ULH (cf. tableau 2). Ceci ne constitue pas une surprise car cette classe de navigation décline les différents plans de l'organisation concernée et représente des dimensions informationnelle et communicationnelle. À l'intérieur de cette classe, les catégories « formation » et « recherche » sont celles ayant la plus forte représentation lexicale (avec respectivement 139 et 121 ULH recensées à ce jour) ce qui se révèle en adéquation avec les deux missions phare de l'organisation universitaire. Vient ensuite la rubrique « Ressources documentaires » avec 79 unités lexicales, ce qui là aussi se révèle cohérent avec la réalité organisationnelle, la documentation étant d'une part un service considéré comme public, important pour les publics des enseignants-chercheurs et étudiants et la dimension numérique ouvrant d'autre part des possibilités d'accès à des ressources et des services en ligne multiples et adaptées aux différentes spécialités.

Tableau 2 : nombre total d'ULH recensées par classes de navigation sur la base du corpus des universités de Bordeaux 1, 2 et 3

Navigation	Structurelle	Outils	Profils
Nombre total d'ULH	577	161	41

22 La classe de navigation « Outils » présente un panel d'ULH assez important, recouvrant des fonctions pratiques diversifiées et illustrées par les catégories recensées à l'intérieur de cette classe. Quant à la classe de navigation « Profils », elle est assez faiblement représentée en termes de nombre d'unités lexicales. Mais c'est aussi, contrairement aux deux autres et notamment à la classe de navigation structurelle, celle qui est le plus amenée à évoluer au niveau de sa représentation par catégories. Seuls ont été traités ici les profils rencontrés sur la base de notre corpus test : si les profils d'usagers sont relativement complets à l'issue de ce premier balayage, il n'en est pas de même pour les profils linguistiques qui sont évidemment appelés à être complétés en intégrant d'autres cibles linguistiques (espagnole, italienne, allemande, etc.) selon la zone géographique des établissements.

23 Actuellement, toutes les ULH recensées ne peuvent être intégrées à la taxonomie : certaines d'entre elles ont été rejetées car étant inopérables et inclassables (« âme du bâtiment »...). Les autres, ULH dites de « spécialité »[7] et ULH « ambivalentes »[8] ne sont actuellement pas classées mais pourront être intégrées ultérieurement, par renvoi sur des référentiels de spécialités ou analyse contextuelle des ULH. Cela réduit évidemment le pouvoir de recouvrement[9] de la taxonomie sur la page d'accueil, mais c'est en l'état présent d'avancée de nos travaux une limite incontournable.

24 Toutefois, le pouvoir de recouvrement de la taxonomie, malgré les restrictions actuelles, reste satisfaisant et surtout, encourageant pour poursuivre dans cette voie. Sur une sélection aléatoire de huit universités françaises (soit au total 589 sites web), une moyenne globale de 68 % au plan du recouvrement par notre taxonomie a été obtenue.

Les récurrences lexicales observées démontrent les régularités de structuration informationnelle et éditoriale sur les pages d'accueil des sites web de type universitaire et permettant de valider notre modèle d'investigation et d'analyse. Ces tests valident notre hypothèse initiale, à savoir la possibilité d'utiliser les signes passeurs lexicaux représentés par les ULH comme marqueurs caractéristiques de pages d'accueil, voire de sites web, transférables sur des situations organisationnelles similaires.

25 Un autre intérêt de la taxonomie réalisée réside dans sa flexibilité. Si les classes sont stables, il est par contre possible d'en varier les regroupements pour mettre l'accent sur certaines dimensions. On peut ainsi changer les vues et perspectives sur le site web en ciblant sur des dimensions particulières, comme « Activités », « Outils », « Profils », etc., et ce, par l'agrégation des classes adéquates.

Enrichissement de la taxonomie : fonctionnalités d'un dispositif en ligne collaboratif

Principes généraux de fonctionnement

26 La taxonomie réalisée à partir du corpus de sites web d'université se donne ainsi à voir sous forme de liste de classes dans le prototype du dispositif en ligne de gestion et utilisation de la taxonomie. La figure 2 représente un extrait de la partie navigateur web et montre l'interface d'accès du prototype.

Figure 2 : interface du dispositif en ligne présentant la liste des classes avec les ULH associées

27 À gauche, la rubrique « Classes » recense toutes les classes qui ont été générées à partir du corpus d'ULH des sites web d'universités. Au centre, la rubrique « Unités lexicales » permet d'afficher la liste des ULH d'une classe sélectionnée, en l'occurrence dans cet exemple, la classe « International » avec 14 ULH à l'heure actuelle. Enfin à droite, apparaît la liste des derniers ajouts d'ULH validés.

28 Ce prototype de dispositif en ligne pour la taxonomie offre de nombreux avantages et possibilités. Tout d'abord, il permet une gestion collaborative de la taxonomie. Une des difficultés majeures réside en effet dans le contrôle de la dynamique des ULH. Les sites web évoluent, leurs pages d'accueil et les termes choisis pour thématiser le contenu aussi. Or, il est indispensable que la taxonomie soit le plus exhaustif et le plus fidèle possible. Plusieurs pistes sont envisagées : idéalement, l'objectif serait de passer d'un mode manuel d'enrichissement à un mode automatisé s'appuyant sur des référentiels, langages, dictionnaires, ontologies appropriés, afin de tendre vers une représentation optimisée et fiable du système de marqueurs lexicaux d'un type d'organisation (ici, les universités).

29 Cependant, plutôt qu'une alimentation manuelle nécessitant un protocole lourd et laborieux, nous suggérons de déporter cette tâche en la répartissant sur les responsables éditoriaux tout en la régulant par un contrôle expert. Ainsi, pour la construction d'une nouvelle taxonomie ainsi que sa maintenance, nous proposons un dispositif en ligne offrant les services utiles au contrôle de l'alimentation et à la maintenance d'une taxonomie, depuis la structuration des données en format ouvert jusqu'à la gestion des flux de données. Le dispositif prend en charge les interfaces permettant d'alimenter et de gérer les thèmes et les contenus d'une taxonomie ainsi que ses usagers. Afin de bénéficier d'une utilisation généralisée, le système s'appuie sur une technologie Web. Ceci nous permet d'inviter des contributeurs potentiels à alimenter eux-mêmes les contenus d'une ou plusieurs taxonomies.

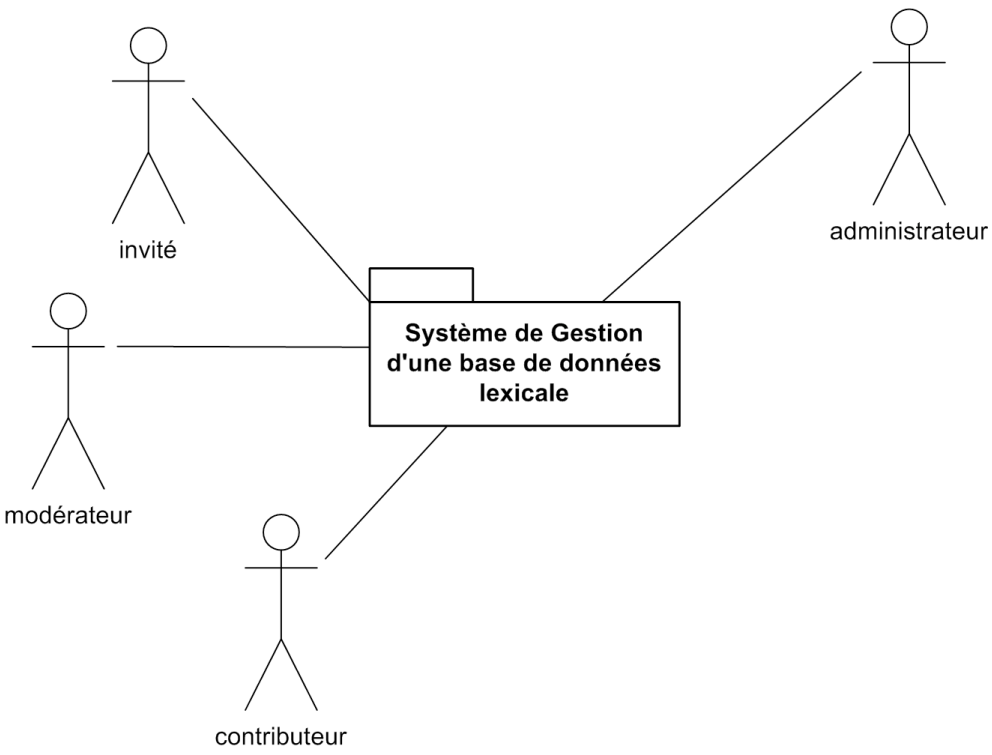
30 L'autre intérêt de ce dispositif en ligne sera de proposer aux usagers collaborateurs des applications[10] permettant par exemple de :

- choisir une unité lexicale pour les menus de son site à partir de la taxonomie en ligne ou encore de mesurer la marginalité des ULH choisies,
- visualiser les résultats de l'application de la taxonomie pour la création du profil de la page d'accueil de sites (Reymond et Pinède, 2010a).

Gestion des rôles dans le dispositif en ligne

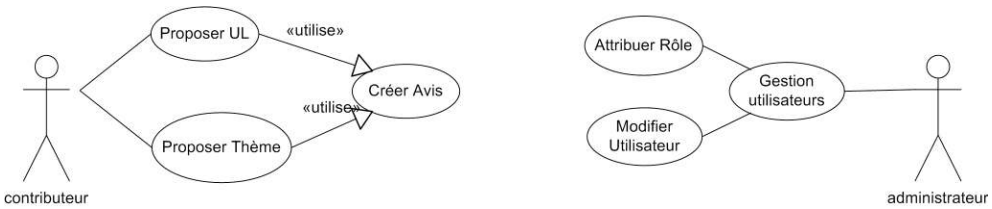
31 La mise en place d'un dispositif collaboratif en ligne pour l'enrichissement des contenus de la taxonomie suppose une gestion des rôles clairement définie. La figure 3 représente les quatre rôles du dispositif : l'administrateur pour la partie gestion des usagers et des taxonomies, les trois autres rôles associés à une ou plusieurs taxonomies, depuis l'invité (un internaute quelconque), le contributeur et le modérateur.

Figure 3 : différents rôles des usagers du dispositif de gestion et alimentation de la taxonomie



32 La figure 4 montre les cas d'utilisation du dispositif. À droite de la figure, l'utilisateur *administrateur*, responsable de la gestion peut attribuer des droits et des rôles aux différents utilisateurs. Un usager (invité par courriel, ou simple curieux motivé) envoie un courriel à l'administrateur pour solliciter une utilisation du dispositif. Le choix de sélection des candidats est réalisé en fonction de leur rôle d'éditeurs de contenus web dans le domaine de la taxonomie en construction. L'administrateur octroierait par exemple le rôle de contributeur à cet usager.

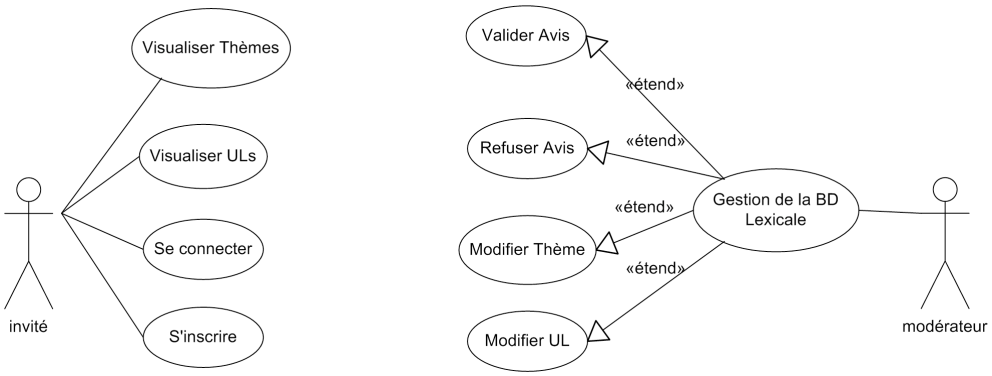
Figure 4 : cas d'utilisation de deux rôles spécifiques (Contributeur/Administrateur)



33 À droite, l'administrateur du dispositif qui gère les rôles en changeant les propriétés des utilisateurs du système. À gauche, les actions possibles des contributeurs : proposer un thème nouveau ou une unité lexicale. Ces deux possibilités sont un avis puisque ces propositions sont soumises au modérateur.

34 Enfin, la figure 5 montre les différentes actions des usagers invités et modérateurs. Les usagers peuvent s'inscrire, visualiser les thèmes de la taxonomie, puis au sein de chaque thème les unités lexicales qui le composent. Ils peuvent en outre proposer leur inscription pour accéder au rôle de contributeur. Le modérateur a le rôle de gérant des différents avis émis par les contributeurs, en acceptant ou refusant les propositions. Chaque rejet est accompagné d'une motivation. Il peut également simplement amender la taxonomie aux deux niveaux : unités lexicales ou thèmes.

Figure 5 : actions possibles des rôles « invités et modérateurs »



35 Ce système de rôles élaboré, avec modération à la clef, est nécessaire au bon fonctionnement de l'ensemble.

Applications offertes via le dispositif en ligne

36 Le dispositif en ligne proposera plusieurs services appuyés sur la taxonomie, permettant à un responsable éditorial de site web, non seulement de participer à l'enrichissement des contenus de la taxonomie sur la base de ses choix lexicaux propres, mais aussi de pouvoir

comparer ses choix personnels avec ceux réalisés dans le même domaine. Il peut aussi avoir une vue de son site web tel qu'il se donne à lire à travers les marqueurs informationnels de sa page d'accueil.

Régularités terminologiques et degré de marginalité

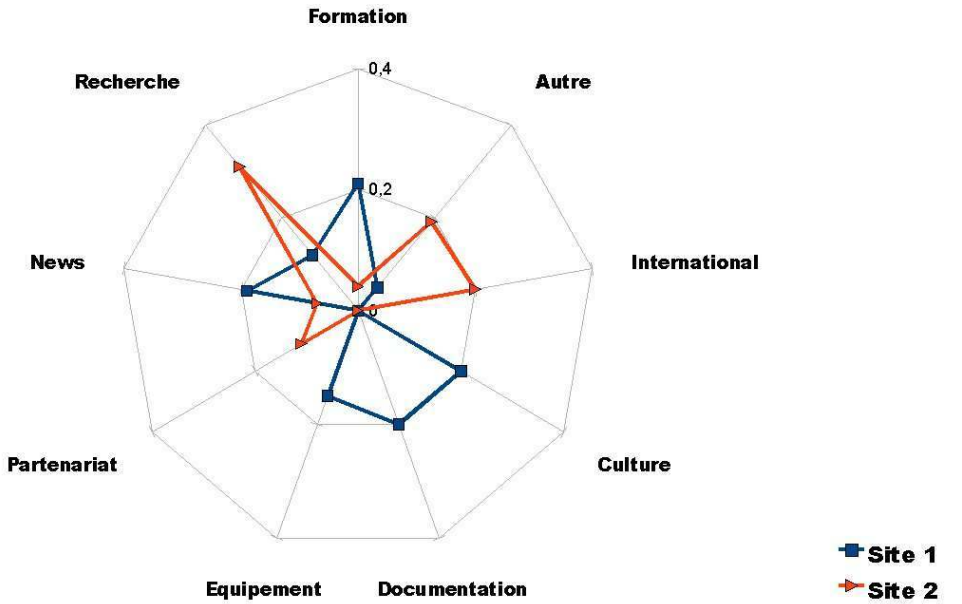
37 Il est possible de calculer pour un corpus établi (la liste des sites ayant servi à la construction de la taxonomie) les occurrences de chaque unité lexicale sur la totalité du corpus et ce pour chaque thème donné. Ainsi, chaque ULH est associée à son occurrence dans le corpus, ce nombre constituant sa fréquence (popularité) dans le corpus donné. En rapportant à la somme de la popularité de chaque ULH de la taxonomie nous construisons son taux d'apparition. Réciproquement, pour une page du corpus, son degré de « marginalité » se définira comme le choix d'unités lexicales peu fréquentes pour composer la page d'accueil et en conséquence comme la somme des popularités de chaque UL la composant.

38 Sur la base des choix lexicaux qui auront été faits pour qualifier le contenu, l'objectif est de pouvoir déterminer les similitudes/écarts (homogénéité/marginalité/originalité) d'un site par rapport à d'autres sites du domaine. À un niveau macro, ces calculs de fréquence et de popularité permettent ainsi de mettre en évidence des phénomènes de normalisation naturelle (Rouquette, 2009) dans les choix de termes ou d'unités lexicales lors de la composition d'un page web. Ce choix conduit à un lissage informationnel et à des effets de masse, par rapport auxquels des écarts en termes d'usages terminologiques peuvent nuire à la visibilité du site et à son accessibilité (par exemple, « com presse » plutôt que « communiqué de presse » ou encore « cult. Sctif » plutôt que « culture scientifique »...). À un niveau micro, cela peut donc être un outil d'aide à une « régularisation » lexicale, afin d'atténuer l'effort cognitif des usagers pour la compréhension des menus de navigation et augmenter sa visibilité. Cela peut aussi constituer un outil d'aide à la décision en matière de termes de référence pour la qualification de contenus, voire un outil d'appui à l'originalité, dans un souci de se démarquer par rapport à des sites similaires.

Visualisation de profils informationnels

39 À ces éléments de gestion contributive se rajoute la possibilité de visualiser les profils informationnels d'une page soumise au dispositif. Celui-ci se charge alors d'aller collecter les ULH de la page, les « textes passeurs », puis les compare avec les unités lexicales de la taxonomie. Chaque expression retrouvée augmentera le nombre d'occurrences du thème associé par la taxonomie. La représentation obtenue est alors une représentation vectorielle de la page, représentation classique en traitement automatisé des documents, dans l'espace informationnel de la taxonomie (les thèmes). Ce dernier point, que nous qualifions par l'ensemble des éléments informationnels essentiels fournis par le domaine auquel s'applique la taxonomie, assure une description homogène de la page soumise. La visualisation s'effectue alors sous forme de radars (fig. 6) indiquant l'importance de chacun des thèmes de la taxonomie, ou dimensions vectorielles de la représentation.

Figure 6 : Profil de deux sites web dans la dimension « Structure » de la taxonomie



40 L'un des intérêts possibles de cette représentation synthétique des contenus web est de permettre aux éditeurs de sites web d'obtenir un résumé synthétique de leur site tel qu'il se donne à voir/lire à travers les choix terminologiques opérés. La production d'une synthèse sous forme de tableau de bord permet de juger de l'adéquation de la mise en visibilité web avec la réalité des activités et domaines de prédilection de sa structure, et à ce titre, constitue un outil d'ajustement stratégique.

Conclusion

41 Nombreuses sont les potentialités ouvertes par cette caractérisation informationnelle. Cette méthodologie s'appuyant sur les contenus web édités d'accès aux sous-strates d'un site permet de projeter une vue multidimensionnelle du site web et favorise de fait la production de sens dans un contexte organisationnel donné. Cette homogénéité de corpus constitue la garantie nécessaire pour, au-delà des hétérogénéités syntaxiques constatées sur les pages d'accueil, établir des constantes, représentations et classifications significatives de ces ensembles organisationnels. Ce postulat initial permet dès à présent d'envisager des translations vers d'autres corpus de SWO, comme les sites de collectivités locales et territoriales.

42 Les applications dérivées de cette méthode sont multiples, par exemple en termes d'analyse stratégique, d'optimisation de la lisibilité par lissage terminologique ou encore de gestion collaborative de contenus. D'autres travaux sont nécessaires notamment pour étendre les domaines applicatifs (comparaisons interculturelles ou linguistiques) et améliorer l'organisation des unités lexicales en thésaurus ou ontologies pour tendre vers les objectifs du web sémantique. Ce travail d'agrégation lexicale contribue dès lors à une réduction du bruit et participe de fait, non seulement à des pratiques éditoriales rénovées, en phase avec des éléments d'ordre stratégique, mais propose aussi une nouvelle forme de représentation sémantique sur le web.

BIBLIOGRAPHIE

Aussenac-Gilles, N. et, Condamines, A., (2004), *Documents électroniques et constitutions de ressources terminologiques et ontologiques*, in Information-Interaction-Intelligence, vol. 4, n° 1, novembre, pp. 75-93.

Dupuy, J.-P., (2008), *Structure de la page Web : texte et paratexte*, in Revue des interactions humaines médiatisées, vol. 9, n° 1, pp. 25-42.

Jeanneret, Y., (2000), *Y a-t-il (vraiment) des technologies de l'information ?*, Villeneuve d'Asq, Presses Universitaires du Septentrion.

Lainé-Cruzel, S., (2004), *Documents, ressources, données : les avatars de l'information numérique*, in Information-Interaction-Intelligence, vol. 4, n° 1, novembre, pp. 105-119.

Nielsen, J. et Tahir, M., (2002), *L'art de la page d'accueil*, Marsat, Eyrolles.

Reymond, D., Rebai, B.-K. et Bordier, A. E., (2010), « AnCaraS : logiciel d'analyse de l'édition web », *Actes du 6e colloque international VSST (Veille Stratégique Scientifique et Technologique)*, octobre, Toulouse.

Reymond, D., Pinède-Wojciechowski, N. et Vieira, L., (2007), *Vers une instrumentation qualitative d'évaluation des dispositifs web : corrélations entre les discours de la gouvernance universitaires et les traces numériques*, in Actes du colloque international EUTIC 2007, Médias et diffusion de l'information, vers une société ouverte, Université d'Athènes, 7-10 novembre.

Reymond, D. et Pinède, N., (2010a), *Website and communication strategy alignment : a librarian science approach to webometrics tools*, in Sixth International Conference on Webometrics Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting, 19-22 October, Mysore, India.

Reymond, D. et Pinède, N., (2010b), *Using a taxonomy based fingerprint : classification and recognition of the academic Webspace*, in Sixth International Conference on Webometrics Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting, 19-22 October, Mysore, India.

Rouquette, S., (2009), *L'analyse des sites internet. Une radiographie du cyberesp@ce*, Bruxelles, De Boeck, INA.

Souchier, E., Jeanneret, Y. et Le Marec, J. (dirs.), (2003), *Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés*, Paris, BPI, 349 p.

NOTES

1. Définis comme l'ensemble des liens internes pointant sur des zones profondes du même site.
2. Restriction en adéquation avec les recommandations d'accessibilité du W3C, puisque tout « signe » d'écran Web autre que textuel et porteur de sens doit être augmenté d'une balise textuelle ALT le définissant.
3. Ce corpus n'intègre pas les composantes de l'université ayant choisi une inscription de leur site web hors DNS de l'organisation mère...

4. En septembre 2009, et au sens DNS seulement. D'autres sites existent probablement mais ne sont pas associés à la zone DNS (en .com ou labo.fr par exemple).
5. Des requêtes des zones DNS des universités concernées puis le lancement du collecteur AnCaraS (Reymond *et al.*, 2010) depuis le site principal (www) ont permis de déterminer un corpus initial qui s'est réduit ensuite à un certain nombre de sites utilisables. Plusieurs raisons à cela : erreurs 404, redirections, problèmes d'accessibilité...
6. Si ce découpage peut-être réalisé avec une relative facilité pour les pages d'accueil des sites principaux il n'en est pas toujours de même pour les autres sites de la zone DNS dont les structurations éditoriales se révèlent pour le moins hétérogènes.
7. Par exemple, « nanostructures organiques ».
8. Par exemple, « programme », « équipe »... de formation ? de recherche ?
9. Défini à l'échelle d'une page ou d'un corpus (par sa moyenne sur l'ensemble des pages le composant) comme le nombre d'ULH de la page présente dans la taxonomie divisée par le nombre d'ULH de la page.
10. Les applications seront accessibles à partir d'un seuil minimum de 60 % de recouvrement de la taxonomie par rapport à une page quelconque du domaine. Ce seuil est à la fois le garant de la représentativité de la taxonomie pour l'URL donné et réciproquement un indice de stabilité de la taxonomie en regard des éditions web de son domaine.

RÉSUMÉS

Nous proposons ici une approche expérimentale appuyée par une méthodologie hybride (quantitative et qualitative) pour caractériser les contenus des interfaces. Nous générons une taxonomie en utilisant la dimension sémantique portée par les hyperliens de pages d'accueil de sites web dont les unités lexicales associées ouvrent l'accès aux contenus sous-jacents et constituent de fait des métadonnées informatives pour les usagers des sites. Après avoir présenté notre démarche méthodologique appliquée à un corpus de sites web d'universités, nous proposerons un dispositif semi-collaboratif en ligne d'enrichissement de notre taxonomie et de génération dynamique de profils de sites web.

We define an experimental approach using qualitative and quantitative methods to characterize Web content interfaces. Taking into account the semantic dimension hosted by the homepage's anchor text, we organize the entire set obtained from a corpus of university website domains into a taxonomy. Such lexical units can also be considered metadata as they refer to the inner content of the sites for their users. Showing the meaning produced using the taxonomy as an informational space for synthesizing the content of the home pages, we suggest a collaborative platform for administering our taxonomy and generating website fingerprints and profile views.

INDEX

Mots-clés : hyperlien, page d'accueil, profil informationnel, site web, taxonomie
Keywords : hyperlink, homepage, informational profile, taxonomy, website

AUTEURS

NATHALIE PINÈDE

Laboratoire MICA-GRESIC – Université de Bordeaux

Nathalie Pinède, Laboratoire MICA-GRESIC, Université de Bordeaux, IUT Michel de Montaigne. Ses axes de recherche concernent les logiques informationnelles autour du numérique, notamment la problématique de la construction du sens à partir de marqueurs lexicaux sur les sites web et l'étude de l'articulation entre stratégies organisationnelles et formalisation via le web. Adresse électronique : nathalie.pinede@iut.u-bordeaux3.fr.

DAVID REYMOND

Laboratoire MICA – GRESIC – Université de Bordeaux

David Reymond, Laboratoire MICA-GRESIC, Université de Bordeaux, IUT Michel de Montaigne. David Reymond développe de nouvelles approches en webométrie par l'utilisation de méthodes hybrides qualitatives et quantitatives de description et d'interprétation des éditions et applications hypertexte et de leurs usages. Les outils et méthodes développés s'inscrivent dans les disciplines de l'informétrie. Adresse électronique : david.reymond@iut.u-bordeaux3.fr.