QUALIFIER LA PRESENCE D'UNE VILLE SUR LE WEB PAR DES INDICATEURS CYBERMETRIQUES SPATIO-TEMPORELS : Une validation expérimentale pour 2 villes moyennes de la région de Tunis

Eric Boutin, Peggy Cadel

QUALIFIER LA PRESENCE D'UNE VILLE SUR LE WEB PAR DES INDICATEURS CYBERMETRIQUES SPATIO-TEMPORELS :

Une validation expérimentale pour 2 villes moyennes de la région de Tunis

---

Eric Boutin,
Maître de conférences en Sciences de l'information – communication IUT de Toulon
boutin@univ-tln.fr + 33 4 94 14 23 56

Peggy Cadel,
Enseignant chercheur IUT de Saint Raphaël
cadel@univ-tln.fr +33 4 94 19 66 00

Adresse professionnelle
Laboratoire I3M
Université de Toulon-Var ★ BP 132 ★ F-83957 La Garde Cedex

Résumé: L'information web ne comporte pas toujours de référent spatial et temporel. Rares sont les pages web qui portent une mention explicite de leur date de création, de leur date de mise à jour, de l'adresse de leur concepteur. Or, la validation d'une information quelle qu'elle soit (information scientifique, technique, économique, stratégique) passe par l'ancrage de cette information dans le temps et l'espace.
Partant de ce constat, l'objectif de cette communication est d'estimer la date et le lieu de création d'une page web et de montrer l'intérêt que présente ces notions pour l'analyse cybermétrique. Cet article fait l'objet d'une validation expérimentale dans le domaine de la veille territoriale à partir de l'étude de la présence sur le web de 2 villes tunisiennes. Ces validations expérimentales seront analysées au regard d'études similaires réalisées à partir de corpus de villes françaises.

Mots clés : temps, espace, indicateur, cybermétrie, information massive, traitement automatique

Lorsqu'on effectue une recherche d'information, quelle que soit la source privilégiée, il est important de pouvoir ancrer cette information dans le temps et dans l'espace : quand cette information a –t-elle été publiée ? et par qui ?. La réponse à ces deux questions est essentielle et va entrer en ligne de compte dans le processus complexe permettant de qualifier la pertinence de l'information collectée. Dans le cas de l'information disponible sur internet, la mise à disposition d'indicateurs spatiaux et temporels revêt la même importance. Toutefois, on se heurte au niveau opérationnel à la difficulté d'avoir accès à cette information de manière systématique. Pour obtenir l'âge ou le nom de l'auteur à l'origine d'une page web, il faut bénéficier d'une expertise dans la maîtrise de bases de données dispersées.
Si le fait d'affecter des indicateurs spatio-temporels à une page web est un élément utile dans le processus de qualification de pertinence de cette page web, ces deux familles d'indicateurs présentent également un autre intérêt au niveau macroscopique cette fois. En effet, lorsqu'on s'intéresse à un corpus de page web, les indicateurs spatio-temporels vont permettre d'établir des statistiques permettant de répondre aux questions suivantes :
- quelles sont les caractéristiques identitaires des auteurs des pages web dans le domaine donné ?
- quelle est la pyramide des âges de ce corpus web ? Cette notion consiste à transposer au monde du web des indicateurs démographiques comme ont pu le faire Pitkow et Pirolli.
Nous avons souhaité appliquer ces indicateurs cybermétriques spatio-temporels à deux corpus web correspondant chacun à la présence de deux villes tunisiennes moyennes sur le web. Ayant eu la chance il y a 20 ans de vivre 4 années entre Ez Zahra et La Marsa, la perspective de cette validation exploratoire par les indicateurs spatiotemporels était un clin d'œil à l'espace et au temps !
La présentation des indicateurs spatio-temporels retenus fera l'objet de la première partie de ce travail. La validation expérimentale qui s'en suivra s'intéressera à la caractérisation de ces 2 villes tunisiennes sur la toile. Les résultats expérimentaux seront comparés entre eux et rapportés à des travaux analogues conduits pour le compte de villes moyennes françaises.

I. Comment caractériser une page web par un indicateur spatio-temporel.

L'objectif de cette partie n'est pas de développer une nouvelle méthode de caractérisation de pages web par des indicateurs de temps et d'espace mais d'utiliser une méthode éprouvée depuis deux ans en la mettant au service d'une validation expérimentale originale. La méthodologie retenue sera donc présentée dans un premier temps de façon didactique en même temps que les limites intrinsèques à une telle caractérisation.

A. La détermination d'indicateurs spatiaux pour une page web

Lorsqu'une page web est créée, celle-ci est construite sous la responsabilité d'une personne physique ou morale. Pourtant, rares sont les pages web dans lesquelles ces auteurs apparaissent ou sont cités dans la page web elle-même. Ce constat varie selon les spécialités mais demeure valide au niveau général. Ainsi, il apparaît hasardeux de chercher à extraire l'auteur d'une page web dans la page elle-même pour des raisons tenant à la disponibilité de cette information et au fait que cette information, quand elle existe, est égarée quelque part dans la page web.
Pour connaître l'auteur d'une page web, nous proposons de recourir à une autre source d'information non pas intrinsèque (contenue dans la page elle-même) mais extrinsèque (contenue dans d'autres espaces du web).
Lorsqu'un acteur du web souhaite déposer un nom de domaine, il doit fournir des informations identitaires à la société qui va ensuite lui octroyer ce nom de domaine. Ces informations sont ensuite mises à disposition des autres internautes à travers l'utilisation d'une commande whois sous unix ou en utilisant les services de sites web gratuits de type www.leregistrar.com spécialisés dans ce type de service. L'exemple présenté figure 1 illustre l'information renvoyée par un registrer pour le site www.fahrplancenter.com .

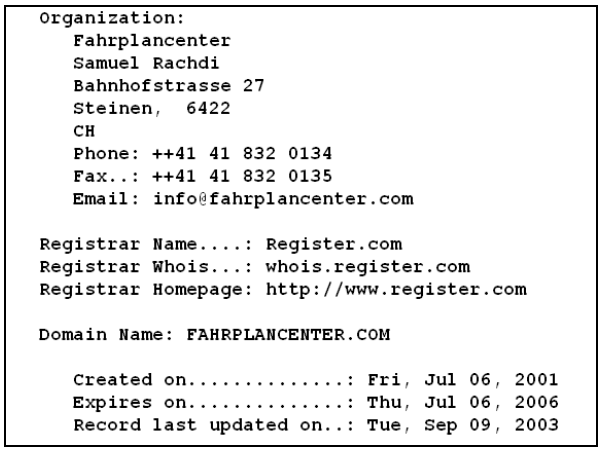

Figure1 : exemple de données obtenues par une requête whois

En exploitant les informations contenues dans ce fichier, on peut récupérer différentes informations identitaires sur le déposant du nom de domaine. Ainsi, peut-on observer que ce site, qui parle du train de banlieue sud de Tunis, a un nom de domaine qui a été déposé par une entreprise Suisse.
L'utilisation de cette information soulève plusieurs problèmes :

- Un problème de pertinence dans le cas de site communautaire par exemple : en effet, si le site web est déposé dans un espace communautaire, le nom du déposant ne correspond pas au nom de celui qui a déposé la page personnelle mais au nom de celui qui a déposé le site communautaire. Il nous sera donc impossible d'estimer par cette méthode le nom du déposant d'une page personnelle hébergée par un site communautaire.
- Un problème de disponibilité : Parfois, il est impossible d'obtenir les informations du registrer soit parce qu'on l'a interrogé à des intervalles de temps trop contigus soit parce que le déposant du nom de domaine a fait en sorte qu'on ne puisse pas accéder à cette information soit parce que le registrer utilisé ne gère pas cette extension pays.
- Un problème juridique dans la collecte de l'information puisqu'il est interdit de faire un usage systématique ou marchand des informations récupérées sur les registrers. La collecte automatique par une routine de l'information des registrers se heurte de ce fait à des obstacles : certains registrers autorisent le requétage successif de leur base après un certain laps de temps ce qui est bloquant pour une analyse d'un grand nombre de données.
- Un problème dans le traitement de ces données du fait du formatage spécifique de chaque registrer : celui-ci renvoie en effet l'information correspondant au nom de domaine demandé en suivant une mise en forme spécifique qui ne présente aucun caractère homogène. De plus, ce formatage des données n'est pas toujours très explicite. A ce titre, les fichiers renvoyés par la Frnic sont peu exploitables par des routines automatiques pour extraire la date de dépôt du site web.

Les expérimentations que nous avons réalisées nous ont conduit à estimer que le nom du déposant pouvait être connu avec pertinence pour environ 50% des sites web. Il est donc délicat de prétendre conduire une analyse exhaustive à partir de cette information.

B. La détermination d'indicateurs temporels pour une page web

Différents indicateurs temporels sont envisageables. Nous en avons identifié 3. Chacun permet de répondre à la question suivante :

⇒ Quand la page a-t-elle été mise sur le web pour la première fois ?
⇒ Quand la page a-t-elle été mise à jour pour la dernière fois ?
⇒ Quand le site a-t-il été déposé sur le web ?
Selon que l'on souhaite répondre à l'une ou l'autre de ces questions, on va privilégier une source d'information particulière.

1. La date de mise à disposition de la page sur le web

La récupération de l'état d'une page web à différentes dates est rendue possible par l'utilisation du moteur *web.archive.org*. Ce moteur de recherche est le fruit d'un travail d'archivage du web afin, selon Feise (2000) d'en garder la mémoire. Ainsi, le moteur de recherche dispose de robots qui ont scanné le web à divers moments de son histoire et ont conservé des images fidèles à intervalle régulier. Il est ainsi possible de récupérer ces différentes versions de pages web pour les analyser.
L'utilisation de l'interface de web.archive.org est assez intuitive. Lorsqu'on tape l'adresse d'une page web, on obtient un tableau qui renvoie une liste de dates rangées par année. Il s'agit des différentes instances d'une page stockée. La date la plus ancienne correspond à la date à laquelle le moteur a scanné la page pour la première fois. Cette date est postérieure à la date de mise à disposition de la page sur le web. Pour pouvoir utiliser cet outil, il faut s'assurer que l'écart de temps entre la date de création de la page web et celle de son référencement dans web.archive.org est limité et que ce moteur a une couverture satisfaisante du web.
La figure 3 donne l'exemple du résultat obtenu pour le site ***http://www.fahrplancenter.com/SNCFT_Bilder_Banlieue.html***. Le problème principal de cette source d'information est qu'elle n'est pas toujours disponible. Chaque mise à jour de la page est mentionnée par une « * » à droite du lien hypertexte mentionnant la date. Ainsi, peut on observer que cette page n'a pas été mise à jour depuis sa première date de prise en compte dans le moteur web.archive.org le 22 Janvier 2003.

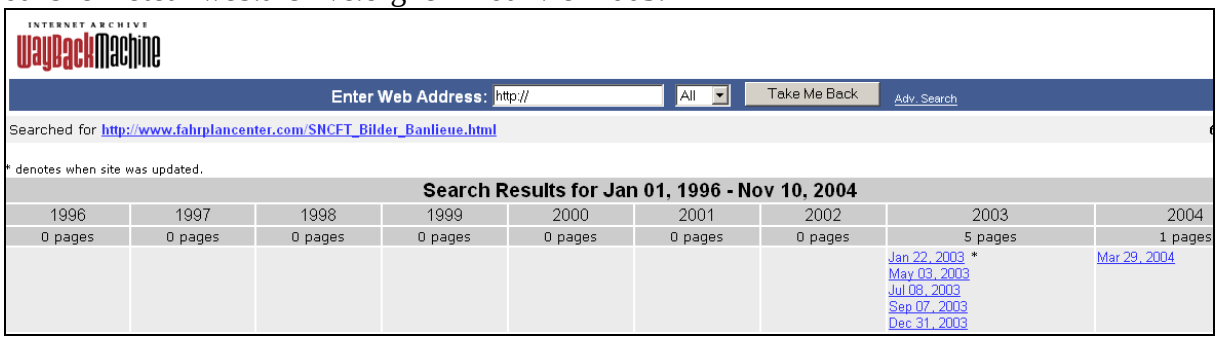

Figure 3 : capture écran de web.archive.org sur la requête mairie-avignon.fr

2. La date de dernière mise à jour de la page

Cette information est laissée à la discrétion du concepteur du site web : dans la pratique elle est peu renseignée. Etant donné que cette information est utilisée dans les algorithmes des moteurs de recherche pour qualifier la fraîcheur d'une page web, certains sites peuvent être rafraîchis automatiquement de façon artificielle pour augmenter leur pertinence sur les moteurs si bien que cette information perd de son intérêt. De plus, avec le développement du web dynamique, les pages web sont créées à la demande de l'internaute. La date de création de la page se confond alors avec la date de la requête et n'a donc plus de sens. Ces trois raisons nous ont conduit à ne pas exploiter cette information

3. Date de dépôt du site sur le web

Cette information est fournie par les registrers de la même manière que l'information spatiale exploitée précédemment. Elle est donc soumise aux mêmes réserves que celles qui ont été précédemment définies. A ces réserves s'ajoute un problème de précision : La date de dépôt du nom de domaine ne correspond pas forcément à la date à laquelle la page a été mise sur le web pour la première fois. Celle-ci lui est généralement postérieure.

C. Protocole expérimental retenu :

Le protocole retenu comporte les étapes suivantes :

- On s'intéresse à 2 villes moyennes de la grande banlieue de Tunis (Ez Zahra, La Marsa,).
- Pour chacune de ces villes, l'expérience consiste à récupérer un échantillon de pages web permettant de caractériser la présence de la ville sur le web. Cet échantillon est constitué à partir du moteur de recherche Google. Pour chacune des villes, on lance la requête *nom de la ville*. On s'intéresse donc exclusivement au web francophone sur ces deux villes.
- L'échantillonnage consiste à sélectionner les 100 premiers sites différents renvoyés par le moteur Google.
- Pour chacun de ces 100 premiers sites, on lance une requête whois grâce à une commande Linux ad hoc qui permet d'accéder aux bases de données des registrers. Ces données sont ensuite parsées automatiquement à la recherche de la date de dépôt du nom de domaine et de l'adresse du déposant.
- On analyse également pour chacune des pages la thématique sous jacente.
- A partir des dates de dépôt des divers noms de domaine et de l'adresse des déposants, il est possible de reconstituer deux indicateurs spatio-temporels
    - la pyramide des âges du territoire. La pyramide des âges est traditionnellement un indicateur démographique. L'objectif de ce papier est, à la manière de ce qu'ont pu faire Douglis et al [3] et Pitkow et Pirolli. [10], de transposer au monde du web des indicateurs démographiques. La pyramide des âges d'un ensemble de pages web est un graphe qui représente la statistique de date de dépôt d'un nom de domaine en fonction de l'année de dépôt. Cet indicateur présente un intérêt en tant que tel mais il peut aussi être comparé entre plusieurs corpus.
    - le graphe d'ouverture du territoire. Ce graphe s'intéresse au lieu de dépôt des noms de domaine pour un territoire considéré. Ainsi peut-on observer si les sites du territoire sont déposés en Tunisie, en France, à l'étranger.

II. Validation expérimentale sur 2 villes tunisiennes moyennes de la région de Tunis.

A. La pyramide des âges :

La *Figure 4* permet de visualiser sur le même graphique la pyramide des âges des deux villes étudiées. Il est à noter que les dates de dépôt des noms de domaine ne sont pas fournies dans le cas des sites déposés en Tunisie (extensions .tn) ou en Allemagne (extension .de), ce qui constitue une limitation forte de notre analyse quand on voit le poids de ces deux pays dans le dépôt de noms de domaine. Les établissements gestionnaires des noms de domaine de ces pays ne donnent pas accès à ces informations temporelles. Les statistiques de la *figure 4* ont donc été établies sur 59 sites dans le cas de la ville d'Ez Zahra et de 77 sites dans le cas de la ville de « La marsa ».

Avant d'analyser la similarité existante entre ces deux courbes, il est important de prendre en compte le fait que les données temporelles restituées par cette pyramide des âges sont biaisées par l'indicateur de pertinence du moteur de recherche Google (Page et Al 1999) qui a été notre source d'information. En effet, l'analyse que nous avons conduite ne prend pas en compte l'intégralité du corpus web caractérisant ces deux villes (environ 50 000 pages dans l'index de Google pour la requête « La Marsa » et environ 11 000 pages dans l'index de Google pour la requête « Ez Zahra ») mais uniquement les 100 premières pages les mieux classées par le moteur de recherche Google. Or, l'indicateur de pertinence de Google valorise les pages en fonction du nombre de liens entrants qu'elles reçoivent. Cette capacité à recevoir des liens entrants dépend de critères temporels : à pertinence égale, une page ancienne aura une probabilité plus forte d'obtenir un grand nombre de liens hypertextes entrants qu'une page récemment mise sur le web. Les statistiques temporelles ont donc intrinsèquement tendance à vieillir le corpus comme l'ont montré Ntoulas et al [2004], Fetterly et al [2003] et Lim et al [2001].
L'observation de cette pyramide des âges fait apparaître une grande similarité dans l'allure générale des deux courbes. La date la plus prolifique pour le dépôt des noms de domaine est 1999. Dans les périodes récentes, il semble se produire une différence entre une pente continue observée sur les résultats des sites de « la Marsa » et une structure beaucoup plus cyclique observée sur le corpus d' « Ez Zahra ».

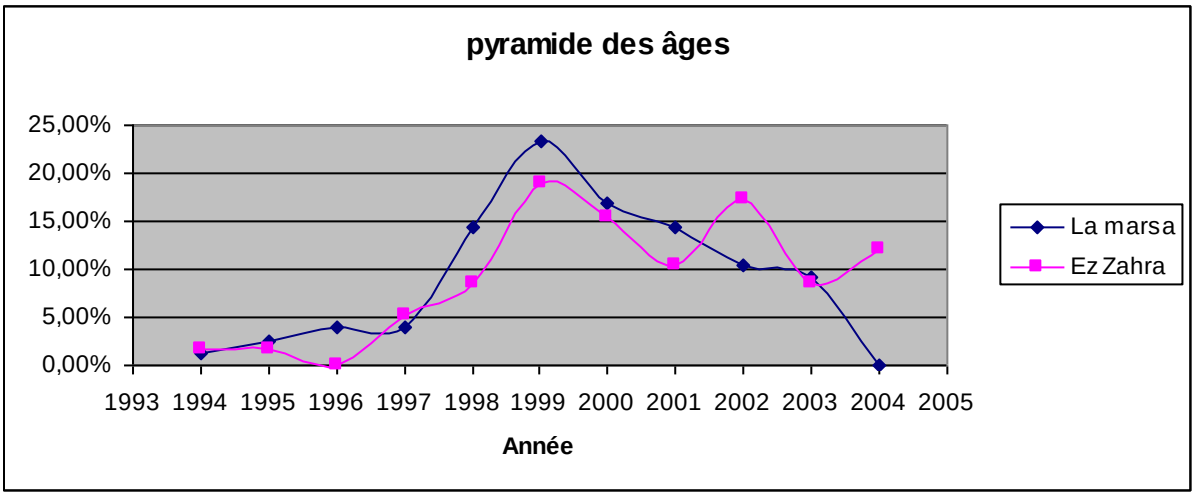

Figure 4 : pyramide des âges de La Marsa et d'Ez Zahra

Nous avons essayé de chercher des explications à cette similitude dans les courbes. Plusieurs explications sont possibles.

- La première piste nous conduit à identifier entre ces corpus un commun dénominateur constitué des sites web qui seraient présents dans les deux corpus. Pour faire ce travail tout en élargissant l'étude à un plus grand nombre de villes tunisiennes, nous avons considéré 10 villes tunisiennes. Pour chacune de ces 10 villes, nous avons récupéré les 100 premières réponses de Google et avons ensuite analysé, pour chaque couple de villes, le nombre de sites web qu'elles avaient en commun dans leurs réponses. Les résultats de ce travail sont contenus dans les graphes de *la figure 5 et 6*. Ces graphes illustrent le fait qu'il y a peu d'intersection entre les sites de Ez Zahra et de La Marsa. Par contre, on observe que 5 villes (Bizerte, Nabeul, Sfax, Tozeur, Gabès) ont une très forte interaction puisque plus de 20% de leurs sites web sont communs. Le nombre de sites communs entre Ez Zahra et La Marsa est en tout état de cause trop faible pour que cette hypothèse soit retenue pour expliquer la similitude des courbes.

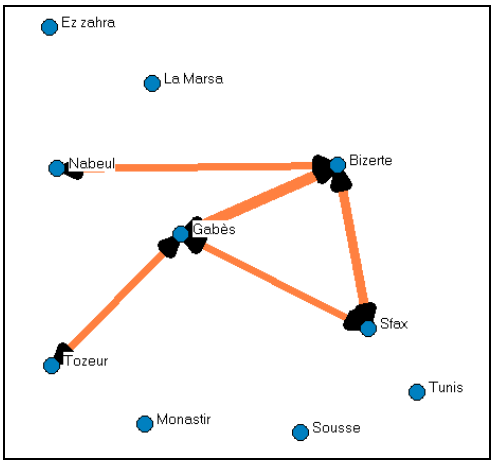

Figure 5 : interaction entre les villes ayant au moins 20% de leur sites web en commun

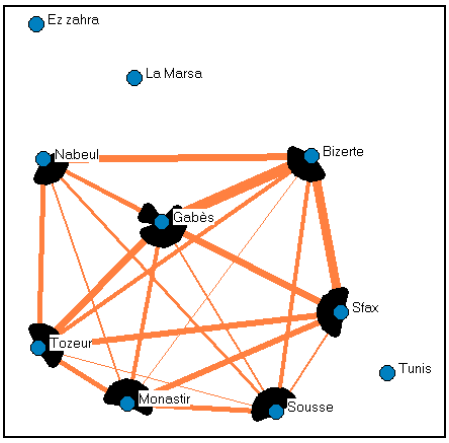

Figure 6 : interaction entre les villes ayant au moins 10% de leur sites web en commun

- La seconde piste pour expliquer les similitudes dans ces deux courbes fait intervenir des comparaisons avec d'autres études similaires que nous avons réalisées (Boutin, [1] et [2]) sur des villes françaises.

La *Figure 7* permet de visualiser la pyramide des âges de 10 villes françaises. Cette pyramide des âges a été construite en utilisant la même méthode que celle que nous avons retenue. Cette figure illustre une superposition forte et une allure fortement cyclique reposant sur des cycles de deux ans.

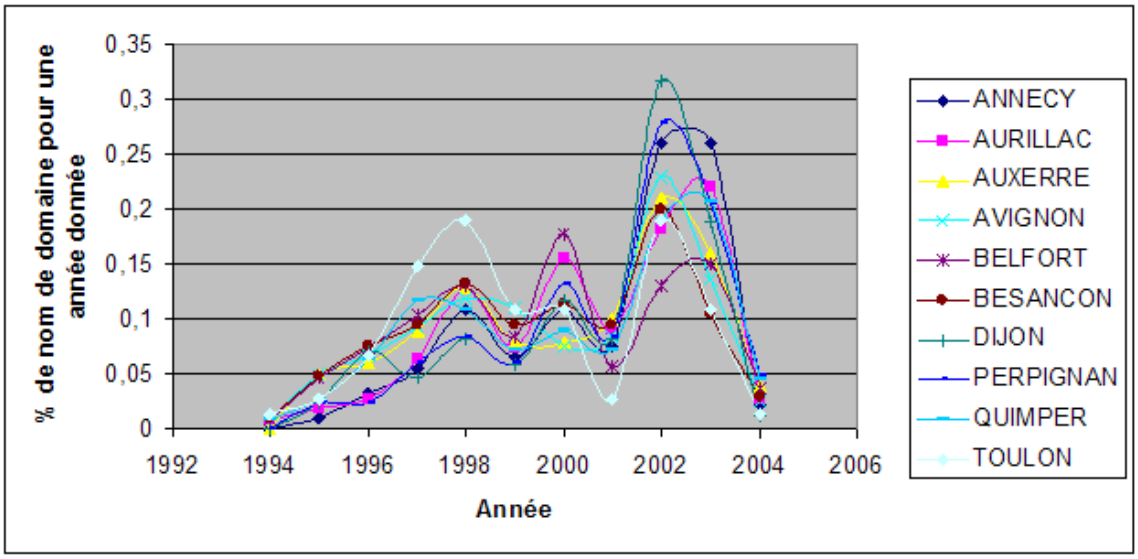

Figure 7 : pyramide des âges de 10 villes françaises

Ces comparaisons ne permettent pas d'identifier de structures récurrentes et de similitudes particulières avec les courbes établies sur les deux villes tunisiennes.

- Il semble très difficile d'expliquer les structures de ces deux courbes car les noms de domaine sont déposés par des acteurs qui n'ont pas une logique territoriale particulière. Lorsqu'on s'intéresse à la présence de villes françaises sur le web, on observe un réseau d'interaction fort entre les acteurs institutionnels de la ville considérée. Dans le cas de ces deux villes tunisiennes, la présence institutionnelle de ces villes est beaucoup plus diffuse. Du coup le champ est libre pour la représentation d'intérêts très divergents appartenant à des logiques marchandes, historiques, culturelles, personnelles.

B. Analyse du lieu de dépôt du nom de domaine :

Dans cette partie, l'objectif consiste à s'intéresser au pays dans lequel le nom de domaine a été déposé. Cette information est plus systématiquement fournie par les registrers de noms de domaine. Nous avons choisi de répartir les pays de provenance des noms de domaine de façon ternaire : Europe, Tunisie et Reste du monde. Le *tableau 1* fournit la part relative de chaque catégorie pour les deux corpus. On observe que dans la majorité des cas, pour les deux villes, les noms de domaine sous jacents sont déposés par des résidents européens. Les noms de domaine déposés par des tunisiens représentent selon le cas 9% ou 18%.

	La Marsa	Ez Zahra
Europe	59%	61%
Reste du monde	33%	20%
Tunisie	9%	18%

Tableau 1 : ventilation des noms de domaine par origine géographique

La *figure 8* est la traduction graphique du *tableau 1*. Selon ce graphique en triangle, plus une ville est située près d'un pôle, et plus cela signifie que les déposants des noms de domaine ont pour origine géographique le pôle considéré.

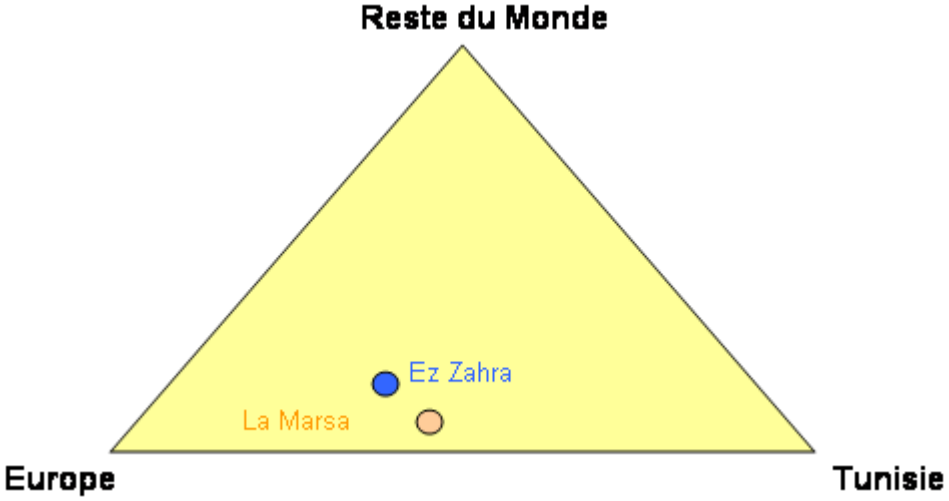

Figure 8 : ventilation des noms de domaine par origine géographique

Nous avons choisi de rentrer dans le détail de ces trois pôles pour mieux cerner la façon dont se répartissent les noms de domaine entre ces grandes masses. Ce qu'illustre la Figure 9.

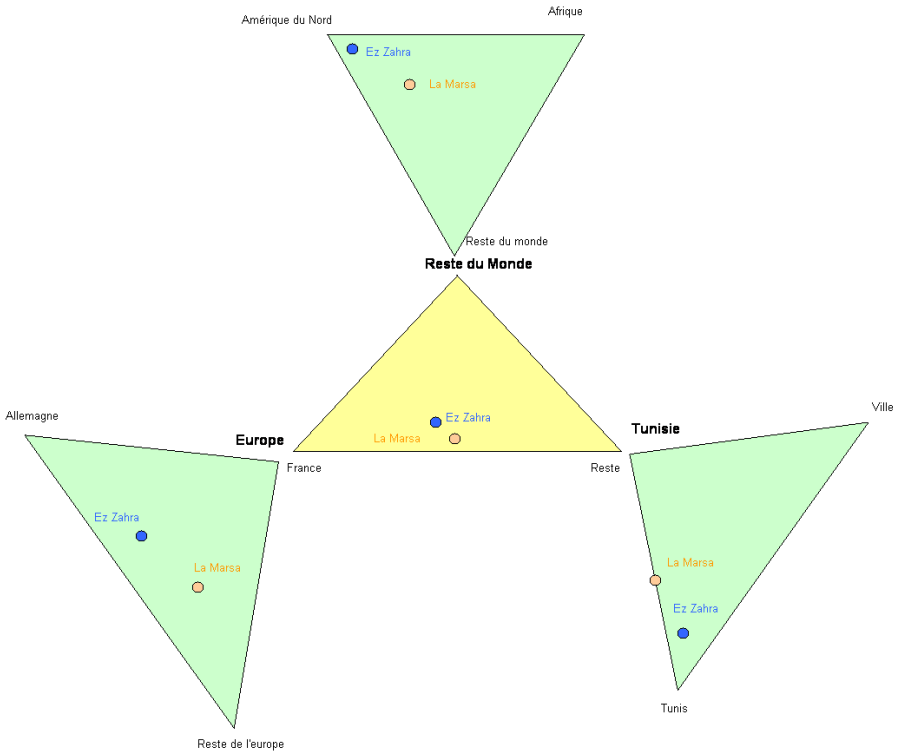

Figure 9 : détail de la ventilation des noms de domaine par origine géographique

En ce qui concerne les noms de domaine déposés en Tunisie, on observe une concentration forte des dépôts dans la ville de Tunis. Il faut convenir que les deux villes étudiées se situent l'une et l'autre dans la proche banlieue de Tunis. Ce type d'observation est assez différent de ce que nous avions pu constater sur des villes françaises.
En ce qui concerne les noms de domaines déposés dans des pays européens, on note une différence entre La Marsa et Ez Zahra. A Ez Zahra, de nombreux noms de domaine sont déposés en Allemagne : les sites correspondants sont des sites axés surtout sur le tourisme. Le poids de l'Allemagne est plus fort que celui de la France. De ce point de vue, l'origine des noms de domaine des déposants de page parlant de La Marsa est mieux répartie entre les pays d'Europe. Concernant les dépôts dans le reste du monde, on observe la prédominance des noms de domaines nord américains, USA notamment.

Conclusion :
Ce travail expérimental a pour objectif, à partir de deux exemples concrets, de montrer l'intérêt que peuvent représenter des indicateurs spatio-temporels pour l'analyse cybermétrique. Cette étude est un point de départ qui devrait déboucher sur des procédures d'automatisation autorisant la systématisation de l'analyse de ce type de corpus. Les données collectées sont pour l'instant difficilement comparables avec des analyses semblables réalisées sur des corpus français. Une piste intéressante consisterait à multiplier ce type d'analyse sur d'autres villes et à élargir les analyses menées au domaine relationnel. Cela permettrait de mieux comprendre les interactions entre les sites d'un corpus.

Bibliographie :

[1] BOUTIN E « L'exploitation de l'âge d'une page web : quelles perspectives pour l'analyse cybermétrique », acte du colloque, VSST 2004, tome B, P.439-449, Octobre 2004

[2] BOUTIN E, Qualifier la présence d'une ville sur le web par des indicateurs cybermétriques dynamiques : une validation expérimentale sur 10 villes françaises. » Acte du colloque TIC et Territoires, Lille, Juin 2004

[3] FETTERLY D. MANASSE M., NAJORK M., WIENER JL. A large-scale study of the evolution of web pages. In Proceedings of the Twelfth WWW Conference, Budapest, Hungary, 2003.

[4] LIM L., WANG M., PADMANABHAN S., VITTER JS., AGARWAL RC., Characterizing web document change. In Proceedings of the Second International Conference on Advances in Web-Age Information Management, pages 133–144. Springer-Verlag, 2001.

[5] NTOULAS A., CHO J., OLSTON B. What's New on the Web? The Evolution of the Web from a Search Engine Perspective, WWW2004, May 17–22, 2004, New York, New York, USA, page 1

[6] PAGE L., BRIN S., MOTWANI R., WINOGRAD T. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999).

[7] PITKOW J., PIROLLI P., Life, death, and lawfulness on the electronic frontier. In Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia, 1997.