Chapitre 1

Recherche d'information, Web et interopérabilité

1.1. Recherche d'information : de la théorie aux pratiques

La généralisation de l'accès au réseau Internet et à ses services les plus populaires tels que la messagerie, le Web et plus récemment les réseaux sociaux numériques, a conduit à une banalisation des pratiques de recherche d'information (RI) [GRI 11, DIN 14, DIN 07, CIA 05, ASS 02], réservées jusqu'à un passé récent, à des spécialistes de l'information (journalistes, documentalistes, veilleurs, archivistes, bibliothécaires, etc.) [CAT 01, DUF 01, LEF 00]. Introduite auprès du grand public par les moteurs de recherches généralistes libres d'accès, la RI a été réduite en quelques années à une délégation aveugle des internautes à la toute-puissance de ces systèmes d'indexation et de recherche automatisés aux mécanismes basiques[1], capables de répertorier avec précision et rapidité une grande partie de la production documentaire visible du Web [LEW 08, CHI 07, RIE 06, LEL 99].

1. « In Google, the web crawling is done by several distributed crawlers. There is a URLserver that sends lists of URLs to be fetched to the crawlers […] The storeserver then compresses and stores the web pages into a repository. Every web page has an associated ID number […] The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. » Page L., Brin S., *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, http://infolab.stanford.edu/~backrub/google.html.

Une enquête menée en 2008[2] auprès de 2 218 doctorants de Bretagne sur leurs besoins de formation à la maîtrise de l'information scientifique a révélé que les ressources utilisées à 96 % étaient les moteurs de recherches[3] (73 % les utilisent très souvent et 23 % régulièrement), les portails spécialisés arrivant en seconde position avec 53 % d'utilisation.

La simplicité algorithmique de ces outils de recherche, servie par des architectures informatiques aisément modulables (*clustering, cloud computing*) a permis de s'adapter à la versatilité des formats de documents et d'absorber l'accroissement exponentiel des requêtes[4]. Cette performance technologique a longtemps contribué à ce que les moteurs de recherche soient perçus comme le nec plus ultra en matière de systèmes de recherche. Les choses, en matière de RI simplifiée, en seraient probablement restée là, si le tournant social (Web 2.0) qui s'est massivement emparé du Web n'avait pas troublé la spectaculaire progression technologique de l'accessibilité des documents, en réintroduisant des variations dans les pratiques informationnelles et numériques.

Si la démarche conceptuelle de la RI [DEN 03, MAN 02, LEF 00] repose sur un modèle systématique et méthodique aisé à appréhender (figure 1.1), elle prendra en pratique de multiples formes en fonction des ressources primaires ou secondaires utilisées et surtout des moyens techniques sollicités (langages documentaires, langages d'interrogation, outils techniques, etc.) [TAS 14, ZOU 13, PIR 10, REP 11]. En peu de temps, la multiplicité des vecteurs technologiques en matière de recherche documentaire sur des informations désormais nativement numériques, a engendré une complexité inédite qui a exigé la mobilisation de stratégies de recherches s'appuyant sur une connaissance précise des fonctionnements techniques des systèmes de recherche d'informations quels qu'ils soient : « D'ores et déjà, les dispositifs technologiques proposent des solutions de recherche que le commun des mortels ignore. Or le constat est étonnant, c'est souvent à la convergence de ces différentes solutions qu'il faudra faire appel pour s'informer qualitativement » [MOE 98, p. 67].

2. Enquête menée par l'URFIST de Rennes et le Service commun de documentation de l'université de Bretagne-Ouest.

3. « Quelles ressources utilisez-vous pour vos recherches d'information ? » Les réponses possibles étaient : documentation de laboratoire, catalogue de bibliothèque, SUDOC, bases de données, moteurs de recherches, portails spécialisés, blogs, etc.

4. Selon www.internetlivestats.com, 40 000 requêtes sont adressées en moyenne par seconde à Google. En 2012, ce sont plus de 1 200 milliards de requêtes qui ont été transmises au leader mondial de la recherche sur le Web (en 1998, date de la création de Google, ce sont 10 000 requêtes par jour qui étaient traitées).

En démocratisant la production de contenus perpétuellement évolutifs (articles, billets, commentaires) enrichis de données multimédias [PAP 03, PAP 99], la blogosphère[5] a ébranlé l'ordre technologique établi par l'ingénierie des moteurs de recherche. Ces systèmes de publication ont facilité la création de contenus sur des logiques essentiellement rédactionnelles [ANG 11, DES 09, SOU 03] et l'ont libérée des indispensables habiletés informatiques qui en restreignaient l'accès à des experts et aux agences de communication spécialisées rompus aux techniques de publication/ écriture en ligne [PAP 14, AMG 08, TAR 07].

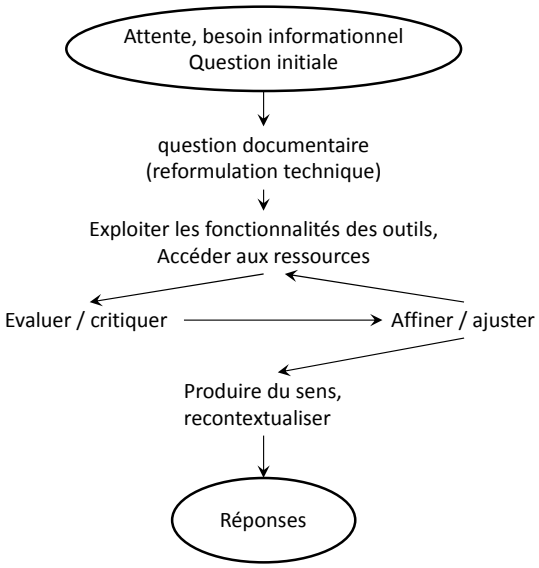

Figure 1.1. *Boucle de recherche d'informations (source : B. Pochet, « Savoir rechercher et interroger », Licence Creative Commons, 2008)*

5. Avant sa dissolution et sa réorientation vers des activités de conseil, la plate-forme Technorati recensait plus de 100 millions de blog en 2008. Grâce aux données agrégées sur sa plate-forme de publication de blogs, Wordpress.org estime que plus de 50 millions de nouveaux billets (ainsi que des pages virtuelles) et plus de 50 millions de commentaires sont publiés chaque mois. Sur le site Skyrock.com, fondé en 2002, le nombre de Skyblogs s'élève aujourd'hui à 26 millions où 650 millions d'articles publiés sont associés à plus de 4 milliards de commentaires.

COMMENTAIRES SUR LA FIGURE 1.1. Le point d'entrée « la question initiale », sous-entend une approche cognitive de l'interrogation avec ce qu'elle revêt de personnel et d'implicite. La « question documentaire » confronte la question initiale à une première explicitation (sans toutefois la spécialiser pour un système technique particulier) en mots, termes ou expressions à utiliser (forme canonique ou non), termes exclus, termes à associer obligatoirement ou facultativement (opérateurs booléens). Elle permet surtout à l'utilisateur de circonscrire le vocabulaire qu'il a à sa disposition pour décrire de différentes façons ses attentes informationnelles. C'est une étape importante durant laquelle l'utilisateur établira une forme de stratégie lexicale dans l'identification de termes génériques/spécifiques, usuels ou de spécialité, le choix de synonymes ou d'antonymes, etc.

1.2. Sources d'information, ressources documentaires

Dans les années 2000, l'engouement mondial pour les réseaux sociaux [BOU 10, REB 07], les approches participatives et collaboratives ont bousculé encore davantage la RI orchestrée par les grands moteurs du Web : « On sait par exemple que Facebook dispose aujourd'hui de plus d'un million de développeurs (des tiers) répartis dans plus de 180 pays dans le monde. Aujourd'hui, Facebook compte plus de 550 000 applications tierces développées, et plus de 70 % des utilisateurs interagissent avec ces applications » [TCH 11, p. 60].

Le fonctionnement de la blogosphère et des réseaux sociaux numériques se caractérise par un mode permanent d'instabilité des contenus et des données paratextuelles[6]. Un billet ou une page virtuelle dans un blog ou un réseau social n'atteint jamais une forme définitive. L'objet peut évoluer à tout instant au fil de la mise à jour (insertion, modification, effacement) du contenu réalisé par un ou plusieurs auteurs en fonction des droits d'accès attribués (figure 1.2). L'instabilité naît également des interactions sociales qui produisent des données d'accompagnement de la ressource primaire et qui l'enrichissent alors comme des facettes dès que l'information est disponible en ligne. Les commentaires, les appréciations, les votes (*rating*), l'identité, le statut, la qualité des lecteurs, la date, etc., sont autant de données sur lesquelles des recherches ultérieures peuvent être conduites.

6. Comme le sont les notes de bas de page, les citations, ou les références bibliographiques associées à un paragraphe ou une phrase.