Des machines pour chercher au hasard : moteurs de recherche et recherche d'information

Olivier Ertzscheid, Gabriel Gallezot

XIVe Congrès SFSIC, Béziers 2004
Questionner l'internationalisation : cultures, acteurs, organisations, machines

Des machines pour chercher au hasard : moteurs de recherche et recherche d'information

Ertzscheid Olivier. Université de Toulouse. O.Ertzscheid@voila.fr
Gallezot Gabriel. Université de Nice. gallezot@unice.fr

Quels sont les modèles et comportements qui sous-tendent la manière dont s'organise la recherche et l'accès aux informations (numériques) à une échelle qui est désormais celle de la planète ? Chacun d'entre nous est aujourd'hui confronté à de nouvelles machines à communiquer dans le cadre de sa recherche d'information (moteurs et annuaires de recherche). Ces nouvelles "bibliothèques" de l'internet reposent sur des logiques de classement des documents très éloignées de celles en vigueur dans nos bibliothèques classiques. Pour autant, ces nouvelles technologies intellectuelles permettent des découvertes informationnelles qui se transformeront, après un processus créatif, en connaissance. De nouvelles pratiques apparaissent, qui utilisent la *sérendipité* (hasard, fortuité) comme adjuvant explicite de la recherche. L'interaction constante entre ces nouvelles pratiques et ces nouvelles logiques d'organisation de l'information fait émerger et permet de caractériser de nouveaux contenus.

Introduction

Nous choisissons de questionner l'internationalisation en prenant la recherche et la diffusion d'information comme angle d'approche. La question nodale ici posée est celle des modèles qui sous-tendent la manière dont s'organisent les informations et les connaissances, à l'échelle de la planète, et sous des modalités essentiellement numériques, modalités chaque jour plus contraintes par des logiques marchandes en lutte avec un certain nombre de principes objectivables de la recherche et de l'accès à l'information.

La caractéristique principale de la toile mondiale n'est pas tant qu'elle a permis de rendre disponibles des millions d'informations, mais aussi et surtout qu'elle a amené des millions d'utilisateurs à faire de la recherche d'information une tâche quotidienne. Chacun d'entre nous, usager ou producteur de l'information, est aujourd'hui confronté à de nouvelles machines à communiquer et conduit à chercher de l'information, notamment avec les moteurs et annuaires de recherche. Ces nouvelles "bibliothèques" de l'Internet reposent sur des logiques de classement et d'organisation de l'information et des documents bien éloignées de celles en vigueur dans nos bibliothèques classiques, logiques pour lesquelles l'usager non-expert ne dispose d'aucune clé de lecture, "découvrant" plus qu'il ne "recherche" de l'information.

Nous parlons alors de "*sérendipité*" pour désigner "la découverte par chance ou par sagacité de résultats que l'on ne cherchait pas" (Horace Walpole, 1754).

Nous replaçant dans le contexte des technologies intellectuelles (1), nous montrons comment, avec l'internationalisation des outils de recherche, des logiques marchandes prennent le pas sur des logiques classificatoires (2), mettant en exergue le phénomène de sérendipité. Les modèles de l'IR (*Information Retrieval*) constituent dès lors une aide précieuse pour mieux comprendre l'ampleur de ce phénomène et permettre d'atténuer certains de ses biais (3). Nous montrons enfin comment l'interaction constante entre ces nouvelles pratiques et ces nouvelles logiques d'organisation de l'information fait émerger et permet de caractériser de nouveaux contenus.

1 - Découvertes informationnelles et technologies intellectuelles.

« Repérer/collecter/traiter/diffuser » l'information. Les actions de cet ensemble sont réalisées par un binôme indissociable d'outils et de méthodes : les technologies intellectuelles [Fayet-Scribe, 00]. Celles-ci permettent des découvertes informationnelles qui se transformeront, après un processus créatif, en connaissance. Ainsi nous distinguons la recherche d'information de l'*épistèmè*, mais soulignons leur appartenance au même processus.
Le processus de création conventionnel fonctionne sur le principe de divergence/convergence où la reconnaissance d'un problème est introduite par une divergence pour converger vers une nouvelle solution. Le processus de création par *serendipité* est le contraire : bien que la solution à un problème soit attendue, il y a divergence dans les parcours qui conduisent à un problème différent ou, plus fréquemment, à la solution d'un problème dont nous n'avions aucune connaissance [Figueirado, 01]
Comment dès lors garder une emprise sur ce phénomène ? Si l'on considère que les découvertes n'arrivent jamais par chance, il faut donc insister sur le rôle de la préparation intellectuelle et/ou l'intensité de l'observation et de la recherche [Boursier, 92]. On peut aussi penser que l'IR prenne en compte les phénomènes de sérendipité en complément des requêtes (*querying*) et de la navigation (*browsing*) pour stimuler la curiosité et encourager l'exploration [Toms, 01].
L'apport de l'internationalisation des outils de recherche doit, pour être compris, être replacé dans le cadre historique de l'évolution des techniques d'indexation, depuis le "Traité de Documentation" d'Otlet (1934) jusqu'au développement des premiers outils "hypertextuels" (Nelson, Engelbart) en passant par le changement de paradigme énoncé par [Bush, 45] s'efforçant de mécaniser le fonctionnement caractéristique de l'esprit humain pour que l'accès aux connaissances et leur indexation se fasse de manière "associative". [Ertzscheid, Gallezot, 03].

Les nouvelles machines à chercher de l'information s'inscrivent dans cette filiation, tout en lui assignant de nouveaux enjeux.

2. Internationalisation des outils de recherche : quand les logiques marchandes prennent le pas sur les logiques classificatoires.

2.1. Approches classificatoires

Les moteurs de recherche, dans l'utilisation qu'ils font des liens comme principes de classification, ne sont pas de simples interfaces de recherche prenant exclusivement en compte les mots (clés ou non) et les occurrences de ces mots. À l'inverse, faire le choix des liens comme principe de classement, de tri et d'organisation de l'information[1], c'est revendiquer clairement le choix de l'immatériel ou à tout le moins le choix de l'information comme mesure « *d'une différence qui produit une autre différence* » [Bateson, 77].
Quand nous consultons une page de résultat de *Google* ou de tout autre moteur utilisant un algorithme semblable, nous ne disposons pas simplement du résultat d'un croisement combinatoire binaire entre des pages répondant à la requête et d'autres n'y répondant pas ou moins (*matching*). Nous disposons d'une vue sur le monde (*watching*) dont la neutralité est clairement absente. Derrière la liste de ces résultats se donnent à lire des principes de

[1] Comme ce fut le cas pour la révolution entraînée par l'algorithme PageRank du moteur *Google* (www.google.com) qui considéra que la pertinence d'une page était liée en priorité au nombre de pages la référençant (liens entrants) et non plus exclusivement à des mesures d'occurrence linguistique. Ce critère (inspiré de Garfield et de la bibliométrie) est actuellement pris en compte par la plupart des outils de recherche.

classification du savoir et d'autres encore plus implicites d'organisation des connaissances. C'est ce rapport particulier entre la (re-)quête d'un individu et la (re-)présentation d'une connaissance qui était présente dans les bibliothèques de la Haute-Egypte, pour en être évacuée avec l'arrivée des principes de classement alphabétiques.

Une nouvelle logique se donne à lire. Moins « subjective » que les principes classificatoires retenus par une élite minoritaire (clergé, etc.) elle n'en est pas moins sujette à caution. Les premières étaient douteuses mais lisibles, celles-ci le sont tout autant parce qu'illisibles[2], c'est-à-dire invisibles : l'affichage lisible d'une liste de résultats, est le résultat de l'itération de principes non plus seulement implicites (comme les plans de classement ou les langages documentaires utilisés dans les bibliothèques) mais invisibles et surtout dynamiques, le classement de la liste répondant à la requête étant susceptible d'évoluer en interaction avec le nombre et le type de requêtes ainsi qu'en interaction avec le renforcement (ou l'effacement) des liens pointant vers les pages présentées dans la page de résultat.
Des pratiques émergent alors qu'il serait intéressant d'analyser du point de vue de la sociologie des usages. Nombre d'usagers ayant connaissance des grands principes de l'algorithmie de Google (pages pivot et pages d'autorités, pertinence d'une page liée au nombre de liens pointant vers elle), se servent de cette connaissance pour fausser ces résultats. Le "*Google Bombing*" désigne ainsi le fait de créer une page (le plus souvent un weblog) dans laquelle on va associer le nom d'une personnalité (politique le plus souvent) à une expression visant à la discréditer. Il suffit alors de faire référencer cette page par des sites disposant d'un bon PageRank (indice de classement de *Google*) pour qu'en quelques jours l'entrée de l'expression associée à la personnalité en question soit considérée comme allant de fait par ce moteur de recherche et renvoie sur des pages "officielles" de la personne[3].
Au-delà de la logique subversive qui sous-tend ces pratiques à l'échelle individuelle, il faut s'interroger sur la position de leader de *Google*, le constituant de fait comme une formidable machine à façonner l'opinion internationale.

2.2. Logiques marchandes

L'internationalisation des outils de recherche pose un certain nombre de questions économiques, questions d'actualité à l'heure où le premier de ces outils (*Google*) annonce depuis maintenant près d'un an sa prochaine introduction en bourse. De nouvelles logiques - marchandes - viennent interférer avec les logiques classificatoires déjà biaisées décrites plus haut. Ainsi devant l'ampleur de la toile mondiale et la difficulté du recensement de l'ensemble des informations disponibles, de plus en plus d'acteurs majeurs de la recherche d'information sur le web fusionnent et se regroupent, ce qui donne lieu à un échange ou à une vente de tout ou partie de leurs bases d'index et de leurs bases de données.
Ainsi, une recherche dans la partie annuaire de *Google* donnera exactement les mêmes résultats qu'une recherche similaire sur l'*Open Directory*, ce dernier fournissant à *Google* sa base d'annuaire[4]. Ces pratiques ne constitueraient pas en elles-mêmes un danger pour l'objectivation de la recherche d'information sur le web si les usagers (novices) en avaient connaissance ou pouvaient le soupçonner, ce qui est loin d'être le cas. D'autant que ces logiques marchandes, qui s'expliquent par des nécessités autant techniques que pragmatiques sont à leur tour biaisées par des aspects de politique marketing également déroutantes pour le chercheur.

---

[2] Pour les utilisateurs non spécialistes.

[3] Ainsi la requête "*Miserable failure*" renvoyant sur le site officiel de Georges W. Bush ou l'expression "*Go to Hell*" renvoyant la page d'accueil de Microsoft. Dernière victime en date, le sénateur Jean Charest dont le nom se retrouve associé à l'expression "mouton insignifiant".

[4] Source : `http://docs.abondance.com/portails.html`

Mentionnons également au premier rang de ces logiques marchandes, l'arrivée de l'indexation payante qui s'affirme comme le seul modèle économiquement viable pour les différents outils de recherche. Il va sans dire que cette logique et les pratiques qui lui sont associées (achat de mots-clés auprès de certains moteurs, garanties de "positionnement" dans la liste affichée de leurs résultats, etc.) constitue là encore un biais évident pour l'objectivation du déroulement d'une procédure de recherche d'information, même si - le monde francophone ayant exprimé sa défiance vis à vis de ce système - les résultats relevant de "liens sponsorisés" ou de "mots-clés achetés" apparaissent de manière différente et plus ou moins repérable sur les sites francophones de ces outils, et qu'il est donc en théorie toujours possible de faire fi de ces résultats pour aller consulter ceux ne relevant pas de cette approche.

3 - Les modèles de l'IR et la serendipité.

L'apport des modèles de l'IR permet d'établir une sériation plus robuste de la sérendipité (emblématisés par exemple, par le bouton "*I'm feeling Lucky*" de Google).

Le tableau suivant se propose d'établir une grille d'analyse en préalable à cette sériation.

Etat Initial	Processus	Modèles
Je sais ce que je cherche	Querying	Computationnel
Je ne sais pas ce que je cherche	Searching Sérendipité structurelle	Utilisateur
Je sais que je ne sais pas ce que je cherche	Learning Sérendipité associative	Environnementaliste

Le premier cas repose sur l'idée que dans la majorité des démarches de recherche d'information, l'utilisateur sait déjà (partiellement) ce qu'il cherche. Il lui reste alors à mettre en place une série de requêtes (querying) correspondant au modèle computationnel classique autorisé par les systèmes documentaires (booléens, langages documentaires, etc.). L'utilisateur est dans une logique de consultation et cherche à savoir ce que peut lui apporter comme résultats (matching) le système d'information qu'il est en train d'utiliser (browsing). Cet utilisateur met en place un raisonnement de type hypothético-déductif. La sérendipité est alors quasi-nulle ou ne relève en tout cas d'aucune démarche volontariste ou consciente.
Le second cas correspond à l'objectif de l'IR selon [Belkin, 00], à savoir : "*Helping people find what they don't know.*" Le processus alors appelé est de type exploratoire (searching). L'utilisateur va, à partir de ce qu'il sait, raisonner par inférence et abduction en fonction de son but ou de son "profil". La sérendipité qui se met ici en place est de type structurelle : si je suis à la recherche d'une thèse en sciences de l'information sur un site mettant à disposition ce genre d'ouvrages, je peux trouver une autre thèse qui recoupera mes préoccupations de recherche (en mathématiques par exemple) mais il s'agira encore d'une thèse (identité de forme) et non pas d'un article de journal.
Le dernier cas est celui qui peut le plus bénéficier du phénomène de sérendipité. L'utilisateur ayant formalisé et explicité qu'il "ne sait pas ce qu'il cherche" se met alors consciemment en situation d'adopter le comportement le plus simple, le plus intuitif et associatif possible, et ce quelle que soit la complexité des systèmes qu'il consultera. Nous sommes alors dans le cadre d'un authentique processus "d'apprentissage périphérique" tel que défini par [Lave & Wenger, 91].

Dans ce processus, l'information qui sera prioritairement "captée" par l'utilisateur et servira de base aux associations qu'il échafaudera pour aller au bout de sa quête, cette information donc, relève en premier lieu des propriétés invariantes de l'environnement : de la même manière que je peux utiliser un stylo comme un stylo si je veux écrire, ou comme un marteau si je veux planter un clou, je peux utiliser la liste des 10 premiers résultats d'un moteur de recherche de manière systématique (et aller voir chacune des pages vers lesquelles ils pointent) ou de manière associative pour repérer aléatoirement (dans le texte de description fourni pour chaque page par exemple) de nouveaux mots-clés, de nouveaux noms de personnes qui vont m'engager sur une autre piste de recherche ou vont en l'état constituer une réponse/solution à ma question/problème. La sérendipité est ici de type associative : c'est ce type de processus qui est systématisé par la plupart des outils de recherche ayant fait le choix de représentations graphiques (Kartoo, Mapstan, etc …) pour optimiser l'instrumentalisation de ce type de sérendipité [Ertzscheid, 03].

La sérendipité peut-être passagère ou devenir un mode privilégié d'accès à l'information dans le cadre d'un processus de recherche ou de l'une de ses itérations. Elle atteste qu'il n'est pas nécessairement plus facile de trouver de l'information dans un système ordonné, structuré et formaté que, comme cela semble être le cas pour le web, dans un système d'information caractérisé par une forte entropie et ne disposant en tout cas d'aucun niveau de contrôle unique[5]. Pour autant, il semble essentiel de se donner les moyens de penser la diffusion d'information et la structuration de contenus numériques en des termes qui prendront en compte, à la source, les sauts conceptuels et autres ruptures d'arborescence dont se nourrit la sérendipité. Les principales voies de recherche œuvrant actuellement dans ce domaine sont celles du web sémantique, des hypermédias d'apprentissage et bien entendu des approches théoriques de la recherche d'information (IR).
Voici pour conclure cette partie un rappel des différents biais intervenant dans le cadre d'un processus de recherche mettant en relation un (des) usager(s), un (des) outil(s) de recherche et un (des) document(s) et favorisant les phénomènes de sérendipité.

Entre l'usager et l'outil de recherche	Logiques sociétales (position de leader de Google)
	Logiques marketing (interfaces modifiées pour favoriser l'accès à telle ou telle base de résultats)
	Logiques représentationnelles (cartographies, clusters dynamiques)
	Logiques commerciales (bases de données partagées)
Entre l'outil de recherche et le document	Logiques algorithmiques (algorithmes différents, variabilité de la notion de "pertinence" selon l'algorithme utilisé
	Logiques commerciales (indexation payante)
	Logiques informationnelles (seules 25% des pages Web datent de plus d'un an : l'exceptionnel taux de renouvellement de la toile mondiale accentue la part de la sérendipité. (Brooks 03))
	Logique de récurrence (taux de recouvrement (taille et nombre de documents recensés) de la base d'index du moteur
	Logique temporelle (possibilité de documents accessibles uniquement dans le "cache" de certains moteurs alors qu'ils n'existent plus physiquement sur la toile)

[5] Un protocole expérimental est actuellement en cours à l'Université de Toulouse 1, auprès d'étudiants en première année de DEUG de droit pour évaluer les usages novices en recherche d'information. Le phénomène de sérendipité peut ainsi être "expérimentalement" observé.

4 - Pratiques cognitives, sérendipité, modèle heuristique.

Appréhender dans leur complexité les phénomènes, les objets de recherches semble une tâche impossible. Pour sortir de la « boucle récursive » ou graphe complexe le chercheur doit entreprendre différentes stratégies et opter pour les choix qui s'offrent à lui. La troisième voie est introduite par la sérendipité. Elle est une approche socio-cognitive de la recherche d'information et impose l'abduction comme heuristique.
Pour tenter d'expliquer l'influence de la sérendipité en matière de construction de connaissances, il semble que deux dimensions soient à retenir : l'importance du contexte et le transfert de compétences dans une situation nouvelle (métaphore). Le contexte est notamment composé de la connaissance et des technologies intellectuelles qui la manipulent. Le transfert de compétences dans une situation nouvelle est lié à l'appropriation d'une culture technique et informationnelle, de savoirs par les chercheurs et leur capacité à transposer, transfigurer des phénomènes, des problèmes.
La sérendipité se réalise alors par l'appropriation individuelle du contexte socio-technique, une *lecture* spécifique, créative du réservoir cognitif et instrumental. Les chercheurs les plus en phase avec le contexte socio-technique favorisent ainsi leur perspicacité et la mise en œuvre d'artefacts informationnels qui permettent de faire apparaître des éléments stochastiques apte à changer « la vision » du monde.

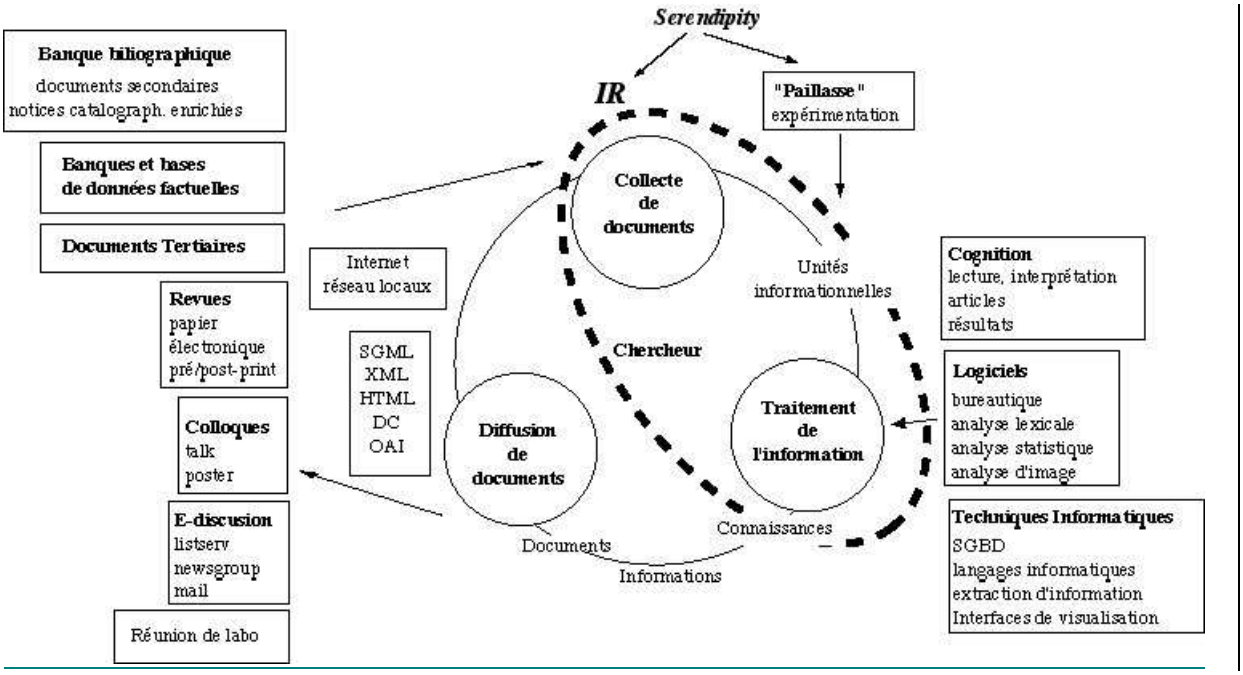

Fig. : le cycle de l'IST, contexte socio-technique, IR et sérendipité

La partie gauche du schéma représente les sources et les entrepôts qui alimentent ou sont alimentés par le processus informationnel scientifique représenté par la collecte, le traitement et la diffusion de l'information scientifique et technique essentiellement médiatisés par des réseaux et des formats de données standardisés très utilisés. Les données issues de la paillasse sont volontairement distinguées de cette dernière partie pour signifier leur statut primaire et endogène aux laboratoires. La partie droite du schéma présente les dispositifs de traitement de l'information en distinguant ce qui relève du seul acte intellectuel du chercheur, puis des logiciels et des techniques qui l'aident dans cette action. La partie IR (en pointillé) indique les étapes du cycle relevant de la recherche d'information et les champs de la sérendipité sont désignés par les flèches… pour montrer au final un modèle heuristique d'analyse de

l'ensemble des actants qui interviennent dans le processus de construction des connaissances. Ces actants sont mondialisés, la manière dont s'organise l'information et la connaissance aussi [Gallezot, 02a].

Si l'avenir et le rôle des moteurs de recherche sur la toile internationale ne semblent pas pouvoir être remis en question, la part que ceux-ci seront disposés à faire à un traitement "objectivable" de l'information à laquelle ils permettent d'accéder paraît en revanche être ramenée vers des portions de plus en plus congrues. Or plus que jamais dans ce contexte international, le document numérique peut-être finement décrit par la manière dont il est repéré et accessible sur les réseaux. Dans le même temps, le même document, vu cette fois comme "résultat de recherche" bénéficie, du point de vue de l'IR, d'une perception radicalement changée, comme cet article s'est efforcé de le montrer.
Ajoutons à cela que faute d'une acculturation aux outils de recherche, la plupart des usagers continueront de percevoir l'information affichée par ces outils comme objectivée sans jamais se poser la question des modalités d'objectivation retenues ou sans pouvoir disposer de méthodologies d'objectivation qui ne soient directement dépendantes de l'un de ces outils et donc non applicable aux autres. Dès lors, la recherche d'outils capables de maîtriser l'information dans cet espace réticulaire constitue pour beaucoup un enjeu majeur. S'il est évident qu'il faille tendre vers une appropriation informationnelle exhaustive pour édifier l'*épistèmè*, la tâche est incommensurable. Que reste-t-il au chercheur devant cette entropie informationnelle ? Se servir des outils *ad hoc* pour repérer au mieux l'information pertinente, borner son référentiel documentaire, expérimenter, observer, évaluer et produire ses résultats à l'aide de méthodologies éprouvées, de protocoles heuristiques… passer des achoppements aux paradigmes scientifiques. Il existe un « raccourci » : la serendipité. Elle s'offre et se révèle lors de découvertes informationnelles [Gallezot, 02b].

Bibliographie.

Bateson G., *Vers une écologie de l'esprit*, T. 1. Paris, Seuil, 1977.

Belkin N., *Helping People Find What They Don't Know*, in Communications of the ACM, August 2000, Vol. 43, No. 8.

Boursier & Van Andel, « Serendipity : expect also the unexpected », *Creativity and innovation management*, vol 3, p.20-32, 1992.

Brooks T.A., "Websearch : how the web has changed information retrieval", *Information Research*, 8(3), paper n° 154. [en ligne] http://informationR.net/8-3/paper154.html.

Bush V., « *As We May Think.* », pp. 101-108, in The Atlantic Monthly, vol.1, n°176, Juillet 1945. [en ligne] `http://www.isg.sfu.ca/~duchier/misc/vbush`, consulté le 07/02/1998.

Engelbart D.C., « Augmenting Human Intellect : a Conceptual Framework », *Summary Report*, AFOSR-3233, Stanford Research Institute (SRI), Contract AF49(638)-1024, SRI Project N° 3578, Octobre 1962. [en ligne] `http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html`, consulté le 03/03/2002.

Ertzscheid O., "Syndrome d'Elpenor et sérendipité : deux nouveaux paramètres pour l'analyse de la navigation hypermédia." in *Actes du colloque H2PTM'03*. Editions Hermès, septembre 2003

Ertzscheid O., Gallezot G., "Chercher faux et trouver juste : sérendipité et recherche d'information." 1ère conférence internationale francophone en Sciences de l'Information et de la Communication. Bucarest. Juillet 2003, http://archivesic.ccsd.cnrs.fr/sic_00000689.html

Fayet-Scribe S. *Histoire de la documentation en France : Culture science et technologie de l'information*, CNRS éditions, 2000.

Figueirado A. Dias de, Campos J., "The Serendipity Equations". *Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning*, ICCBR 2001, Technical Note AIC-01-003. Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence [en ligne] `max.ipv.pt/pub/AdeFigueiredo01.pdf`

Gallezot G., « La recherche in silico » In : Chartron G. (dir.*) Les chercheurs et la documentation électronique : nouveaux services, nouveaux usages*, Edition du cercle de la Librairie, Coll. Bibliothèque, juillet 2002.

Gallezot G., "Exploration informationnelle et construction des connaissances en génomique", *Les Cahiers du numérique,* Hermès, vol.3, n°3, novembre 2002.

Lave G., Wenger E., Situated Learning : *Legitimate Peripheral Participation*. New-York, Cambridge University Press, 1991.

Toms Elaine G. « Serendipitous Information Retrieval », `http://www.ercim.org/publication/ws-proceedings/DelNoe01/3_Toms.pdf`