I

Appréhender la navigation sur le Web : questions, méthodes, données, outils

Qu'est-ce qu'un parcours sur le Web ? Comment appréhender cet objet de recherche articulant les logiques de production des contenus et celles de leur usage en situation ? Quelles données et quels outils va-t-on mettre en œuvre pour en construire des représentations fidèles ? C'est à cet ensemble de questions que répond cette première partie. Dans un premier temps, nous proposons une approche praxéologique des parcours fondée sur la description et l'interprétation des situations de navigation et des régularités construites par l'usage, en s'appuyant sur des données de trafic centrées-utilisateur. Après avoir décrit le format et les contraintes de ce matériau brut, nous exposons les méthodes que nous avons élaborées pour l'enrichir et le manipuler : représentation des contenus visités, indicateurs topologiques et temporels des parcours, outils de visualisation. Ce n'est qu'ainsi outillés que nous pouvons envisager ensuite une analyse des parcours sur le Web tenant compte de leur forme, de leur contenu et de leur inscription dans les pratiques individuelles.

Chapitre 1

Appréhender les parcours sur le Web

Le travail que nous présentons ici se propose d'analyser les parcours sur le Web sur la base de données de trafic centrées-utilisateur. Deux questions sont soulevées : d'une part, la définition de ce qu'est un parcours et en quoi cet objet s'articule dans l'activité de navigation en général ; d'autre part, à quels questionnements sur cet objet les données de trafic permettent-elles de répondre, et quelles méthodologies cela implique-t-il.

1.1 Le parcours comme objet d'analyse

Nous entendons poser les bases méthodologiques et empiriques d'une sémantique des parcours ; nous estimons qu'il y a là un champ de recherche à part entière dont la spécificité ne se construit pas uniquement sur la particularité sémiotique de son objet, mais également sur les modes d'interaction entre l'utilisateur et les contenus qu'il appréhende.

1.1.1 Le parcours au centre de l'activité de navigation

La diffusion en milieux domestique et professionnel des outils informatiques en général et de l'accès à Internet en particulier s'accompagne d'une banalisation de ces outils et de l'expérience de leur usage. Socialement attesté et délimité, le fait d'« aller sur le Web » est à ranger au même rang que « prendre un café » ou « passer un coup de téléphone » : si la dénomination n'en réduit pas la complexité ni la diversité en termes de pratiques et de situations, elle désigne une activité identifiée et bornée dans le temps qui légitime son étude en tant que telle.

Comprendre les parcours en situation

L'appréhension de cette activité particulière d' « aller sur Internet[1] » intéresse plusieurs champs disciplinaires, qui ne l'interrogent pas de la même manière, n'y valorisent pas les mêmes éléments – lorsque, travaillant sur le même aspect, ils n'y portent pas des vues divergentes du fait de prémices diamétralement opposées.

Puisqu'il s'agit d'activité, situons tout d'abord son cadre : on se limitera à l'accès à Internet par un terminal « classique » (informatique personnelle, de type PC), c'est-à-dire en excluant les accès *via* la téléphonie mobile et les assistants personnels. L'évolution des terminaux eux-mêmes et de leur interopérabilité bouleversera peut-être cette division, en autorisant des modes d'accès nomades (dans la rue, dans le train, etc.), semi-nomades (on pense par exemple aux bornes WiFi dans le cadre de l'accès à un réseau d'entreprise) ou mobiles (au sein de l'espace domestique). Posons qu'aujourd'hui encore, pour aller sur Internet, il faut un ordinateur, c'est-à-dire un terminal relativement volumineux composé d'un écran, d'un clavier et d'un dispositif de pointage – souris ou autres sur ordinateurs portables.

Cette question est d'importance dans le courant de l'action située : en appréhendant la navigation comme pratique incorporée, on ne peut manquer de s'attacher à décrire non seulement le contenu de l'écran, mais également le couplage de l'individu avec les dispositifs techniques de médiation (le clavier, la souris, etc.) et son entour, en termes de perturbation, de sollicitation ou de co-construction de l'activité[2]. Notre approche laissera de côté ces aspects, même si nous partageons avec les tenants de l'action située le souci de placer les pratiques en contexte. En centrant notre analyse non pas sur la pratique incorporée, mais sur la pratique à l'écran, on laissera de côté la question du couplage de l'individu avec l'outil technique, mais ce faisant, on met l'accent sur ce qui conduit l'activité, la guide et l'organise. Nous ne cherchons pas à minimiser les déterminations qu'induisent sur le cours d'action les sollicitations auxquelles l'individu est soumis lors d'une activité de navigation ; mais gageons qu'elle est au moins autant visible à l'écran que devant, et que c'est au sein même du passage de page en page et de site en site, dans le choix des contenus visités, dans la longueur des séquences de visionnage, leur rythmique, leur agencement interne, que se trouve l'essentiel d'un parcours sur le Web.

On traitera ici de la navigation sur le Web uniquement ; reconnaissons qu'il s'agit d'une réduction de l'utilisation d'Internet, qui permet également de faire de la messagerie asynchrone ou instantanée, des jeux, du téléchargement de fichier, etc. Nous laissons en particulier de côté les éléments d'entrelacement entre ces différents supports de l'accès aux ressources disponibles sur le réseau : un utilisateur peut suivre un lien à partir d'un mail, puis répondre par mail à l'expéditeur pour lui donner son avis sur le site en question ; dans le cadre d'une séance de *chat*, le Web peut être mobilisé comme ressource externe et support de conversation de manière

---

[1] « Internet » recouvre l'ensemble des applications, ressources et outils disponibles *via* le réseau : Web, messagerie, *peer-to-peer*, chat, etc. Le Web se restreint à l'accès à des sites par le biais d'un navigateur (voir Glossaire).

[2] Voir notamment les observations rapportées dans [Relieu & Olszewska 2004].

ponctuelle[1] ; dans une pratique de *peer-to-peer*, la recherche de fichiers à télécharger sur les réseaux d'échange peut se faire *via* une interface Web ; etc. Ces éléments ne doivent pas nous interdire de restreindre l'étude au Web : considérons que les autres outils Internet sont à mettre sur le même rang, pour ce qui est de l'analyse des comportements de navigation sur la Toile, que les éléments externes à Internet – programmes et documents présents sur l'ordinateur, matériau à la disposition de l'internaute en situation de navigation, interactions avec d'autres personnes en coprésence ou par téléphone.

Nous nous concentrerons donc sur l'activité de navigation « dans l'écran » et posons que, au titre d'activité, elle prend sens et se construit *en contexte*. En observant les parcours de page en page et de site en site, on se place à la croisée de chemins très divers, de situations variées, et dans des types de pratiques très différentes. De la même manière qu'une étude de la lecture doit prendre en compte les situations sociales dans lesquelles celle-ci se produit autant que ses supports et ses contenus, l'analyse des parcours se situe dans une diversité de pratiques sociales qui en modifient profondément le sens. Notre travail s'attachera à décrire des régularités et des modalités particulières de la pratique du Web, et les spécificités de ces modalités en regard des situations. Dans cette perspective, nous posons que l'appréhension des contenus est contextuelle, et ce à double titre : d'une part, deux personnes n'appréhenderont pas de la même manière une même page ou un même site, et d'autre part, un même contenu pourra être appréhendé différemment par un utilisateur donné dans deux contextes différents.

En ce sens, la sémantique des parcours que nous proposons se situe dans la perspective de la Sémantique Interprétative textuelle de F. Rastier[2], en reprenant au palier de l'appréhension des contenus du Web les éléments que la Sémantique Interprétative a mis à jour du côté des textes : détermination du local par le global, inscription des pratiques dans des genres et des situations, construction contextuelle du sens. Elle ne peut toutefois s'y réduire, pour deux raisons : d'une part, la dimension hypertextuelle du média tend à briser les unités textuelles en privilégiant le fragment et la recomposition d'un ensemble à partir de sources au sein du parcours. Au palier méso-sémantique, le déploiement des isotopies s'en trouve bouleversé, et ne peut être étudié qu'à travers l'infinité de corpus produits par la pratique : le site, qui pourrait correspondre au livre dans la mesure où il correspond à une unité éditoriale assumant une thématique et une topique spécifiques, n'est plus une unité d'analyse systématiquement pertinente en termes de réception et donc de plan interprétatif. D'autre part, les contenus Web ne peuvent être envisagés sous l'angle des textes uniquement : il ne s'agit pas uniquement du caractère multimédia des contenus (images, bandeaux, menus interactifs, etc.), qu'une sémantique textuelle est à même de prendre en compte, mais du fait que le Web propose, outre des contenu à « lire », de l'outillage. Celui-ci se compose d'outils de recherche, de communication (WebMail, forums, WebChat, etc.), de jeux, d'achat en ligne, etc. qui

---

[1] Voir par exemple [Beaudouin & Velkovska 1999].

[2] Voir notamment [Rastier 1987] et [Rastier 1989].

valent en tant que support d'activité et induisent des séquences d'action bien distinctes des contenus qu'ils véhiculent[1]. Ce n'est donc pas, une sémantique textuelle, mais une théorie de l'action qui permettra d'appréhender les parcours sur la Toile ; on posera ainsi, en suivant F. Rastier dans le projet d'une Sémiotique des cultures, que :

> Les théories de l'action ont privilégié l'axe de la représentation, et notamment le rapport entre les représentations et la motricité, c'est-à-dire les deux niveaux périphériques de la zone identitaire. Le niveau sémiotique est resté peu questionné en tant que tel. Si l'on tient compte maintenant de l'axe de l'interprétation, il faut prendre en considération les trois disciplines qui s'y articulent, d'ailleurs non sans de notables différences de statut.
> (i) Au niveau des présentations, la phénoménologie et la psychologie rivalisent pour décrire le flux de conscience comme activité.
> (ii) Au niveau sémiotique, l'herméneutique matérielle, entendue comme organon de la sémiotique des cultures, décrit les cours d'action sémiotiques. Discipline subordonnée, la sémantique interprétative prend spécifiquement pour objet le sens des textes.
> (iii) Au niveau physique, la praxéologie comprend une kinésique, mais aussi une technologie qui inclut les techniques du corps. Ainsi se dessinerait un regroupement de disciplines considérées comme désuètes ou marginales : mais l'ergonomie, la technologie, la sémiotique, l'herméneutique, la phénoménologie, toutes conçues comme des disciplines de la raison pratique, pourraient y trouver le lieu de rencontres nécessaires. Un de leurs premiers enjeux pourrait être de redéfinir la notion de pratique.[2]

Une sémantique des parcours développera donc une approche se réclamant d'une *herméneutique matérielle qui se concentre sur la dynamique production / réception* et sur sa manifestation et son déroulement temporel et ordonné dans le cadre du parcours. Partant, nous faisons l'hypothèse que *forme et contenu des parcours sont liés*, et que c'est au sein de cette dynamique que se mettent en place des structures actantielles où l'on cherchera à trouver des invariants et des situations prototypiques, à travers l'observation des pratiques effectives et de leur diversité.

Les différents paliers de l'analyse

Un parcours sur le Web s'apparente à un parcours de lecture et d'action dans le cadre d'une navigation hypertextuelle. Il se présente, pour un utilisateur donné, comme un cheminement régi par une série de contraintes internes (le projet général de l'utilisateur, ses compétences, les contenus visités) et externes (le dispositif technique hypertextuel qui sous-tend la navigation, les contenus proposés, leur organisation et leur présentation par les producteurs, leur accessibilité) au sein du

---

[1] On rejoint d'ailleurs ici F. Rastier, qui propose dans [Rastier 2003] une classification des « choses » en trois types : « les outils (en comprenant par là aussi les outils de communication comme les médias et les commandes, informatiques par exemple) ; les signes (linguistiques ou non : mots, symboles, chiffres, etc.) ; enfin les œuvres, qui sont issues d'une combinaison de signes ».

[2] [Rastier 2001b], pp. 212-213.

Web. Le terme de « parcours » inclut dans son étymologie des éléments qui recouvrent à notre sens les différents aspects des parcours sur Internet : 1) courir sans s'arrêter, courir en toute hâte ; 2) parcourir, traverser ; 3) parcourir du regard, lire, voir ; 4) parcourir par la parole, passer rapidement sur (un sujet), passer rapidement en revue, glisser sur, effleurer. Dynamique, transversalité, lecture multimédia et interaction sont des éléments spécifiques et fondamentaux des parcours sur le Web.

On peut décomposer l'analyse d'un parcours en cinq échelles distinctes, bien que celles-ci soient étroitement entremêlées (nous reviendrons par la suite sur ces interactions), avec au sein de chacune d'elle une spécificité de la confrontation entre l'internaute et le matériau multimédia à sa disposition.

1. *niveau micro* : une suite de pages vues.
   La sémantique des parcours prend comme palier inférieur de l'analyse de la page Web en tant qu'unité ergonomique et navigationnelle élémentaire. À ce niveau d'analyse, on s'attache à décrire le contenu de la page, que ce soit par des ressources externes ou par l'analyse de son contenu textuel et de ses propriétés (texte statique / dynamique, requête, page simple / complexe, utilisation de scripts côté client, etc.). Il est à noter ici que l'on se place, en terme de « pages vues », du point de vue de l'utilisateur sur le plan ergonomique, ce qui implique une reconstitution de la page à partir des différents éléments qui peuvent la composer, dans la mesure où une page telle qu'elle est vue peut être le résultat d'une série de requêtes. La page Web apparaît, de ce point de vue, comme un assemblage dynamique plus que comme un objet figé et clos sur lui-même.

2. *niveau mini* : une visite au sein d'un site
   Le niveau *mini* s'attache à décrire le parcours d'un utilisateur au sein d'un site donné, et, dans ce cadre, la rencontre dynamique entre l'ensemble que forme le site en termes de discours et d'unicité éditoriale et le chemin qu'y suit un utilisateur, la manière dont il appréhende, sélectionne son contenu et participe à son élaboration.

3. *niveau méso* : un parcours cohérent de lecture et d'action.
   À ce niveau d'analyse, on s'intéresse à l'articulation entre forme et contenu du parcours au sein de la temporalité de la session. Ici, une sémantique des parcours s'attachera d'une part à analyser et représenter ce qui relève de la « topologie » du parcours, à savoir les phénomènes de revisite, de détour, de retour en arrière, etc. ; d'autre part, elle étudiera l'articulation des contenus des différentes pages et sites visités au cours de la session, ce qui l'amènera à s'interroger sur les principes de cohérence alors à l'œuvre (notions d'activité intentionnelle, de projet, de surf, etc.).

4. *niveau macro* : un ou plusieurs cours d'action au sein d'une session.
   Dans le cadre borné dans le temps que constitue la session, on s'attachera à décrire ici le nombre, l'articulation, l'enchaînement ou l'entrelacement des différentes séquences homogènes de navigation. À ce niveau d'analyse, on cherchera à voir quels peuvent être les enchaînements typiques entre les

différents cours d'action (par exemple : portail de FAI[1] – WebMail – recherche), et les formes qu'ils peuvent prendre en fonction des contenus visités.

5. *niveau méga* : un parcours d'utilisateur parmi d'autres.
   La méga-sémantique des parcours replace le parcours dans le cadre de projets et de contextes d'usage particuliers qu'elle s'attache à décrire. Elle vise à découvrir et analyser des invariants au sein des différentes sessions envisagées au niveau *macro*, et examine les corrélations que ces constantes entretiennent avec des éléments extérieurs au parcours, tels que l'expérience de l'utilisateur, sa connaissance du « thème » du parcours, le fait qu'il ait déjà visité tel ou tel site auparavant, etc. À cette échelle, la sémantique des parcours observe l'articulation de l'ensemble des sessions entre elles pour un utilisateur donné, et, symétriquement, les différences et les similitudes entre sessions d'utilisateurs différents.

Ce que nous pouvons résumer dans le tableau suivant :

Tableau 1.1 - Sémantique des parcours Web : grille analytique

	niveau d'analyse du support de l'action	*niveau d'analyse de l'action*	*éléments assemblés*	*objet décrit*
micro	page	appréhension de l'interface	composition des requêtes formant les pages	contenu en termes thématique et fonctionnel
mini	site	navigation à l'intérieur d'un site	pages visitées sur le site	contenus proposés / accédés
méso	assemblage de pages sur un ou plusieurs sites	chaîne opératoire mobilisant pages et sites	une/plusieurs pages, sur un/plusieurs sites	routine de navigation
macro	session	séquence d'activité Web bornée dans le temps	groupes de pages regroupées en sites	organisation séquentielle des routines
méga	utilisateur	activité inscrite dans les pratiques, routines	les sessions d'un utilisateur	pratiques de l'utilisateur

À travers cet appareil analytique allant de la page à l'utilisateur, une sémantique des parcours dispose d'un cadre de travail pour l'analyse des usages du Web qui permet de prendre en compte l'ensemble des phénomènes en jeu du côté de la production des contenus comme de leur réception.

[1] Fournisseur d'Accès à Internet.

1.1.2 Un champ d'études encore nouveau

L'analyse de parcours sur le Web demeure un champ de recherche assez peu exploré ; elle hérite certes des travaux menés auparavant sur les hypertextes et les interactions homme-machine, mais dans le cadre du Web, elle est bien souvent réduite à l'observation de la navigation sur un site en particulier, ou à des pratiques ciblées « en laboratoire », et bien peu d'études ont pu travailler sur des données d'usage à grande échelle, et sur des utilisateurs observés en situation naturelle.

Comportement d'utilisateur

Les questions que nous soulevons ici ont peu été traitées jusqu'alors, ou sous un angle qui ne nous satisfait pas entièrement. D'un côté, les sciences cognitives ont mis en avant, au sein d'un paradigme sujet/objet, un modèle de l'activité humaine comme « système de traitement d'information » ; appliqué à la navigation, cette approche se retrouve dans les nombreux travaux issus du champ de la Recherche d'Information et de l'Intelligence Artificielle. L'approche informatique commune des parcours réduit ainsi les contenus Web à des informations, ou au mieux à un espace documentaire, dont le principal défaut est de n'être ni structuré, ni hiérarchisé.

Ces postulats se retrouvent dans la plupart de travaux sur les hypertextes, antérieurs au Web, qui portent alors principalement sur la modélisation de l'utilisateur à travers l'étude de ses parcours dans un système hypermédia donné ; les applications sont alors tournées vers les recommandations de conception et surtout la mise en place d'hypermédias adaptatifs (*adaptive hypermedia*). Nous renvoyons à la lecture de [Brusilovsky 1996] pour un panorama très complet des problématiques, des méthodes et des applications relatives aux hypermédias adaptatifs avant l'émergence du Web, complété en 2001 dans [Brusilovsky 2001] pour les études centrées sur le Web. Dans leur ensemble, les travaux décrits ne visent pas tant la description des pratiques que, à travers une modélisation du comportement, une meilleure conception des systèmes hypertextuels, en particulier dans le champ des sciences de l'éducation (application à des encyclopédies, des méthodes d'apprentissage multimédia) et de la recherche d'information, proche de l'ingénierie documentaire.

Avec l'apparition du Web, toute une série de travaux a suivi cette voie en conservant les paradigmes issus des études sur les hypermédias ; ces recherches se situent dans le champ des sciences cognitives et s'orientent vers la modélisation de l'utilisateur en situation de navigation sur le Web. Nous renvoyons à la lecture de [Modjeska 1997] pour un panorama certes un peu daté des travaux effectués dans ce domaine, mais qui rend bien compte des problématiques soulevées par cette approche, qui fait la part belle aux perceptions, aux « structures cognitives » et aux « modèles mentaux » de l'utilisateur. Dans la plupart des cas, il s'agit d'études centrées-utilisateur sur des échantillons restreints, parfois tournées vers l'« usabilité » d'un site en particulier et le problème de la « désorientation » des utilisateurs, mais le plus souvent orientées vers la recherche d'information. Ce paradigme, directement hérité de l'ingénierie documentaire, domine encore la recherche sur la navigation Web, où les contenus sont assimilés à des documents contenant des « molécules informationnelles », avec en arrière-plan une vision orientée « exécution de tâche » et résolution de problème. On trouve ainsi un certain nombre de travaux sur les

stratégies des utilisateurs en recherche d'informations (Choo sur les *knowledge workers*[1], Jansen sur les usages du moteur de recherche Excite[2]) ou encore sur l'usage de certains types de sites particuliers[3].

Ces travaux ne sont pas dénués d'intérêt, car ils permettent d'isoler, dans un contexte particulier, des questions précises sur les comportements : soulignons ici une étude particulièrement intéressante, *Web Search Behavior of Internet Experts and Newbies* menée par Hölscher et Strube en 2000 ([Hölscher & Strube 2000]), qui montre, dans un contexte de recherche d'information, l'importance de la double expertise des utilisateurs en termes de maniement des outils de recherche et de navigation sur le Web, mais aussi de connaissance du domaine sur lequel porte la recherche. Pour autant, l'approche est ici réductrice, car elle isole l'utilisateur de son cadre habituel d'activité, avec toutes les variations et perturbations que celui-ci peut comporter, et elle lui impose des activités qui ne sont peut-être pas du tout représentatives de ses pratiques.

L'étude menée par Byrne en 1999 tente de répondre à ces problèmes en proposant une approche globale de l'usage s'appuyant sur un dispositif de recueil de données original dans le champ de l'analyse de « tâches » des utilisateurs. En 1999, Byrne *et alii* proposent une *taskonomy* de l'usage du Web[4], qui rend compte de l'analyse à l'aide de la vidéo de l'activité Web de dix personnes pendant une journée, et a pour but de comprendre les tâches engagées par l'utilisateur quand il navigue au quotidien. Les participants, des utilisateurs expérimentés du Web, sont soumis à un double enregistrement vidéo, pointé sur eux et sur leur écran, qu'ils mettent en marche lorsqu'ils naviguent. Ils sont en outre invités à commenter oralement leurs actions pour faciliter le travail de dépouillement des données par la suite. L'analyse des vidéos décompose la navigation en tâches à deux niveaux de codage ; au premier niveau, les actions de base comptent huit catégories : *use information*, *locate information*, *provide information*, *find on page*, *navigate*, *configure browser*, *manage window* et *react to environment* ; chacune de ces catégories se décompose ensuite en sous-catégories plus fines. Le résultat majeur de l'étude est l'imbrication des tâches entre elles : une tâche peut générer n'importe quel autre type de tâche, et la plus fréquente tâche, *Use Information*, génère d'autres tâches du type *Locate Information*, *Navigate* et *Find On Page*. En outre, l'étude fournit des observations avancées sur certaines tâches :

- dans la tâche *Use Information*, décomposée en *reading*, *print*, *duplicate*, *view*, *listen*, et *download*, la sous-tâche la plus fréquente est *reading*, ce qui replace la lecture au centre de la navigation et de l'appréhension des contenus ;
- pour la tâche *Locate Information*, le moteur de recherche s'impose comme le point de départ privilégié ;

---

[1] Voir [Choo *et al.* 1999] et [Choo *et al.* 2000].

[2] Voir [Jansen *et al.* 1998a] et [Jansen *et al.* 1998b].

[3] Par exemple, [Jones *et al.* 1998] sur les bibliothèques électroniques.

[4] Voir [Byrne *et al.* 1999a] et [Byrne *et al.* 1999b].

- pour *Find On Page*, le sous-type le plus fréquent est de loin *related*, par rapport à *image*, *interesting*, *string* et *tagged* ; mais en durée, les cinq sont équilibrés.

Enfin, les auteurs observent que dans la manière d'accéder aux pages, le lien hypertexte fait plus de la moitié des requêtes, tandis que les actions de type *back* et *autre* se partagent le reste.

L'entreprise taxinomique de Byrne est d'autant plus intéressante qu'elle se fonde sur des observations de la vie « de tous les jours », et qu'elle tente d'embrasser la diversité des pratiques. Toutefois, nous ferons à cette étude la même critique qu'à l'approche cognitiviste des parcours sur le Web : contenus proposés, modes d'accès et activité de navigation sont enfermés dans le paradigme de la recherche d'information et, par extension, les contenus du Web sont valorisés sous cet angle unique. En arrière-plan, se dessinent les approches mentalistes issues des sciences cognitives qui réduisent les parcours sur le Web à la réalisation d'un projet par un sujet à l'aide de l'outil technique que constitue le Web. Ce faisant, l'approche cognitive conclut à des équivalences entre motifs de navigation, tâche et motivation de l'utilisateur qui sont réductrices dès lors que l'on examine la diversité de l'offre de contenus sur le Web autant que les usages qui en sont faits.

Dès lors, nous laisserons volontiers de côté ces approches orientées modélisation pour notre analyse, et y opposons une approche descriptive pragmatique qui s'attache à replacer les modes de navigation dans le cadre de pratique avérées, et à prendre en compte la singularité des situations, des contenus et des individus. On cherchera certes des invariants dans l'ensemble des parcours observés, mais sans les relier à de quelconques modèles mentaux ou psychologiques. Pour cela, nous empruntons à Leroi-Gourhan[1] la notion de « chaîne opératoire », définie comme un processus de travail qui mène d'une matière première à un objet fini. Se décomposant en une série d'étapes, la chaîne opératoire intègre un projet, un savoir-faire, un geste, une matière première, un outil ; elle s'articule et s'imbrique avec d'autres chaînes, qu'elle peut croiser et influencer. Appliqué plus spécifiquement aux parcours sur le Web, ce concept de chaîne opératoire permet de rendre compte à la fois des aspects techniques liés au maniement de l'outil informatique en général et du Web en particulier, de l'importance des connaissances et du savoir-faire de l'utilisateur, et de l'implication de ces deux éléments dans des « projets » qui sous-tendent la navigation. Par opposition à la tâche, le projet engage et construit les savoir-faire, peut être décomposé en plusieurs chaînes opératoires, et ne cesse de s'élaborer, de se redéfinir et de s'alimenter au fur et à mesure qu'il se réalise. L'ambition d'une sémantique des parcours sera dès lors de mettre à jour des chaînes opératoires récurrentes, et d'examiner les structures, les savoir-faire, l'outillage et le déploiement temporel et rythmique des mouvements et de la gestuelle navigationnels.

Web Mining et données de trafic

Avec le développement du Web tant du côté de l'offre de contenus que de l'accès, s'est constitué depuis la fin des années 90 un champ de recherche autour du « Web

---

[1] [Leroi-Gourhan 1943] et [Leroi-Gourhan 1964].

Mining ». Plutôt orienté vers l'analyse de données, les méthodes statistiques et les aspects applicatifs, le Web Mining se divise en trois domaines : Web Content Mining pour l'analyse des contenus, Web Structure Mining pour l'étude globale de l'organisation hypertextuelle de la Toile, et Web Usages Mining sur le plan des usages[1]. Dans ce dernier, qui a fait des données de trafic son matériau privilégié, on distingue communément les approches centrées-serveur (*site-centric*), qui traitent de données recueillies sur un site particulier, et centrées-utilisateur (*user-centric*) qui se basent sur des informations collectées du côté de l'utilisateur.

Le travail que nous présentons ici, basé sur l'analyse de traces de navigation collectées sur les postes des internautes, se place résolument dans cette dernière approche. Bien qu'elle soit la seule approche appropriée pour l'analyse des usages, elle est difficile à mettre en œuvre, et demeure peu pratiquée : très peu d'études ont comme matériau des données de navigation enregistrées du côté de l'utilisateur en situation d'usage réelle – nous en comptons au total quatre.

La première d'entre elles est celle de Catledge et Pitkow en 1995, *Characterizing browsing strategies in the World-Wide-Web*[2]. Travail fondateur et unique, cet article présente l'analyse de données recueillies au niveau du navigateur M*osaic* pendant trois semaines auprès de cent sept utilisateurs. Celles-ci contiennent non seulement les URL visitées par l'utilisateur et la date, mais aussi l'ensemble des actions sur le navigateur ; l'ouverture d'une page est ainsi décomposée en :

- *Selection of hyperlink in document*
- *Go back one document*
- *Open file via a URL*
- *Go to document via Hotlist*
- *Go forward one document*
- *Open local file*
- *Go to the Home document*
- *Go to document via Window History*

L'étude montre un certain nombre de résultats intéressants :

1. Découpage en sessions : le temps moyen entre deux actions sur le navigateur est de 9,3 minutes. Partant, le temps d'inactivité retenu pour définir la fin d'une session est de 25,5 minutes.
2. Séparation par protocole : 80 % de la navigation se fait par HTTP, dont 4 % générés par CGI (contenu dynamique).
3. Méthode d'interaction : 52 % des actions de navigation sont faites par le suivi de liens dans les pages, et 41 % par des *Back*. Les raccourcis clavier ne sont presque jamais employés.
4. Séquences répétées : une corrélation linéaire est observée entre le nombre moyen de pages vues sur un site dans une visite et le nombre de visites du

---

[1] Pour une vue générale et synthétique du champ du Web Mining, on trouvera dans [Kosala & Blockeel 2000] et dans [Srivastava *et al.* 2003] des informations claires et une sélection d'articles.

[2] [Catledge & Pitkow 1995].

site : beaucoup de sites sont vus sur une faible longueur, peu sur un longueur importante.

5. À l'intérieur d'un site, les auteurs constatent une stratégie *spoke and hub*, c'est-à-dire une très forte utilisation du *Back* pour des séries d'avant-arrière autour d'une page-pivot. Ceci suggère que cette forme de navigation est indépendante du nombre de liens proposés dans une page.
6. Autre méthode de navigation souvent observée : les pages personnelles comme une sorte d'index vers des pages intéressantes.

Ces résultats commencent à être un peu anciens en regard de l'évolution du Web et de la diffusion de son accès ; cela étant, les conclusions et la méthode n'en sont pas moins intéressantes, en particulier dans la capacité à lier les pages vues à des modes d'interaction sur le navigateur.

À la même époque, Crovella et Bestavros publient *Characteristics of WWW client-based traces* en 1995 avec Cunha[1], et *Self-similarity in World Wide Web traffic - Evidence and possible cause* en 1996[2]. Les deux articles présentent l'analyse de traces de navigation côté-client sur près d'un million de requêtes ; comme dans [Catledge & Pitkow 1995], le navigateur *Mosaic* sert de support au recueil de données : 37 ordinateurs partagés (stations de travail Sun) sont équipées du dispositif de recueil dans des salles d'une université d'informatique. Entre autres résultats intéressants, nous notons le constat que le trafic répond à une loi de puissance (*power-law distribution*) : une faible part des documents concentre la majorité des requêtes, tandis qu'un grand nombre de pages ne sont vues que très rarement. Cette loi est également observée en ce qui concerne la taille des fichiers, ce qui intéresse particulièrement les deux auteurs dont la perspective est d'améliorer les techniques de *cache* afin d'augmenter la rapidité d'accès aux pages Web.

La troisième étude, présentée en 1997 par L. Tauscher et S. Greenberg[3] se penche sur la « revisite » de pages Web : l'objectif est d'élaborer des modèles de revisite, et d'en tirer des conclusions pour la conception des systèmes d'historique des navigateurs. Les données analysées sont constituées par les traces de navigation de 23 utilisateurs observés pendant six semaines. L'analyse montre qu'en moyenne 58 % des requêtes pour un utilisateur donné pointent vers une page qu'il a déjà visitée, en même temps que le « vocabulaire » des URL ne cesse de croître avec le temps. Des entretiens ont été menés par la suite avec les utilisateurs, qui montrent que les causes de la visite de nouvelles pages sont essentiellement : 1) le besoin de nouvelles informations ; 2) le désir d'explorer un site en particulier ; 3) la page est recommandée par un collègue, et 4) la page a été trouvée en cherchant autre chose. L'étude croise également ces données avec un enregistrement des actions sur le navigateur similaire à celui pratiqué dans [Catledge & Pitkow 1995].

---

[1] [Cunha *et al.* 1995].

[2] [Crovella & Bestavros 1996].

[3] Voir [Tauscher & Greenberg 1997a] et [Tauscher & Greenberg 1997b].

Enfin, en 2000, Cockburn et McKenzie présentent *What do Web users do ? An empirical analysis of Web use*[1]. La méthode de recueil de données consiste à utiliser les fichiers *history.dat* et *bookmark.html* du navigateur Netscape pour 70 utilisateurs (membres de la faculté) sur 4 mois (d'octobre 1999 à janvier 2000) ; pour chaque page, sont notés, outre l'URL, les dates de premier et de dernier accès et le nombre de visites, avec une collecte chaque jour. Dans les données récupérées, n'apparaissent que les URL explicitement demandées par l'utilisateur, les paramètres des CGI se trouvent tronqués, et les différentes pages des *frameset* sont regroupées sous une seule entrée. Cockburn et McKenzie font plusieurs constats :

- croissance assez régulière du vocabulaire (les pages) dans le temps, de manière globale ainsi que pour chaque utilisateur.
- forte corrélation entre le nombre de visites et le vocabulaire. En moyenne, quatre visites pour une page (« for each new URL added to the overall vocabulary, four pages are revisited »).
- taux de revisite : sur l'ensemble du panel, le taux de revisite est de 81 %.
- pour chaque utilisateur, peu de pages sont visitées très régulièrement (en moyenne, 24 % du vocabulaire de chaque individu) : on retrouve un comportement de type « loi de Zipf ». Souvent, les utilisateurs ont des raccourcis vers les deux pages les plus vues (*home page* ou *bookmark*).
- la majorité des pages sont vues une seconde ou moins : « browsing is a *rapidly* interactive activity ».
- *bookmarks* : ils sont nombreux chez tous les sujets. Dans le temps, le nombre d'ajouts est supérieur au nombre de suppressions, de sorte que la taille des *bookmarks* ne cesse d'augmenter.
- les utilisateurs, bien qu'appartenant au même département de l'université, voient des espaces très différents sur le Web : 91 % des pages visitées hors du site de l'université n'ont été vues que par un seul utilisateur.

Les auteurs concluent sur les implications de ces résultats pour la conception de navigateurs, ce qui ne nous intéresse pas directement ici, mais leurs résultats empiriques sont tout à faire réutilisables pour l'analyse des parcours.

Ces études sont très précieuses, en premier lieu parce qu'elles proposent des pistes pour l'exploitation de données de trafic, ce qui nous intéresse particulièrement étant donné notre matériau, mais également parce qu'elles traitent de la navigation du côté de l'utilisateur et en situation « réelle ». Le faible nombre de travaux « centrés-utilisateur » tient à la rareté des données de ce type, et ces quatre travaux fournissent d'importants résultats en termes statistiques et méthodologiques. Ils esquissent une image de la navigation ouvrant l'utilisateur à un nombre toujours renouvelé de pages, au sein d'une pratique complexe où les fonctionnalités des navigateurs autant que les contenus des pages entrent en ligne de compte. Deux reproches peuvent toutefois leur être faits : d'une part, les données concernent toujours des catégories particulières d'utilisateurs (étudiants en informatique le plus souvent), et pour une durée relativement limitée (quatre mois au maximum). D'autre part, le contenu des pages n'est jamais abordé : il semblerait particulièrement intéressant de savoir quelles corrélations peuvent exister entre stratégies de navigation, visite de nouveaux sites et

[1] [Cockburn & McKenzie 2000], repris dans [McKenzie & Cockburn 2001].

thème ou service proposés par les pages accédées. Il entrera donc dans les objectifs d'une sémantique des parcours d'avoir une approche dynamique tenant dans un même temps les logiques de production et de réception.

1.2 Au croisement de deux dynamiques

Le travail que nous présentons ici se base principalement sur l'analyse de traces de navigation recueillies du côté de l'utilisateur. Dès lors, à l'enjeu descriptif des modes de parcours, s'ajoute celui de l'exploitation des traces d'usage : pour remplir les objectifs d'une sémantique des parcours, nous devons à la fois être capables de décrire finement leur double contexte, celui des contenus comme celui de l'internaute, mais également donner sens à ce mouvement en intégrant ces descriptions au sein de la dynamique du parcours.

La distinction méthodologique en paliers que nous avons proposée ainsi que leur ordre de présentation ne doivent en rien masquer les interactions fortes qui existent entre les cinq échelles d'analyse : de la même manière que la Sémantique Interprétative pose la détermination du local (thématique, dialectique, dialogique, tactique) par le global (genres, discours, pratiques), la sémantique des parcours que nous proposons postule la double primauté du projet et du savoir-faire de l'utilisateur (niveaux *macro* et *méga*) sur les contenus visités et leur articulation (niveaux *micro*, *mini* et *méso*). Entre ces cinq paliers d'analyse, se jouent des influences réciproques qu'une sémantique des parcours se doit d'explorer et d'expliciter. À titre d'exemple, on peut citer :

- *méga vers micro* : les centres d'intérêt de l'utilisateur définissent la thématique et la fonction des pages visitées. L'expertise ergonomique de l'internaute, sa connaissance de la sémiotique des pages Web modifie son appréhension des pages. Certaines études[1] ont ainsi montré chez les primo-accédants, la confusion entre adresses Web et adresses de messagerie, ou que les bannières publicitaires ne sont pas identifiées en tant que telles, ce qui induit des incompréhensions et des détours au sein du parcours.
- *macro vers micro* : l'appréhension du contenu d'une page est elle-même fonction de la position de la page dans la session, et de la façon dont elle s'inscrit dans le projet en cours : la lecture d'un contenu donné met en jeu un processus interprétatif qui dépend de l'utilisateur et de la chaîne opératoire dans laquelle il se place. Par exemple, un site comme celui de la Fnac peut être appréhendé dans sa fonction première de catalogue de vente en ligne, mais aussi comme répertoire d'opinions d'internautes sur un produit.
- *méso vers méga* : les contenus et services proposés par un site peuvent amener l'utilisateur à des pratiques régulières. Par exemple, un site fournissant le programme télévisé va plaire à l'utilisateur, qui y reviendra fréquemment.

---

[1] Voir par exemple [Relieu & Olszewska 2004] ou [Cotte 2002].

- *méso vers mini* : l'ensemble du contenu du site a une incidence directe sur le niveau *mini* (un site ne peut pas proposer ce qu'il n'a pas). Les chemins des visiteurs au sein d'un site sont susceptibles d'entraîner, à court ou moyen terme, une réorganisation de ce site ; ceci est particulièrement le cas pour les sites marchands qui souhaitent améliorer leur navigabilité.
- *méso vers macro* : la propension des sites à pointer vers d'autres sites, ainsi que le type de sites vers lesquels ils renvoient, ont une influence sur l'ensemble des sites visités dans la session : moteurs de recherche, sites personnels avec page de liens et site commerciaux ont à cet égard des influences très contrastées.
- *micro vers méso* : le contenu peut renvoyer à des activités hors Web, que ce soient encore des activités sur Internet (mail, *chat*, etc.) ou dans de tout autres domaines et sur de tout autres supports (le Web comme ressource d'informations). Un site peut proposer à l'utilisateur des informations sur la navigation, les outils du Web, ou des services d'aide à la navigation qui auront une influence sur son comportement général.
- *micro et mini vers méso et macro* : le type de pages et de sites visités contraignent à telle ou telle forme de navigation (par exemple : l'utilisation du WebMail passe par une authentification de l'utilisateur). Le parcours se trouve encadré dans une forme de schéma actantiel défini par le site qui guide son parcours. Les liens présents sur une page définissent des possibilités de navigation hypertextuelle. Certaines pages contiennent des scripts qui opèrent des actions d'ordre navigationnel : redirection, ouverture automatique d'une ou plusieurs fenêtres, etc. La nécessité pour l'utilisateur de disposer de *plugins* (briques logicielles additionnelles) au sein du navigateur, ou même d'un navigateur particulier, peut lui interdire l'accès à certaines pages ou certains sites.

Il ne suffit pas, bien évidemment, d'énumérer ces influences réciproques entre différents paliers, il importe de les quantifier, de les ordonner, et d'évaluer les cas où tel élément prend le pas sur tel autre et de quelle manière. Notre travail vise à décrire ces interactions, ainsi qu'à évaluer dans quelle mesure leur connaissance peut permettre d'élaborer un système d'analyse des parcours sur le Web en vue de l'étude des usages d'Internet. Dans la mesure où l'on traitera de données volumineuses pour y chercher des phénomènes récurrents, on cherchera à se doter de représentations synthétiques aux différents niveaux d'analyse des parcours, de la page à l'utilisateur.

1.2.1 Décrire les contenus

À l'échelle de la page, une sémantique des parcours doit prendre en compte la nature spécifique des contenus Web. Dans la littérature, ceux-ci sont très généralement considérés en termes d'informations ou de connaissances : c'est de cette manière que Vennevar Bush imaginait en 1945 le système de nature hypertextuelle qu'il baptisa Memex : « Un memex est un appareil dans lequel une personne stocke tous ses livres, ses archives et sa correspondance, et qui est mécanisé de façon à pouvoir être consulté de manière très rapide et très flexible. Il s'agit d'un

supplément agrandi et intime de sa mémoire. »[1]. Aujourd'hui encore, on retrouve l'assimilation des contenus Web à de simples molécules informationnelles au sein du projet du Web Sémantique, que son initiateur Berners-Lee définissait ainsi en 2001 : « Le Web Sémantique est une extension du Web actuel, où l'information a un sens bien déterminé, permettant une meilleure coopération entre les machines et les individus »[2].

Cette conception traverse la majorité de la littérature sur l'hypertexte, on la retrouve par exemple chez Balpe en 1996 :

> Mais c'est plus en termes de finalités que l'hypertexte se comprend […]. En effet, on ne peut oublier que l'hypertexte permet avant tout de mettre les capacités de calcul et de présentation d'un ordinateur au service de l'information structurée ou non, en réalisant des associations entre des éléments de nature différente, associations conduites par l'intelligence ou l'intuition de l'utilisateur.[3]

On la retrouve également dans l'approche documentaire du Web, ce qu'illustre notamment l'ensemble des interventions présentées dans le colloque en ligne Text-e organisé par la BPI en 2001, en particulier dans [Chartier 2003] ou dans [Broadbent & Cara 2003] ; elle se retrouve également dans l'approche pourtant orientée vers l'ethnométhodologie de [Ghitalla *et al.* 2003], où le Web est envisagé en termes d' « architecture documentaire numérique » (terme repris à Broadbent et Cara), et la dimension interactive des pages centrée sur la manipulation des interfaces[4].

Cette définition des contenus du Web en termes d'informations et de connaissances nous paraît doublement réductrice. Tout d'abord, si le Web, en tant que système hypertextuel, s'apparente dans certains cas à une collection de textes ou de documents multimédia, cela n'implique pas que l'on puisse assimiler leur contenu à la mise en forme d'informations ou de connaissances : le Web n'est pas simplement un objet hybride entre encyclopédie déstructurée et annuaire. Plus encore, le paradigme du document nous paraît insuffisant pour décrire la diversité des contenus accessibles sur le Web : outils de communication (WebMail, WebChat, etc.), espaces coopératifs, jeux en ligne, sites communautaires, tous ces éléments invitent à

---

[1] « A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. » ([Bush 1946]).

[2] « The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation » ([Berners-Lee *et al.* 2001]).

[3] [Balpe *et al.* 1996], p.18.

[4] Par exemple : « On pourrait croire, parfois que le document à l'écran revisite à sa façon les deux procédés essentiels qui ont façonné l'histoire des supports d'écriture : le *scrolling* épouse le principe du déroulement du parchemin, le tourne pages électronique rappelle l'agencement des pages d'un livre. L'écran lui-même peut épouser les contours d'une page autonome, la « page écran » faisant alors coïncider espace d'affichage et géographie du document. Mais ni l'écran, ni même parfois l'espace de la « page » comme ensemble signifiant, ne sont la mesure de l'activité. C'est d'abord de la maîtrise du fenêtrage que dépend cette dernière […]. » ([Ghitalla *et al.* 2003], p. 164).

envisager également la Toile comme un fournisseur de services et plus généralement d'outillage. L'attention particulière des travaux sur les écrits numériques portée aux interfaces Web, aux règles complexes de leur composition, aux différentes zones et leur statut sémiotique, à la valorisation des textes insérés dans ce contexte, semble faire presque oublier le « travail des serveurs » et, pourrait-on dire, ce que « le clic renvoie ». C'est dans cette perspective qu'on regrettera que dans les réflexions de Souchier ou Jeanneret sur les « écrits d'écran »[1], l'attention soit tellement portée sur la sémiotique des pages Web qu'elle en néglige d'entrer pleinement dans les éléments interactifs, et ne s'attache qu'à ce qui relève du texte et de la lecture ; certes, il est toujours question de lecture quels que soient les contenus, mais ce palier ne réduit ni n'explique la navigation elle-même. À titre d'exemple, lorsqu'un internaute joue aux échecs sur le site Yahoo, il lit la page, la position des pièces, l'horloge, les classements, mais avant tout, il joue aux échecs. L'appréhension des contenus Web se base donc sur la lecture, mais ne s'y réduit pas toujours, et une sémantique des parcours doit se doter d'une représentation des contenus qui tienne pleinement compte des types de contenus, des types d'activités qu'ils suggèrent et autorisent et des interactions avec les fournisseurs de contenus.

Dans cette optique, on se penchera avec attention sur les études visant à distinguer, selon l'approche retenue, des genres ou des types de pages et de sites Web à partir de corpus. Au sein de l'approche générique, on trouve en particulier les études présentées dans [Karlgren & Cutting 1994] et [Karlgren *et al.* 1998], [Dimitrova *et al.* 2002], [Rehm 2002], [Roussinov *et al.* 2001] et [Dillon & Gushrowski 2000]. Si les divisions génériques varient selon les auteurs, tous partagent le fait de travailler sur des corpus de pages et de chercher à caractériser leurs genres sur la base de leur contenu textuel mais aussi de traits structurels et présentationnels ; les traits retenus donnent ainsi une place importante au support HTML des documents et à leur dimension hypertextuelle : nombre de liens, liens interne ou externe au site, nombre et proportion d'images, etc. Si l'on peut discuter les choix faits pour fonder les distinctions génériques au sein de la production Web, ces études ont l'avantage de prendre en compte la spécificité des contenus dans leur dimension hypertextuelle et fonctionnelle (par exemple : Karlgren distingue un genre « FAQ »).

L'approche typologique propose une démarche plus inductive, tout en conservant les aspects fonctionnels des classes constituées : dans *The connectivity sonar*[2], Amitay *et alii* proposent une méthode de classification fonctionnelle des sites sur la base de leur structure interne, en dehors de toute analyse de contenu. Les auteurs font l'hypothèse que le type d'un site est étroitement lié à sa structure (sa taille, l'organisation de ses pages en répertoires et sous-répertoires, les liens internes et externes), et que celui-ci peut être retrouvé à partir de celle-là[3]. Dans une perspective

---

[1] Voir notamment [Jeanneret & Souchier 1999] et [Souchier 2000].

[2] [Amitay *et al.* 2003].

[3] « Since sites are created for different purposes and by different people, it should come as no surprise that they sport different designs: the sizes of the sites, the organization of the pages

différente, et plus large en ce qui concerne les traits retenus pour décrire les sites et les pages, les projets TypWeb et SensNet suivent une démarche similaire dans la description des contenus. L'objectif est de parvenir « faire émerger, de manière inductive, des typologies sur la base des corrélations observées entre des indicateurs portant sur l'outillage grammatical et le lexique, sur la structuration textuelle et hypertextuelle, et sur l'aspect multimédia. »[1]. Le projet s'appuie sur l'extraction de l'ensemble des éléments textuels, structurels et formels des pages et des sites, pour leur analyse à l'aide de traitements matriciels. Ce travail pointe les différences fortes entre types de pages à l'intérieur d'un site, fondées sur leur fonction dans le cadre de la navigation, donc spécifiques aux contenus Web. Il a ainsi mis en avant une distinction entre pages « à contenu » et pages « d'orientation », qui s'avère précieuse pour l'appréciation de la dynamique des parcours.

En somme, rejetant le paradigme informationnel ou documentaire des contenus Web, et privilégiant une approche issue de la linguistique de corpus et adaptée aux spécificités du Web, nous appuierons une analyse des contenus au sein des parcours sur la Sémantique Interprétative telle qu'elle a été définie par François Rastier[2]. En tant que textes, les pages Web s'inscrivent dans des pratiques d'écriture, des discours, des genres, et incluent en leur sein les dimensions thématique, dialectique, dialogique et tactique. Cela étant, comme nous l'avons déjà dit, nous estimons que cela n'est pas suffisant dans le cadre de la publication et des contenus Web, et qu'il est nécessaire de considérer également le Web en termes d'outillage. Pour répondre à ces objectifs, on tentera de se doter de représentations à partir de sources endogènes, par la constitution de corpus notamment, et exogènes adaptées, à l'aide des annuaires du Web.

En outre, pour compléter cette approche, l'analyse des parcours gagnerait à inclure dans ses représentations une vision globale du Web en termes d'interconnexion : nous retenons notamment la distinction faite dans [Broder *et al.* 2000], sur la base des liens hypertexte entre pages, entre un fort noyau très interconnecté, une zone qui mène à ce noyau mais à laquelle on accède difficilement, et à l'inverse une zone pointée par le noyau central mais dont il est difficile de sortir. De telles observations sont particulièrement intéressantes : dans une navigation de site en site, l'impossibilité de « sortir » d'un site ou au contraire l'impossibilité d'aller sur un site car aucun lien n'y mène sont des facteurs importants. Ramenées aux cheminements des internautes de site en site et aux pratiques observées dans la durée, ces éléments peuvent permettre d'expliquer la régularité des comportements et la construction de territoires personnels et de routines sur le Web.

---

in directories and subdirectories, the internal linkage patterns within the site's pages and the manner in which the sites link to the rest of the Web. »

[1] [Beaudouin *et al.* 2001].

[2] Voir [Rastier 1987] et [Rastier 1989].

1.2.2 Dynamique des parcours et des individus

La séquentialité est un aspect essentiel de la sémantique des parcours : de la même manière qu'un texte ne peut être considéré, sinon au prix d'une réduction importante de ses composants sémantiques, à un « sac de mots » ou de phrases, un parcours sur le Web ne saurait être réduit à une collection de pages. L'ancrage temporel des parcours conforte cette position ; mais plus encore, le parcours s'apparente à une série d'actions (de navigation) dont l'ensemble ordonné seul peut donner le sens. Cette question est perceptible dans bien des travaux : les réflexions autour de la notion de tâche, que l'on retrouve en particulier dans les travaux de Byrne *et alii* ([Byrne *et al.* 1999a] et [Byrne *et al.* 1999b]) et la *taskonomy* qu'ils proposent, les travaux sur les comportements d'utilisateurs en recherche d'informations ([Choo *et al.* 1999]), les études sur la revisite de pages ([Tauscher & Greenberg 1997a] et [Cockburn & McKenzie 2000]) ou les recherches plus générales sur la navigation Web ([Catledge & Pitkow 1995], [Huberman *et al.* 1998]). Cela étant, l'analyse du parcours comme suite ordonnée de pages est dans presque tous les cas réduite au nombre de pages vues sur un site, ou de sites différents visités.

Pour répondre à ce problème, on se tournera volontiers vers les études centrées-serveur, qui font au contraire la part belle à la recherche de motifs de navigation au sein des parcours sur un site donné. Une littérature relativement abondante traite de l'analyse des parcours d'utilisateurs d'un point de vue *site-centric*, sur la base de l'analyse des *logs* des serveurs Web. Un tel engouement s'explique par les enjeux économiques sous-jacents à ces recherches : les sites à vocation commerciale souhaitent disposer de données les plus précises possibles sur leur fréquentation, afin de savoir quelles pages sont les plus visitées, comment les utilisateurs y arrivent, et comment les faire « rester » plus longtemps sur le site. Nous renvoyons à la lecture des articles de Masand en 2000[1] et de Srivastave en 2003[2] pour un panorama de ces recherches. Si le point de vue centré-serveur ne correspond pas aux objectifs d'une sémantique des parcours centrée sur les pratiques d'utilisateurs, certaines méthodes employées pour l'analyse des visites de sites particuliers peuvent être mobilisées.

Dans ce cadre, trois études retiennent particulièrement notre attention : en premier lieu, Borges et Levene proposent en 1998[3] une modélisation des parcours sur un site sous forme de Grammaire Probabiliste Hypertextuelle (HPG : *Hypertext Probabilistic Grammar*) qui permet de fouiller des bases de parcours et d'identifier des séquences récurrentes suivies par les internautes. À la même époque, et poursuivant les mêmes objectifs, Srivastava *et alii*[4] présentent le logiciel *WebMiner*, qui applique des techniques de *data mining* à l'étude des usages du Web ; en complément, le système *WebSIFT* doit permettre d'identifier les motifs d'usages les plus intéressants *via* l'analyse du contenu et de la structure d'un site. Enfin,

---

[1] [Masand & Spiliopoulou 2000].

[2] [Kosala & Blockeel 2000].

[3] Voir [Borges & Levene 1998], ainsi que [Borges & Levene 1999] et [Borges & Levene 2000].

[4] [Cooley *et al.* 1997], [Cooley *et al.* 1999a] et [Mobasher *et al.* 2000].

Spiliopoulou, Faulstich et Winkler présentent en 1999 un outil, *Web Usage Miner*[1], capable d'agréger sous forme d'arbre les différentes navigations suivies au sein d'un site. Un langage proche du SQL, *MINT*, permet de fouiller, sous forme de requête, dans la base des parcours effectués sur un site et de connaître la probabilité qu'un utilisateur voie une page étant donné les autres pages vues avant ou après. De manière générale, ces outils et méthodes centrés-serveur tiennent souvent pour acquis que le contenu des pages est connu, et appliquent des méthodes d'analyse statistique sur des séries de symboles représentant les pages. Cela étant, ces approches sont intéressantes en ce qu'elles traitent réellement de l'aspect séquentiel et parfois même temporel des parcours, et permettent d'identifier des motifs de navigation pertinents. Ce qui leur manque avant tout, c'est une connaissance et un suivi des utilisateurs ; certains proposent des méthodes pour tenter de les deviner, comme [Murray & Durrell 1999] qui fait correspondre informations socio-démographiques et centres d'intérêt. Dans [Chevalier *et al.* 2003], les auteurs proposent d'identifier, au sein des visiteurs d'un site donné, des profils socio-démographiques distincts et des modes de navigation correspondants. Toutefois, le positionnement structurel de ces travaux du côté des sites les rendent inaptes à dépasser des corrélations locales entre certains types de sites et certaines variables relatives à l'utilisateur, et leur interdit d'embrasser la diversité des pratiques. Une sémantique des parcours ne saurait se satisfaire des résultats obtenus dans ce cadre en termes d'usages, et appelle une méthodologie centrée sur l'internaute qui prenne en compte sa dynamique d'usage dans l'analyse.

Les expériences menées en psychologie cognitive et tournées vers la modélisation du comportement des utilisateurs rentrent dans ce cadre, mais ne nous satisfont pas pour les raisons que nous avons déjà évoquées, et c'est bien plutôt vers la sociologie des usages que nous chercherons des éléments de description globaux et de contextualisation des parcours. Ainsi que le rappelle Jouët dans son *Retour critique sur la sociologie des usages*, dans tous les travaux de ce champ de recherche, « l'usage est analysé comme un construit social. […] La sociologie des usages, à l'opposé de la problématique de la traduction, n'étudie pas tant l'amont que l'aval, c'est-à-dire l'usage resitué dans l'action sociale. La construction de l'usage ne se réduit dès lors pas aux seules formes d'utilisation prescrites par la technique qui font certes partie de l'usage, mais s'étend aux multiples processus d'intermédiations qui se jouent pour lui donner sa qualité d'usage social. »[2] L'auteur distingue quatre problématiques principales, quoique souvent entremêlées au sein des études de cas, qui traversent le champ de la sociologie des usages :

- la généalogie des usages met en parallèle l'évolution des outils techniques et leur insertion dans les pratiques et les équipements existants ;
- la question de l'appropriation s'attache à décrire la construction des usages par les individus, les situations interactionnelles induites par l'objet

[1] [Spiliopoulou *et al.* 1999].

[2] [Jouët 2000], p. 499.

technique, et la dimension de construction de l'identité personnelle et sociale induite par les TIC ;
- le questionnement sur le lien social vise l'élaboration ou la modification des liens interpersonnels et des collectifs à l'aide des TIC ;
- enfin, la question des rapports sociaux prête attention au fait que les TIC s'insèrent dans des rapports sociaux et que, en tant qu'objets symboliques, ils constituent des enjeux de pouvoir.

Que retiendrons-nous de ces éléments pour l'analyse des parcours ? La majorité des études menées dans ce champ adoptent des méthodologies plus qualitatives que quantitatives : entretiens, observations, enregistrements vidéo, carnets de correspondants, etc. Comme le note Jouët, « si seule l'approche qualitative peut tenter de dégager la signification des actes de communication au niveau individuel et le sens social des usages auprès de groupes sociaux spécifiques, la démarche quantitative se révèle riche pour donner à l'usage une dimension plus macrosociale »[1]. En travaillant sur des données de trafic, on ne se trouve totalement ni dans l'une, ni dans l'autre des deux perspectives : la finesse de ce type de matériau autorise des analyses très précises sur les modes de navigation, bien qu'elle les désincarne, tandis que la technicité du matériau autorise des analyses globales et des croisements statistiques pour faire émerger des phénomènes récurrents à grande échelle.

Ne perdant pas de vue qu'elle s'intéresse principalement aux parcours « dans » l'écran et à l'activité de navigation, une sémantique des parcours cherchera ainsi à trouver des éléments d'explication des comportements en les rattachant à un utilisateur, mais dans une perspective toutefois endogène : les caractéristiques des parcours Web d'un individu sont sans doute corrélées statistiquement avec des variables socio-démographiques, mais c'est dans le contexte plus précis de la vie de l'utilisateur sur le Web, de ses parcours passés et de son corpus de sites et de pages que l'on peut faire surgir une signification de la navigation perçue comme activité située.

En nous centrant sur l'activité de navigation envisagée comme chaîne opératoire et réalisation d'un projet, nous nous heurtons, avec les données de trafic dont nous disposons, au problème de la reconstruction *a posteriori* d'une intentionnalité de l'utilisateur. Avec ce type de matériau, cette dimension – problématique en elle-même – échappe totalement à notre regard, et la question du projet qui sous-tend le parcours demeure indécidable. Pour répondre à ce type de questions, il parait bien plus approprié de mener des entretiens avec les utilisateurs, de les observer directement ou de leur présenter des parcours qu'ils ont faits en leur demandant de les commenter et d'en expliciter la logique. Figure énigmatique, l'utilisateur sera pour nous à la fois fuyant, car inaccessible, et omniprésent, car principe d'existence des parcours, de sorte que nous ne le considérerons pas tant sous l'angle de ce qu'il est (son âge, sa profession, etc.) que de ce qu'il fait (sa navigation).

C'est alors sur la notion de *territoires personnels* que peut s'appuyer l'appréhension du sens des parcours. Les données de trafic dont nous disposons, enregistrées sur une longue période, permettent de rapporter la visite d'un site au

---

[1] [Jouët 2000], p. 514.

corpus individuel de sessions et d'espaces Web de l'individu, et de la valoriser en regard des visites précédentes. Nous l'avons dit précédemment, une page peut être appréhendée de manière différente par un même individu dans deux contextes distincts ; corrélativement, la primauté du global sur le local nous invite à penser que l'appréhension d'une page ou d'un site est fortement déterminé par l'ensemble des visites précédentes du site ou de la page. Au fil du temps, de l'expérience à l'usage, la pratique individuelle délimite au sein de la Toile un territoire où se distinguent le routinier, l'habituel et l'exceptionnel, et où se dessinent des modes d'activité, des comportements et des temporalités distincts. Sur la base d'un corpus volumineux de traces d'activités en situation naturelle, l'analyse de la structure des ces territoires, de leur contenu et des formes de parcours qui y sont liés alimente, en quelque sorte, une éthologie de la navigation sur le Web.

Conclusion

Le Web n'est pas seulement un lieu de lecture mais également d'activité, et la navigation n'est pas tant un parcours de lecture qu'un régime d'action particulier pouvant supporter différents types d'activités : lecture, écriture, jeu, communication, etc. Le parcours sur le Web s'apparente alors à la rencontre dynamique entre un utilisateur et des contenus, à un moment et dans une chaîne opératoire donnés. Les enjeux d'une analyse des parcours sur le Web ne se déterminent alors plus en termes d'accès à l'information, mais amènent à examiner dans quel régime d'action se situe l'utilisateur à un moment donné en fonction de ses routines, ses motivations (ce qui vient de l'utilisateur, cause exogène) et de ce qui est disponible sur le Web (nature du contenu, cause endogène). La navigation Web s'apparente ainsi à un objet complexe dont l'analyse nécessite un appareillage méthodologique spécifique à même de prendre en compte les dynamiques qui s'y jouent ; le travail que nous présentons ici entend, sur la base de données de trafic centrées-utilisateur, explorer des pistes pour y parvenir.

L'ambition générale d'une sémantique des parcours étant une description la plus fine possible des parcours sur le Web, celle-ci affronte un double problème : elle doit parvenir à remplir la tâche qu'elle s'est fixée aux différents paliers de l'analyse, ce qui implique de parvenir à caractériser les pages tout autant que les panélistes avec la même acuité. Elle doit également être capable de décrire les interactions qui existent entre les différentes échelles et d'en tenir compte dans un système descriptif et analytique complet. L'appareil méthodologique que nous avons exposé montre assez clairement la place primordiale que doit tenir l'activité de navigation proprement dite, rattachée à un utilisateur donné, dans la sémantique des parcours. Quand bien même l'analyse de la production a un rôle à y jouer, c'est bien autour de l'internaute que se noue l'essentiel des éléments du parcours, et dans cette perspective, les données de trafic centrées-utilisateur dont nous disposons sont tout particulièrement adaptées à l'analyse des parcours.

Pour cela, nous ferons bien évidemment appel aux travaux déjà menés du côté tant de l'analyse de la page que de celle des stratégies et des modes de navigation ; nous tentons surtout de mobiliser ces recherches jusqu'ici disjointes autour de l'analyse des parcours. Nous avons recours pour y parvenir à une série de champs

disciplinaires, au premier rang desquels la linguistique informatique et la sémantique interprétative, associées aux outils de la statistique descriptive. Et, par-delà l'appel à tel ou tel outillage analytique susceptible d'éclairer un point précis de nos recherches, nous plaçons résolument nos travaux dans le cadre des sciences humaines et, corrélativement, d'une praxéologie.