Terminologie hypertexte : dynamique temporelle d'une taxonomie

Nathalie PINEDE
Université de Bordeaux, MICA EA4426, MSHA, France
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux, France

David REYMOND
Université du Sud, Toulon-Var, I3M EA 3820, La Garde, France
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux , France

Benoit LE BLANC
Université de Bordeaux, IPB, IMS (UMR5218), Bordeaux, France
ISCC, CNRS, 20 rue Berbier-du-Mets, Paris, France

Véronique LESPINET-NAJIB
Université de Bordeaux, IPB, IMS (UMR5218), 146 rue Léo Saignat 33 076 Bordeaux Cedex

Résumé : Dans la lignée de travaux précédemment réalisés, l'objectif de notre étude ici est de caractériser le contenu d'un site web organisationnel par les marqueurs lexicaux associés aux liens hypertextes de la page d'accueil (unités lexicales hypertextuelles – ULH). Nous nous sommes intéressés à un ensemble de sites web issus d'un domaine organisationnel homogène (i.e le domaine universitaire français), à partir duquel nous avons généré deux corpus d'ULH relevé à deux moments différents (2009 et 2011). Ces deux corpus d'ULH ont été intégrés au sein d'une taxonomie produite à partir des niveaux hiérarchiques et catégoriels des différents modes de navigation. S'appuyant sur un premier niveau d'analyse quantitative, l'intérêt est de mettre à jour, à travers cette dynamique temporelle, les évolutions au niveau des pratiques du web tout en faisant la preuve d'une stabilité certaine des corpus d'ULH.

Mots-clés : Unités Lexicales Hypertextes, approche diachronique, taxonomie, site web organisationnel.

1. Introduction

La notion de « site web » s'impose comme une manifestation emblématique des contenus architecturés du Web, mais elle dissimule aussi une réalité multidimensionnelle qui fonde sa nature complexe. Parmi les différentes facettes qui peuvent contribuer à son approche, nous retiendrons le site web en tant que ressource relevant de l'appellation générique « document numérique » (Leleu-Merviel, 2004). Sa caractérisation en tant que document numérique permet de l'envisager notamment au plan informationnel, selon une approche originale. En effet, nombre de travaux menés autour d'une analyse des contenus de sites web opèrent soit selon une perspective qualitative (évaluation ergonomique d'interfaces, étude d'ordre sémiotique), soit selon une perspective orientée statistico-sémantique à partir du contenu intégral de documents textuels. Notre hypothèse ici est de nous appuyer sur un niveau d'organisation des connaissances déjà réalisé sur le site, à savoir les liens hypertextes des pages d'accueil des sites web, liens hypertextes considérés dès lors comme les entrées d'un sommaire donnant accès aux contenus internes du site. Un travail de recensement automatique de ces liens hypertextes et de classification des unités lexicales associées permet ensuite de disposer d'une taxonomie à partir de laquelle un certain nombre d'éléments d'analyse et d'applications potentielles peuvent être réalisés.

Nos travaux[22] ont déjà amené des résultats intéressants. Constituée à partir d'un corpus réduit, la stabilité et la fiabilité de la taxonomie ainsi obtenue ont pu être vérifiées grâce à son pouvoir de recouvrement sur d'autres sites web du même domaine (Reymond *et al.,* 2011). D'autre part, deux pistes d'exploration sont en cours : générer des profils informationnels de sites web, sur la base des occurrences dans la taxonomie, à partir des dominantes thématiques qui émergent (notamment au plan des activités organisationnelles) ; permettre la reconnaissance et le classement automatique de sites[23].

Si notre postulat de départ reste identique (i.e. caractériser le contenu d'un site web par les marqueurs lexicaux associés aux liens hypertextes de sa page d'accueil), notre positionnement ici est différent. En effet alors que jusqu'à présent, nous avions mené nos analyses dans une perspective synchronique, à un instant t, nous nous positionnons ici dans une perspective diachronique pour tenter de mettre en lumière des éléments significatifs d'évolution. Comment évoluent ces marqueurs lexicaux dans le temps ? Quels enseignements peut-on tirer des

[22] Menés dans le cadre du programme RAUDIN (Recherches aquitaines sur les usages pour le développement des dispositifs numériques) à financement Feder n°31462, Conseil Régional Aquitaine et Université Bordeaux 3

[23] Ces enjeux et perspectives sont synthétisés dans une communication présentée aux journées d'études TICIS 2010.

modifications notoires observées, au plan des pratiques du web mais aussi de la dynamique organisationnelle ?

Les unités lexicales hypertextes (ULH) des pages d'accueil peuvent être assimilées à des fragments informationnels qualifiant les contenus internes du site. A ce titre, elles représentent non seulement les activités phare mais aussi des informations d'ordre structurel ou en prise avec l'actualité et la vie de l'organisation ; elles donnent également accès à des fonctionnalités rendues possibles par le média web (outils logiciels, flux RSS, etc.). Elles donnent donc à voir une certaine image de l'organisation et incarnent de fait une mémoire intéressante, signifiante et significative, des changements d'une structure, de ses choix, de ses représentations, de l'intégration de problématiques telles celle de l'accessibilité, ainsi que des phénomènes de lissage et normalisation terminologiques qui peuvent s'effectuer dans le temps (Rouquette, 2009). Dans une perspective plus opérationnelle, les analyses extraites, aux plans qualitatif et quantitatif, peuvent générer des préconisations en matière de qualification de l'information sur le web, en s'appuyant sur des pratiques consensuelles.

Partant d'un corpus unique de sites web, nous avons donc analysé l'évolution de ces marqueurs lexicaux à deux années d'intervalle[24]. Les premiers éléments de cette étude comparée que nous présentons ici sont essentiellement d'ordre quantitatif, pour tenter de qualifier les évolutions globales au plan des corpus[25]. Dans ce texte, nous développerons tout d'abord les arguments clefs de notre problématique, à savoir l'articulation site web/document numérique, la dimension organisationnelle pour le site web, le rôle déterminant de la page d'accueil et des ULH en tant que représentations informationnelles signifiantes de l'ensemble des contenus du site. Dans un deuxième temps, nous expliciterons le principe d'organisation de nos ULH en taxonomie, en nous appuyant sur une catégorisation en fonction des types de navigation. Ensuite, nous présenterons nos deux terrains d'investigation, choisis dans le domaine organisationnel des universités. Une université en sciences sociales et une université en sciences ont été sélectionnées pour constituer notre corpus d'ULH sur deux temporalités distinctes, 2009 et 2011. A partir de ce premier niveau d'analyse diachronique, la comparaison quantitative de nos corpus d'ULH permettra de dégager des éléments génériques intéressants, qui seront à affiner ultérieurement par l'auscultation de la dynamique des formes et de la sémantique des ULH.

[24] Cet intervalle de temps permet de faire émerger des tendances, à confirmer par l'analyse du même corpus à t+2.

[25] Une approche plus fine des évolutions terminologiques est actuellement en cours.

2. Problématique

2.1. Du site web au document numérique

Le terme de « site web » renvoie naturellement à un carrefour de pages et d'hyperliens, participant à la définition de territoires numériques (Musso, 2009). Mais au-delà de cette première approche, plusieurs déclinaisons peuvent être mises en avant pour tenter de définir un site web.

Tout d'abord on retiendra au premier chef celle d'un dispositif technique s'appuyant sur un serveur informatique et identifié par un nom de domaine de type DNS (Domain Name Server), réalité technique pouvant être rapprochée des contours et limites d'une organisation[26].

Au-delà de cette définition du site web par la mobilisation de ressources informatiques et protocoles techniques d'échanges de données (image du site web peu accessible à l'usager), le site web se donne à voir à travers son contenu. Stockinger définit le site web comme « lieu de prestations, lieu de services à destination d'un certain public » (Stockinger, 2005). Les sites web organisent de façon plus ou moins sophistiquée, diversifiée, cohérente et exhaustive, des ressources de nature informationnelle à destination de certaines catégories d'usagers-internautes. Ils sont dès lors des produits d'information éditée, renvoyant à une responsabilité éditoriale identifiée.

Le site web peut également être appréhendé comme un dispositif communicationnel « qui définit des rapports énonciatifs, attribue des rôles, inscrit des marques pour l'interprétation, à ceci près qu'il renvoie à un espace de communication dit « planétaire », en fait assez indéterminé sur le plan culturel et social » (Souchier et al., 2003). Il s'inscrit dès lors dans une tension entre intentionnalité(s) et usage(s), incluant les contraintes inhérentes au dispositif technique sollicité. Dans cette articulation d'éléments, l'intention (de communication, de proposition de services) se traduit par la production d'un site web au service d'une stratégie à destination de publics supposés.

Enfin, le site achève de prendre sens dans cet espace web à travers le réseau d'hyperliens qui le jalonnent et qui guident l'action des usagers tout en leur autorisant de multiples parcours, dans un paradoxe permanent entre clôture et évasion. Le site web peut ainsi être assimilé à un document numérique dynamique inscrit dans un espace topologique hypertextuel (Leleu-Merviel, 2004). On peut aussi adhérer aux propositions du collectif R.T. Pédauque selon lequel le document numérique s'articule autour de la tridimensionnalité Forme-Texte-Medium et possède les propriétés de mémorisation, organisation, création et transmission, propriétés dont nous nous inspirons pour bâtir notre argumentation présente.

[26] Nous verrons toutefois, à partir des sites de notre corpus, que cette proximité DNS/organisation n'est pas si évidente que cela en pratique.

2.2. Site web organisationnel

Ces dimensions multiples (et non exhaustives) des sites web génèrent de nombreuses possibilités d'approche et d'analyse. Il est toutefois difficile d'appréhender ces sites dans leur globalité, tant est grande l'hétérogénéité des formes, des contextes de production, des intentionnalités de communication, ou encore des données informationnelles. Dans l'optique de traiter, notamment au plan quantitatif, voire automatique, des ensembles de sites, il paraît indispensable de repérer des cohérences et régularités parmi les différents sites web.

Là encore, plusieurs angles de lectures peuvent être choisis. Stockinger propose une typologie des sites sur la base de leurs caractéristiques les plus saillantes : sites personnels « simples », sites d'information, sites d'accès à des ressources (sites portails), sites fac-simile ou dématérialisés de services organisationnels, sites à thème (Stockinger, 2005). On peut bien entendu regrouper les sites selon d'autres perspectives : par secteurs d'activité, selon la finalité, etc.

En ce qui nous concerne et par rapport à nos objectifs de recherche, nous avons choisi de travailler sur des corpus de sites représentant une organisation-type : nous appellerons ces sites « sites web organisationnels » (SWO). Ce choix de constitution de corpus, sous couvert d'identification d'organisations relevant d'un même domaine (au plan de l'activité), garantit un facteur de cohérence et d'homogénéité en termes de missions, services et publics visés. D'une façon plus spécifique, nous nous sommes intéressés aux sites web des universités françaises.

En considérant d'une part la présence historique du milieu universitaire sur le Web et d'autre part l'activité de communication des universités via leur site, le support internet est au cœur de la communication des universités. Ainsi que le rappelle 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). À ce titre, elle le caractérise en tant que ressource, qui s'inscrit dans une double logique de médiation et d'usage.

Par ailleurs, l'intérêt de travailler sur ce domaine organisationnel est d'avoir une architecture du web complexe, puisque sous l'appellation « site web d'université » se dissimule une démultiplication de sous-sites web, autonomes, mais participant tous de l'image organisationnelle de l'université.

2.3. Fonction de la page d'accueil et des ULH pour une discrimination informationnelle des SWO

Vis-à-vis de cet objet web aux facettes multiples, la page d'accueil joue un rôle déterminant. Celle-ci met en effet en représentation une grande partie de l'information gérée par l'organisation ou la structure concernée, tout en traduisant des choix sélectifs et stratégiques, qui en valorisent certaines dimensions plutôt que d'autres (Nielsen & Tahir, 2002). Mais la page d'accueil d'un site web se donne aussi à voir en tant qu'écrit d'écran, pour lesquels « nous disposons de « signes outils », de « signes passeurs » qui nous donnent accès aux multiples modalités du texte » (Souchier *et al.*, 2003). La fonction d'orientation des usagers du site vers des zones informationnelles spécifiques est donc assurée par ces signes passeurs, choisis notamment pour répondre aux contraintes issues de l'écriture pour ce média. Au plan des usagers, les stratégies de navigation web varient en fonction du niveau d'expérience Web (Thatcher, 2008). Il a été montré que les novices du web s'appuient principalement sur la recherche par sommaire (i.e. les menus de navigation), a contrario de la recherche par mots-clés qui s'avère plutôt efficace seulement pour ceux qui ont déjà des connaissances dans le domaine (Dressen-Hammouda & Drot-Delange, 2009) indépendamment de leur niveau d'expertise web. Dans le cadre de notre recherche, nous considérons donc la page d'accueil en tant que « sommaire » du site (au plan lexical), sommaire qui, tout en présentant des régularités de contenus, pourra prendre différentes formes pour un ensemble de sites du domaine.

Nous restreignons également la notion générique (Jeanneret, 2007) de « signe passeur » (incluant par exemple la dimension iconique) à celle de « texte passeur ». Marqués à la fois par des contraintes d'édition et de structuration web, les objectifs stratégiques de valorisation de l'information, et les capacités rédactionnelles des auteurs, les ULH de la page d'accueil, produisent une signature textuelle et sémantique de la page d'accueil (Reymond et Pinède, 2010a) et par extension, du site web concerné. En effet, la majeure partie de ces ULH sont des mots-clés représentant les contenus sous-jacents et constituent des marqueurs thématiques des contenus du site, que l'on peut traduire en signature informationnelle. À ce titre, les ULH d'une page d'accueil de site peuvent être considérées comme représentatives de l'ensemble de l'information portée par le site et, pour certaines, discriminantes (dans leur co-présence) par rapport à des sites web organisationnels.

Partant de ces postulats, nous nous intéressons ici à la comparaison de deux corpus d'ULH issus d'un ensemble de SWO du domaine universitaire. L'exploitation de ces deux collectes s'effectue à plusieurs niveaux :

- Comparaison générique entre ces deux corpus d'ULH permettant d'analyser l'évolution des marqueurs lexicaux aux plans quantitatifs

(augmentation/diminution du volume global de termes, à mettre en perspective avec les évolutions organisationnelles et techniques).
- Comparaison diachronique des entrées de la taxonomie par classe. Il s'agit ici, en superposant les disparitions et nouvelles entrées d'ULH dans les classes de la taxonomie, en prenant également en compte l'évolution des occurrences d'ULH dans les classes, de faire émerger des dynamiques thématiques significatives dans le temps (entre t et t+1, à savoir 2009 et 2011).

3. Méthodologie

3.1. Constitution des corpus d'ULH

Nous avons collecté les ULH présentes sur les interfaces de pages d'accueil[27] de sites web relevant du domaine organisationnel de deux établissements universitaires, se distinguant par leur champ disciplinaire (Sciences et Techniques / Sciences humaines et sociale)[28].
Par rapport à une première extraction de l'ensemble des sites web des zones DNS (Domain Name Service) de Bordeaux 1 et Bordeaux 2, nous avons opéré à une sélection en ne retenant que les sites répondant aux critères d'accessibilité (en langage HTML) ainsi que ceux présents simultanément en 2011 et sur le site webarchive.org pour la collecte 2009. Dans un deuxième temps, n'ont été conservés que les sites dont :
les ULH ont une sémantique propre (les critères d'exclusion sont : hyperliens en soi (http), courriels, nombres et ULH à 1 caractère) ;
le nombre d'ULH collecté par site est supérieur à deux.
Notre corpus de site est constitué au final de 96 pages d'accueil (ou pages de menu lorsqu'il existe une page d'introduction) de sites pour moitié issus respectivement des domaines DNS de l'université de Bordeaux 1 et de Bordeaux 2.
La collecte automatique s'appuie sur un analyseur lexical qui en premier lieu requête la page d'accueil ou son éventuelle redirection. La seconde phase extrait du code source les termes passeurs (ancres) des pages ainsi que le contenu des balises <ALT>[29]. L'extraction opère dans le contenu en langage html tout ce qui se trouve entre les balises et ainsi que dans le champ des balises ALT pour les hyperliens sur images, icônes ou zone (<xxx src=..>) inscrits en tant qu'hyperliens. Selon le format éditorial choisi par les éditeurs, la collecte automatique souffre d'un biais issu de la possibilité de masquer certains éléments du

[27] Par pages d'accueil, nous entendons chaque page d'accueil de tous les sites recensés dans une zone DNS identifiée. Soit une centaine de pages d'accueil en moyenne pour chaque organisation universitaire.
[28] Respectivement universités de Bordeaux 1 et Bordeaux 2.
[29] Ces balises sont destinées à l'accessibilité de la page pour combler la déficience visuelle et sont donc à vocation de décrire textuellement un marqueur de lien ou d'image.

code source lorsque le navigateur en produit le rendu via des feuilles de style (CSS par exemple) ou par programmation (ECMAscript). Le parseur n'interprète pas ces langages et est en conséquence insensible à ces variétés de mise en forme. Inversement, le collecteur peut aussi capter les marques d'accessibilité qui sont des compléments d'information textuels des données multimédia. Nous notons que le collecteur réalisé n'est pas exempt d'erreurs d'analyse lexicale lors du traitement des données. Toutefois après notre traitement automatisé nous avons pu mesurer une erreur peu significative : moins de 5% des ULH sont erronées (par exemple, , <xt javascript>).

3.2. Principes de constitution de la taxonomie

L'objectif est d'organiser les ULH collectées au sein d'une taxonomie (Tableau 1). Cette taxonomie se découpe selon trois catégories de navigation, déclinées à partir de la segmentation de (Nielsen et Tahir, 2002), elles-mêmes scindées en différentes classes :

- Catégorie 1 « Navigation thématique » : correspond aux ULH permettant de donner accès aux contenus des zones profondes des sites. Ces contenus peuvent être directement liés aux activités de l'organisation (5 classes) ou faire référence à des contenus génériques c'est à dire transversaux à différentes organisations (9 classes).
- Catégorie 2 « Navigation fonctionnelle » : correspond aux ULH faisant référence aux liens types outils (8 classes).
- Catégorie 3 « Navigation par profil » : correspond aux ULH permettant d'avoir accès à une recomposition du site pour les usagers en fonction de spécificités linguistiques ou autres (2 classes).

Catégories		Classes	Définition	Exemples d'ULH
Navigation thématique	Activités	Recherche	Activités ayant trait à la recherche	Ecole doctorale Laboratoires
		Formation	Activités ayant trait à l'enseignement	Master 2 Nouvelle licence 2011-2012
		Ressources documentaires	Ressources documentaires générales ou spécialisées accessible en ligne	Livres numériques Thèses électroniques
		Partenariat/transfert/ valorisation	Partenariats affichés de l'université (hors international)	Nos partenaires Collaborations et contrats
		International	Relations et politique internationales de l'université	Etudes à l'étranger Relations internationales
	Génériques	Accueil/présentation	Contenus généralistes sur l'organisation	Bienvenue Accueil
		Actualités	Contenus ayant une dimension temporelle, éphémère	A la une Evènements à venir

		Composantes extérieures	Mention de composantes ou structures extérieures à l'organisation	Accueil CNRS Institut Polytechnique de Bordeaux
		Informations pratiques	Informations, contenus démarches dépendantes de l'organisation	Infos pratiques étudiants Logement Santé-social
		Services dématérialisés	Accès à des services identifiés accessible en ligne (annuaire, etc.)	Intranet Annuaire
		Recrutement	Proposition d'emploi, mode de recrutement (hors formation)	Offres d'emploi Travailler à l'université
		Culture / loisirs	Activités liées au culturel, à l'associatif	Atelier photo Ciné-club
		Composantes de l'organisation	Dimension structurelle de l'organisation (hors activités)	Conseil d'administration Bibliothèque
		Logistique / équipement	Mention d'infrastructures techniques et des modalités de gestion de celle-ci	Equipement Réservation salles
Navigation Fonctionnelle		Accès web	Navigation en termes d'action dans le site	Cliquer ici Aller au pied de page
		Contacts	Renvoi sur un contact	Nous écrire Contactez-nous
		Technologies	Renvoi sur des formats, langages, ou outils informatiques (hors outils web 2.0)	Joomla Template XHTML 1.0
		Accès géographique	Indications géographiques d'accès au site physique	Plan d'accès Localisation
		Outils de communication web	Renvoi sur des outils d'échange et de communication du web 1.0 et 2.0	Flux RSS Liste de diffusion
		Authentification	Procédure d'authentification (sans mention du service associé)	Identifiez-vous Mot de passe oublié ?
		Mentions légales	Renvoi sur des mentions, informations obligatoires et légales	Crédits et mentions légales Législation
		Fonction rechercher	Renvoi sur la fonction de recherche (sur/hors du site)	Rechercher Moteur de recherche
Navigation par profils		Profil utilisateurs	Renvoi sur un espace dédié à un profil précis d'utilisateurs	Espace Entreprise Futur étudiant Accès étudiants
		Profils linguistiques	Renvoi sur l'interface traduite dans la langue concernée	En fr

Tableau 1 : Définition des principales classes de la taxonomie

Il est à noter que certaines ULH ne peuvent être classées actuellement et ce, pour trois raisons principales :
- ULH rejetées : certaines ULH renvoient à des mentions marginales ou anecdotiques (par exemple, « âme du bâtiment ») qui ne peuvent entrer dans un recensement terminologique dont la finalité est d'aboutir à une normalisation.
- ULH de spécialité : certaines ULH renvoient à des termes relatifs à des spécialités disciplinaires (par exemple, « Nanostructures Organiques », « Harmoniques XUV et impulsions attosecondes ») qu'on pourra tenter d'intégrer ultérieurement dans la taxonomie en s'appuyant sur des référentiels dédiés.
- ULH ambiguës : des ULH comme « programme », « équipe » sont évidemment des termes pertinents qu'il serait souhaitable d'intégrer à la taxonomie mais qui, à l'heure actuelle, demeurent ambivalentes sans qualification par le contexte associé : relèvent-elles de la classe « formation », de la classe « recherche » ou d'une autre classe ? Il s'avère donc impossible actuellement de les classer sans risque d'erreur.

3.3. Méthodologie d'analyse et de comparaison des corpus

Concernant le corpus, pour chaque année (2009 et 2011) nous mesurerons les variables suivantes :
- Le nombre d'ULH totales (ie. ULHt) ainsi que le nombre moyen d'ULH par site (nous préciserons l'étendue). Cet indicateur inclut les occurrences.
- Le nombre d'ULH distinctes (i.e. ULHd). Les ULH distinctes correspondent aux unités lexicales se distinguant par la forme. Par exemple, « Présentation » / « présentation » / « présentation » correspondent à trois ULH distinctes même si c'est le même terme désigné.
- La taille moyenne des ULH en nombre de caractères (espace et ponctuation inclus) (par exemple : « contactez-nous » : ULH de taille 14 / « programme de recherche » : ULH de taille 22)

Nous effectuerons des comparaisons de moyennes sur ces différents critères. Afin d'étudier l'évolution temporelle du corpus et de la taxonomie, plusieurs analyses seront effectuées :
- au plan de la caractérisation du contenu des interfaces pour déterminer les évolutions quantitatives,
- au plan des catégories d'ULH de la taxonomie pour transposer les évolutions précédentes aux catégories informationnelles présentées par les dispositifs et, indirectement, par les organisations.

De plus, nous précisons le nombre d'ULH t et d en fonction de leur appartenance à 3 types de corpus :
- corpus exclusif 2009 : ULH n'apparaissant qu'en 2009 et ayant disparu en 2011

- corpus commun aux deux années : ULH apparaissant aussi bien en 2009 qu'en 2011
- corpus exclusif 2011 : ULH n'apparaissant qu'en 2011

4. Résultats

4.1. Analyse sur le corpus

4.1.1 Caractéristique des interfaces

Le propose une comparaison concernant les principales données sur les ULH. Une progression, entre les deux années, a été observée sur le nombre moyen d'ULHt par site avec une progression de 136%. Une diminution en 2011 a été observée sur le nombre maximum d'ULHt par site (diminution de 9%).

	Nb moyen d'ULHt par site	Etendue du nb d'ULHt par site
2009	22	min 1 / max 174
2011	30	min 1 / max 160
Différence	+ 8	Max – 14

Tableau 2 : Caractéristiques des pages d'accueil du corpus en critères d'ULH (présence, taille)

Plusieurs hypothèses peuvent être formulées pour expliquer ce phénomène d'augmentation. On peut l'expliquer par une convergence au plan technique des modes de composition de l'interface page d'accueil. La progression du nombre d'ULHt peut être corrélée aux augmentations de formats d'écran (le plus souvent contraint en 1024x768 en 2009, et à 1280x1024 en 2011), laissant plus de place pour les menus, ou encore à l'utilisation de CSS et javascript pour dispenser via la même page menus ET sous-menus. L'autre hypothèse est de l'attribuer à une augmentation en volume des contenus publiés sur les sites. Pour vérifier cela, il faudrait corréler cette observation avec l'évolution de la taille des sites web sur la période concernée.

4.1.2. Caractéristiques du corpus d'ULH

La comparaison des deux corpus montre des différences importantes. Les évolutions temporelles concernent plusieurs indicateurs et révèlent une progression en faveur de 2011. Le Tableau 3 montre les caractéristiques générales des corpus d'ULH collectés sur les 96 sites. Rapportée au nombre d'ULH par site, l'augmentation entre les deux années (+725, soit 36% d'augmentation) montre la croissance de ces ensembles lexicaux. Si l'on regarde toutefois l'augmentation des ULH distinctes au plan lexical (+418, soit 21% d'augmentation), cela pondère

l'effet d'augmentation et permet de mettre en évidence une croissance relative, en phase au plan des contenus sur l'ensemble du corpus.

	2009	*2011*	*Variation*
Nombre d'ULHt - ULHt	2096	2940	+844
Nombre d'ULHd	1589	1895	+298
Proportion des ULHd / ULHt	76 %	64 %	

Tableau 3 : Caractéristique des corpus d'ULH collectées (nombre, occurrences, taille)

Le Tableau 4 montre la dynamique globale du corpus des ULH. Les ULH communes 2009-2011 représentent l'ensemble des ULH stables sur les deux années. Les ULH exclusives 2009 sont celles qui ont disparu entre les deux périodes de collecte. Les exclusives 2011 sont celles apparues en 2011.

	Nb ULHt (% / au nb total)	Nb ULHd (% / au nb total)	Proportion ULHd / ULHt
Exclusif 2009	841 (16,7%)	779 (29%)	92,7 %
Commun 2009-2011	2762 (54,8%)	818 (30,6%)	29,6 %
Exclusif 2011	1433 (28,5%)	1077 (40,3%)	75,2 %
Corpus complet	5036	2674	53,1 %

Tableau 4 : Dynamique temporelle du corpus d'ULH : les communs aux deux années, exclusifs 2009 et 2011.

Plusieurs points sont à mentionner :
- 55 % des ULHt du corpus complet sont issus des ULHt communes 2009-2011 alors que cette proportion diminue à 29,6% lorsqu'on s'intéresse aux ULH distinctes.
- au niveau des ULH communes 2009-2011, un principe d'occurrence très important est observé : 1 ULH apparait 3 fois en moyenne (proportion ULHd/ULHt).
- Entre 2009 et 2011, nous notons une diminution notoire de la dispersion des ULH (93% en 2009, contre seulement 75,3 % en 2011)..
- Sur le corpus complet, si on s'intéresse uniquement aux ULHd alors on se rend compte que ce corpus d'ULHd est surtout composé d'ULHd issues du corpus exclusif 2011 (43 %).

Le corpus commun illustre une forme de consensus autour de ces unités lexicales, à la fois dans le temps (permanence et stabilité de ces termes), et dans le partage (nombreuses occurrences). Nous pouvons constater une augmentation des occurrences sur le corpus 2011, montrant que si de nouvelles ULH sont utilisées, elles en appellent pour leur majorité à

des ULH existantes dans le corpus, ou encore à de nouvelles ULH mais « harmonisées » puisqu'à occurrences multiples.

4.2. Taxonomie

Nous avons classé chacune des ULH dans les différentes classes de la taxonomie présentée précédemment (cf. Tableau 1). Le Tableau 5 donne les statistiques de classement des ULH pour les deux années en précisant les ULH classées des ULH non classées.

	Corpus complet Nb ULHt / ULHd	Corpus classés Nb d'ULHt / ULHd	Corpus non classés Nb d'ULHt / ULHd
Corpus 2009	2096 / 1589	1385 / 961	711 / 636
Corpus 2011	2940 / 1782	2028 / 1156	912 / 739
Variation entre les 2 années	+844 / +185	+643 / +195	+201 / +103
Progression en %	+40% /+11,5 %	+46,4% / +20,3%	+28%/ +14%

Tableau 5 : Caractéristiques des différents corpus.

Ainsi, une progression du nombre d'ULH entre les deux années s'observe aussi bien sur le corpus classé que le corpus non classé. Il est à noter que sur les 5036 ULH total du corpus complet (2009 et 2011 inclus), 32 % n'ont pas pu être classées au sein de notre taxonomie. Nous retrouvons dans cette catégorie « non classées » les ULH rejetées, de spécialité et ambiguës (cf. 3.2 pour la définition des ULH non classées). Ce qui est important de noter, c'est que l'augmentation des non classés ne suit pas l'augmentation du nombre d'ULH (moins de 24%).

Les données suivantes ne concernent que les ULH classées au sein de notre taxonomie, afin de permettre une meilleure visibilité, nous ne présenterons que les données concernant les ULHt en fonction des catégories.

Concernant les ULH classées, si on reste au niveau des 3 catégories de notre taxonomie, il y a une grande différence en terme d'utilisation des ULH, avec une surreprésentation de la catégorie « navigation thématique » qui reste stable sur les deux années (Tableau 6) : pour les deux corpus exclusifs, 80 % des ULHt appartiennent à la classe « navigation thématique » alors que pour le corpus commun 2009-2011 cette catégorie diminue à 66%. Dans le corpus commun 2009-2011, 32 % des ULHt appartiennent à la catégorie « navigation fonctionnelle ».

Lorsqu'on compare les deux corpus exclusifs catégorie par catégorie, certaines évolutions temporelles apparaissent :

15% des ULHt de la classe « thématique » appartiennent au corpus exclusif 2009 contre 28 % appartenant au corpus exclusif 2011.
5 % des ULHt de la classe « navigation par profil » appartiennent au corpus exclusif 2009 contre 35 % appartenant au corpus exclusif 2011.
Concernant les ULHt de la classe « navigation fonctionnelle » il n'y a pas de différence entre les deux corpus exclusifs (10% pour 2009 et 15% pour 2011).
La catégorie « navigation fonctionnelle » est la plus stable des trois, celle où les ULH restent majoritairement présentes sur les deux années (peu de disparition, peu de nouvelles ULH). On note par contre une prise en compte croissante des profils utilisateurs (apparition importante d'ULH dans cette catégorie en 2011). Quant à la navigation thématique, dont la présence majoritaire reste constante, elle témoigne de la dynamique la plus importante, avec des renouvellements d'ULH importants, ce qui est à rapprocher des évolutions dans le temps de l'organisation concernée.

	Type de corpus d'ULHt						
	Exclusifs 2009		Commun 2009-2011		Exclusifs 2011		
Catégorie de la taxonomie	ULHt (%/type corpus)	% par catégorie	ULHt (%/type corpus)	% par catégorie	ULHt (%/type corpus)	% par catégorie	Total
« navigation thématique »	366 (80%)	15%	1382 (66%)	57%	684 (80%)	28%	2432
« navigation fonctionnelle »	84 (18%)	9,5%	663 (32%)	75%	135 (16%)	15%	882
« navigation par profil »	5 (1%)	5%	59 (3%)	60%	35 (4%)	35%	99
Total	455		2104		854		3413

Tableau 6 : Poids respectif des 3 catégories de la taxonomie en fonction du type de corpus

Par ailleurs, 5 classes (sur les 24 classes de la taxonomie) représentent à elles seules à peu près 60 % des ULHt (Tableau 7). On a ainsi une concentration importante autour de 5 pôles d'intérêt qui concernent à la fois les missions phare de l'organisation concernée (Recherche, Formation, Ressources documentaires) et une aide à la navigation sur le média web. Néanmoins, certaines différences apparaissent en fonction du type de corpus considéré. Pour les deux corpus exclusifs, les deux classes les mieux représentées sont la recherche et la formation avec une priorité à la recherche dans le corpus exclusif 2009 (23%) et une priorité à la formation pour le corpus exclusif 2011 (20%). Alors que les deux

classes les mieux représentées au sein du corpus commun sont respectivement : accueil/présentation (13,5%) et accès web (13%), la recherche n'apparaissant qu'en 3ème position (12%) et la formation en cinquième position (8%).

	% d'ULH classées / aux ULHt		
Ordre d'importance des classes	Exclusif 2009	Commun 2009-2011	Exclusif 2011
Rang 1	Recherche (23%)	Accueil/présentation (13,36%)	Formation (21%)
Rang 2	Formation (18,7%)	Accès web (13%)	Recherche (17%)
Rang 3	Ressources documentaires (9,6%)	Recherche (12%)	Informations pratiques (9%)
Rang 4	Accès web (7,7%)	Ressources documentaires (10,6%)	Ressources documentaires (9%)
Rang 5	Accueil/présentation (5,3%)	Formation (8%)	Accès web (8%)
% de représentation des 5 classes	64 %	57 %	62 %

Tableau 7 : Poids des 5 classes les plus représentées dans la taxonomie en 2009 et 2011

Afin d'observer plus précisément la hiérarchisation de ces 5 classes dans les corpus (exclusifs 2009 et 2011, commun 2009-2011), une carte cognitive30 a été réalisée. L'intérêt de cette carte est de faire apparaître visuellement à la fois les classes les plus fortement représentées et l'évolution de leur positionnement dans le temps. Les classes Formation et Recherche restent prédominantes sur les corpus exclusifs. Par contre, dans le corpus commun, ce sont les classes Accueil / Présentation et Accès web qui sont majoritairement représentées, ce qui illustre bien la stabilité des termes présents dans ces classes. Enfin, on relèvera présence croissante dans le corpus exclusif 2011 de la classe Informations pratiques, ce qui pourrait indiquer un changement dans les pratiques de diffusion d'informations (passage du papier au numérique).

[30] Réalisées avec le logiciel XMind

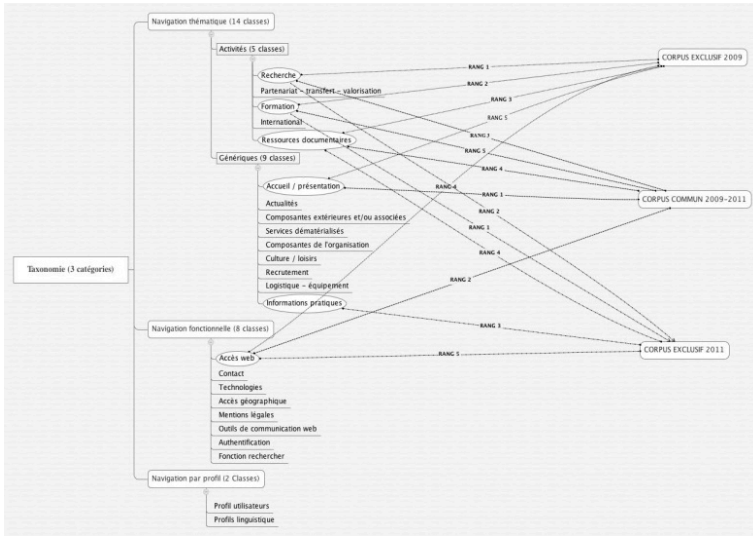

Figure 1 : Carte cognitive de la taxonomie en fonction des 3 corpus

Conclusion

Ces travaux ouvrent donc des perspectives de recherches intéressantes en terme d'analyse des contenus numériques. En l'occurrence, le fait d'observer, à travers le prisme de cette taxonomie, deux corpus de sites web appartenant au même domaine organisationnel permet tout d'abord de mettre en évidence des grands mouvements d'ensemble au plan des choix terminologiques. Ainsi, les deux images (en 2009 et 2011) de la taxonomie gardent les mêmes proportions, avec une relative stabilité des poids des classes informationnelles et des catégories sur les deux années. Les tests montrent une convergence (dans des temporalités courtes) de la gamme des ULH utilisées et de surcroît une normalisation de leur forme. Cela concerne essentiellement les classes de la navigation thématique. On voit notamment un ancrage temporel fort pour ces classes, et notamment pour la partie Activités. D'autres classes semblent moins dépendantes à cette dimension temporelle, notamment les classes Accueil / Présentation et Accès web, avec des ULH de marquage énonciatif plus pérennes. Ces tendances émergentes constatées ici seront à confirmer dans le temps, en observant le même corpus de sites web à échéance de deux ans.

Cette approche permettra aussi d'apporter des éléments d'information sur l'évolution des thématiques et des priorités, en accord avec les changements organisationnels. Cela suppose de descendre à un niveau

micro pour étudier les évolutions terminologiques à l'intérieur des classes informationnelles, tout en les situant par rapport au contexte organisationnel concerné, en l'occurrence l'université. Dans cette optique, il serait également intéressant d'intégrer la dimension sémiotique des pages d'accueil pour enrichir les dimensions d'analyse.

Au plan synchronique, d'autres perspectives sont à creuser, notamment en terme de consolidation de la taxonomie, par inclusion des termes de spécialités et traitement des ULH ambiguës. Même si le pouvoir de recouvrement de la taxonomie actuelle est satisfaisant, dans la perspective d'une alimentation automatique de la taxonomie (Rastier et al., 1994, Tellier, 2009), il est nécessaire d'améliorer la structure de la taxonomie. In fine, ces travaux nous permettront, à l'aide d'une typologie de sites web appuyée sur des secteurs d'activités (collectivités territoriales, santé), de transposer cette approche pour des applications à d'autres domaines (Reymond, Pinède, 2010b).

La démarche déployée ici, incluant la dimension temporelle, offre l'intérêt de proposer un angle d'analyse inédit du site web et de voir les permanences et renouvellements à travers l'image saisie de deux corpus d'ULH. Au prisme de la comparaison de ces deux corpus, l'évolution des marqueurs lexicaux –signes et traces-, tant au plan quantitatif que qualitatif, témoigne des changements, que ceux-ci se situent aux niveaux sémantique, informationnel ou encore strictement langagier. Ils témoignent aussi des inflexions de pratiques, en matière de représentation de l'information sur le web, ainsi que de l'évolution des contextes, d'organisation ou d'usage. C'est à l'intersection, aux interstices de ces différentes sphères que s'inscrit la production de sens pour ce document numérique particulier qu'est le site web.

Bibliographie

DRESSEN-HAMMOUDA D., DROT-DELANGE B. (2009). « Expertise et maîtrise des structures informationnelles : le cas de la documentation professionnelle en ligne pour les concepteurs web », Colloque «Evolutions technologiques et information professionnelle », organisé par le GRESEC, Université Stendhal Grenoble III, 10-11 décembre 2009.

JEANNERET, Y. (2007). *Y a-t-il (vraiment) des technologies de l'information ?* Lille, Presses universitaires du Septentrion.

LAINE-CRUZEL S. (2009). Documents, ressources, données : les avatars de l'information numérique. *Information-Interaction-Intelligence*, vol. 4, n°1, p 105-119.

LELEU-MERVIEL S. (2004). Effets de la numérisation et de la mise en réseau sur le concept de document. *Information-Interaction-Intelligence*, vol. 4, n°1, p. 121-140.

MUSSO P. (2009). « Critique de la notion de « territoire numérique ». Les dilemmes de l'économie numérique, sous la direction de Laurent Gille, Limoges, FYP Éditions (collection Innovation), p. 168-175.

NIELSEN J., TAHIR M. (2002). *L'art de la page d'accueil*. Paris, Eyrolles, 2002.
PINEDE N., Reymond D. (2010). De la diversité au lissage informationnel : création d'une taxonomie inductive pour les sites web universitaires. 17e congrès de la SFSIC: au cœur et à la lisière des SIC, Dijon 23-26 juin 2010.
RASTIER, F. CAVAZZA, M, ABEILLE, A. (1994). Sémantique pour l'analyse : de la linguistique à l'informatique, Paris, Masson.
REYMOND D., PINEDE N., LESPINET-NAJIB V., LE BLANC B. (2011). "Une étude terminologique de la communication hypertexte web. Application au domaine universitaire." 9th international conference on terminolgy and articial intelligence. Paris, 8-10 novembre.
REYMOND D., PINÈDE N. (2010a) "Using a taxonomy based fingerprint: classification and recognition of the academic webspace". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting*, Mysore, India, 2010.
REYMOND D., PINÈDE N. (2010b) "Website and communication strategy alignment: a librarian science approach to webometrics tools". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting*, Mysore, India, 2010.
REYMOND D., PINEDE N. (2010c). "Améliorer la "lecture" du web. Synthèse informationnelle des interfaces web". Communication aux journées d'études TICIS, Paris, 13-15 décembre.
ROUQUETTE S. (2009). L'anayse des sites internet. Une radiographe du cyberspace. Bruxelles, de Boeck , 2009.
SOUCHIER E., JEANNERET Y. et LE MAREC J. (2003). (dir.). *Lire, écrire, récrire. Objets, signes et pratiques des médias informatisés*, Paris, BPI, 2003, 335 p.
STOCKINGER P. (2005). *Les sites Web. Conception, description, évaluation*. Paris, Hermès- Lavoisier, 2005, 270 p.
TELLIER I. (2009). « Apprentissage automatique pour le TAL : Préface », *Traitement Automatique des Langues* 50, 3, p.7-21.
THATCHER A. (2008). « Web search strategies: The influence of Web experience and task type », Information Processing and Management, vol. 44, n°3, p. 1308-1329.