Les enjeux du Web invisible académique: Spécificités et difficultés

Bentenbi CHAIB DRAA TANI
Université d'Oran 1 Ahmed Benbella.

1. Introduction

La partie du Web que l'on nomme d'ordinaire « Web invisible » est pareille à la partie immergée d'un iceberg. Bien que cette première appellation soit la plus couramment admise, on parle d'ailleurs aussi de « Web profond ». En réalité, l'expression désigne dans une perspective générale l'ensemble du contenu Web non indexé par les moteurs de recherche traditionnels tels que Google.

Il s'agira donc plus précisément, si l'on se réfère à la définition donnée par Sherman et Price[1], de pages de texte, fichiers ou autres informations de référence dotés d'une qualité tangible, accessibles à partir du World Wide Web, et que, souvent, les moteurs de recherche usuels s'avèrent incapables ou choisissent d'eux-mêmes – pour des raisons d'ordre technique (et notamment liées à restriction de l'espace de stockage) – de ne pas indexer, c'est-à-dire de ne pas référencer au sein de leurs pages d'index.

Pour affiner la teneur de cette définition du Web invisible, les analyses du chercheur Bergman[2] sont éloquentes. Ce dernier précise en effet que certains moteurs de recherche sont inaptes à saisir ou appréhender certains types d'information: dans les faits, ils sont dans l'incapacité de « voir » ou de collecter, à partir du Web profond, des pages qui n'existeront concrètement qu'une fois qu'elles auront été créées de façon dynamique, c'est-à-dire en tant que résultat d'une recherche bien spécifique. Autant dire qu'il convient de distinguer, au sein même du Web invisible, les données plus brutes issues de bases de données relationnelles d'un contenu généré dynamiquement.

En définitive, si par leur nature même certains types de pages (ex : *disconnected pages* ou pages accessibles en "mode hors connexion") présentent les germes d'une mauvaise indexation (elles peuvent, notamment, être exemptes des liens nécessaires à l'indexation par les moteurs de recherche – comme sur le Web visible, mais dans une bien moindre mesure – la vraie problématique est, elle, d'ordre proprement technologique, dans la mesure où il existe un réel manque d'informations quant à l'existence de « ce contenu invisible ». En

d’autres mots, il est bien évident que si les moteurs de recherche d’aujourd’hui avaient à leur disposition les moyens de trouver ces pages "laissées pour compte", il leur serait beaucoup moins ardu d’indexer ces dernières.

Il n’en demeure pas moins que si le PDF est désormais quasi reconnu de tous, nombre d’applications telles que *Flash* restent encore illisibles de certains instruments de requête. Ce qui suggère une interrogation d’importance : y-a-t-il un intérêt véritable à indexer le contenu du Web invisible ? Et qu'en est-il plus précisément s'agissant de la part du Web invisible que l'on nomme Web Invisible Académique ou W.I.A?

2. Pour une définition constructive du Web Invisible Académique

C’est donc dans un cadre fondamentalement dichotomique que s’inscrit l’existence du Web Invisible Académique. Or, ce dernier désigne la part du Web profond dédiée plus spécifiquement à la recherche universitaire; elle met donc en jeu l’ensemble des fonds documentaires dits « académiques ». Dans ce contexte, le contenu visé a plus précisément trait aux données fournies par les bases de données ; et dans la mesure où ce type de contenu est souvent au format PDF, c’est un ensemble de données techniquement lisible des moteurs de recherche à visée généraliste, et donc un véritable pivot documentaire.

Néanmoins, tous les obstacles à l'indexation de contenu invisible à partir du Web ne sont pas purement techniques, et dans le cas du W. I. A., la distinction proprement liée au format – opérée par Sherman et Price [3]– est essentielle. Car c'est plus particulièrement s'agissant de recherche documentaire académique qu'édifier une ligne de partage entre "contenu au format propriétaire" (*proprietary content*) et "contenu libre de droits" (*open access content*) prend tout son sens.

En effet, le Web invisible orienté requêtes universitaires est dans une grande proportion composé de fonds tirés des bases de données produites par les éditeurs: il s'agit donc de contenu documentaire au format propriétaire. Car le W. I. A. est avant tout un espace de recherche universitaire. Dans ce cadre, il présente surtout des documents textes d'extensions *.pdf, .doc* et *.ppt*, auxquels l'on peut accéder à partir des bibliothèques ou moteurs de recherche spécifiquement académiques. C'est ici la définition du Web Invisible Académique selon D. LEWAMDOWSKI et P. MAYR[4] qui s'esquisse. Pour ces chercheurs allemands en Sciences de l'information, le W. I. A. désigne l'ensemble des contenus disponibles à partir des bases de données et bibliothèques de documents virtuelles à caractère académique, et que l'on peut difficilement atteindre par le biais des

moteurs de recherche généralistes. C'est la définition du Web Invisible Académique que l'on retiendra.

3. Les enjeux du Web Invisible Académique : aspects techniques et applications

Vers une meilleure évaluation des dimensions du W. I. A.

C'est au chercheur BERGMAN[5] que l'on doit d'avoir initié les recherches sur la taille du Web invisible. Après des investigations fructueuses, lesquelles rencontrent l'approbation de nombre de ses collègues, le chercheur met au jour et publie un certain nombre de découvertes d'un intérêt notable eu égard aux tentatives d'évaluation dimensionnelle de cette "portion cachée" du Web.

Tout d'abord, il nous apprend que le Web invisible est près de 550 fois plus étendu que le Web visible, et qu'il présente un total de 550 milliards de documents environ. La plupart des évaluations ultérieures prendront pour point de départ les calculs de BERGMAN ; c'est ainsi qu'à l'instar de LYMAN et al[6]. en 2003, les scientifiques seront quasi unanimes à adopter l'échelle dimensionnelle établie par ce dernier, laquelle fonde la taille du Web invisible sur le ratio **Taille Web visible/ Taille Web invisible**, soit 1:500.

Dans un second temps, le chercheur s'emploie à la réalisation de calculs plus poussés dans le but de venir affiner les évaluations qu'il a entérinées. Il s'appuie sur un ensemble de bases de données (ce sera le *Top 60*), un échantillon représentatif des plus grandes bases du Web profond. S'appuyant alors sur ce matériel de données, il fonde son raisonnement sur la capacité moyenne de chacune des bases de données rencontrées sur le Web invisible – qu'il établit à 5,43 millions de documents – et parvient à évaluer la capacité totale du Web invisible à un ensemble de données s'élevant à 543 milliards de documents. Il convient néanmoins de garder à l'esprit que la taille du Web visible au moment des recherches évoquées (2001) équivaut approximativement à un contenu plus ou moins égal à 1 milliard de documents (cf. : LAWRENCE et GILES, 1999)[7]. Se basant sur ces présomptions, BERGMAN vient à conclure que le Web invisible est 550 fois plus étendu que le Web visible ou Web de surface.

Or, qu'advient-il plus précisément, face aux conclusions tirées, de la part proprement allouée au Web Invisible Académique ou W. I. A. ? Ce qui permettra d'ébaucher quelque embryon de réponse, ce sont les chiffres de quelques chercheurs, qui bientôt se manifestent et qui viennent remettre en cause les dernières estimations… Parmi ces derniers, on trouve SHERMAN (2001)[8] ou STOCK (2003)[9] ; mais ces deux scientifiques ne font qu'énoncer des hypothèses, sans apporter aucune

formule de calcul innovante, ni même d'explication tendant à montrer que la thèse de BERGMAN serait erronée.

Ainsi, certaines analyses viennent contribuer au débat, mais ce sont celles de LEWANDOWSKI et MAYR[10] qui viennent changer quelque peu la donne. Pour ces derniers chercheurs, le caractère biaisé du raisonnement élaboré par BERGMAN réside dans cet usage de la moyenne qu'il a introduit dans son calcul de la capacité totale dévolue au W.I.A. En l'occurrence, la *moyenne* s'avère très élevée, tandis que la *médiane* des toutes les bases de données impliquées dans les investigations est relativement basse, s'élevant à seulement 4 950 documents. En réalité, il s'avère que le *Top 60* est un argument originellement biaisé, puisque la répartition par taille des bases de données qui le composent est dès le départ faussée. De surcroît, Bergman recourt à une taille exprimée en GB (Giga bites), au lieu d'utiliser le nombre d'enregistrements par base de données ! D'où l'impossibilité de déduire le décompte des enregistrements des données dimensionnelles contenu au sein des fichiers, et ce en raison de dimensions extrêmement variées selon les bases (photos, notices bibliographiques, enregistrements en texte intégral). C'est donc à profit que la Recherche bénéficierait d'une nouvelle collecte, effectuée à partir des bases de données les plus grandes du Web invisible, pour un corpus d'expérimentation beaucoup plus fiable.

Toujours est-il que la classification erronée semble un attribut inhérent à certaines compilations de données collectées pour être ordonnées. Pour preuve, on citera de nombreuses collections de données entreposées dans des bases, comme celle du dispositif *DIALOG*, directement accessible par l'intermédiaire d'une requête Internet. Les 347 fichiers de cette base sont agencés en suivant la logique d'une échelle logarithmique, ce qui montre qu'il existe des bases de données ne comprenant pas moins de 100 000 enregistrements si l'on se réfère aux chiffres mentionnés par WILLIAMS en 2005[11]. Pour la majorité d'entre elles néanmoins, on parle d'une capacité inférieure à moins de 100 000 enregistrements. La répartition effective de ces bases se lit à travers une fonction exponentielle dotée d'une corrélation Pearson élevée égale à 0,96. En termes de tailles, la médiane de ces 347 bases de données tourne autour de 380 000 enregistrements. À ce titre, MAYR et LEWANDOWSKI émettent l'hypothèse selon laquelle le W. I. A. est amené à suivre une répartition exponentielle similaire.

Une fois posées ces bases, il paraît plus aisé d'essayer d'estimer la seule capacité du Web Invisible Académique… Et si l'on reprend pour support de base le *Top 60* bergmanien, il semble que 90 % de cet ensemble puisse être considéré comme du contenu académique ; pourtant, si l'on omet délibérément toutes les bases contenant une

majorité de données brutes, la proportion du contenu académique en jeu ne s'élève plus qu'à 4 %. C'est une hypothèse que vient corroborer l'étude du Web de surface conduite en 1999 par LAWRENCE et GILES[12], lesquels parviennent aux mêmes conclusions.

En tout état de cause, les données brutes font partie intégrante du W. I. A. Les éliminer, c'est réduire de façon drastique et sans raison valable un contenu sans lequel des investigations poussées ne sauraient être validées. À cette effet, et comme l'a déjà dit LEWANDOWSKI[13], parier sur l'indexation du W.I. A. dans son entier, c'est abandonner l'action unilatérale d'une institution unique au profit des efforts concertés d'un groupe d'acteurs qui s'organise. Au service de l'examen de ce processus d'indexation/et ou référencement, WILLIAMS (2005)[14] s'appuie sur les fonds documentaires du *Gale Directory*, un répertoire de données qui ne compte pas moins de 13 000 bases et couvre la plupart des bases académiques d'un contenu qualitatif avéré, aussi bien qu'un certain nombre de bases d'intérêt exclusivement commercial.

Après collecte, WILLIAMS[15] divise l'ensemble des bases de données qui composent le *Gale Directory* en 6 groupes : le premier orienté 'verbe', le second axé sur les chiffres, le suivant rassemblant images et vidéo, un quatrième liés aux documents audio, le cinquième autour des services électroniques et le dernier attaché aux logiciels… Or, pour les bibliothèques et moteurs de recherche académique, ce sont principalement les bases de données axées sur les mots, soit environ 69 % de l'ensemble du *Gale Directory*, qui prévalent. C'est ainsi que sur ces 8 994 bases orientées 'verbe', quelque 80 % sont en *full text* ou bien de l'information provenant de bibliothèques. Pour LEWANDOWSKI, ces chiffres constituent donc un point de départ fiable pour qui cherche à estimer la taille du Web Invisible Académique au plus près de la réalité.

Dans les faits, LEWANDOWSKI se fonde sur un corpus qui s'avère défaillant (certaines bases du *Top 60* de Bergman ne s'y retrouvent pas). Cela dit, l'évaluation de la capacité est établie à 18,92 milliards de documents, la taille moyenne par base de données s'élevant à 1,15 millions d'enregistrements, tout en tenant compte d'une répartition largement biaisée – 5 % d'entre elles comptent plus d'1 million d'enregistrements, certaines plus de 100 millions. Cela dit, si l'on omet les bases de données extrêmement vastes, il s'avère que la taille médiane d'une base est d'environ 150 000 enregistrements. L'estimation de la taille totale est alors calculée en ajoutant les tailles des bases de données connues et en supposant la médiane à 150 000 enregistrements pour toute autre base de données.

Dès lors, et sachant que certaines bases sont manquantes, LEWANDOWSKI[16] conclut qu'on ne peut faire sur la taille du Web

Invisible Académique qu'une supposition motivée, dans la mesure où les chiffres issus de *Gale* sont probablement trop bas. Ce dernier chercheur se base néanmoins sur ces quelques chiffres, à nuancer par conséquent, pour évaluer la taille du W.I.A. entre 20 et 100 milliards de documents. En définitive, l'estimation dimensionnelle du Web Académique invisible qu'il établit se situe dans la fourchette de tailles englobant les plus grands moteurs de recherche du Web de surface.

Accéder au Web académique : quelques approches

Pour promouvoir l'accès au Web, et plus précisément dans sa dimension académique, l'étude des systèmes d'indexation de documents et de l'accessibilité à "la Toile" en général – autrement dit les supports de requête ou moteurs de recherche spécifiques existant à l'heure actuelle – doit être au centre des préoccupations. En effet, le Web tel qu'il se présente aujourd'hui offre plusieurs modèles. Or, par leurs caractéristiques inhérentes ou leurs limites, certains d'entre eux sortent du lot. C'est en ce sens qu'ils nécessitent une analyse plus approfondie.

Au nombre des dispositifs à distinguer – qui ont pour axe commun de se focaliser sur l'Information –, se trouvent d'abord *Google Scholar* et *Scirus* [17], des projets initiés par des entreprises commerciales. Au cœur de leurs ressources, on trouve les banques de données nées à l'initiative des éditeurs, auxquels il convient d'ajouter des matériaux disponibles en accès libre.

À l'heure actuelle, c'est bien *Google Scholar* qui demeure l'approche la plus débattue sur la place publique. En effet, la version bêta est en ligne depuis 2004 et indexe une part substantielle des éditeurs internationaux dits STM, c'est à dire orientés Sciences, Techniques et Médecine. À ces derniers se sont joints quelques éditeurs à l'origine du *Cross Reference Initiative*. Et c'est ainsi que Google a pu mettre en place un prototype doté d'un potentiel manifeste, aux dires de LEWANDOWSKI[18], MAYER et WALTER (2005)[19], mais qui présente néanmoins quelques écueils dont on se passerait volontiers… Pourtant, pour pallier ces quelques défauts intrinsèques, *Google Scholar* cherche à intégrer à son système la technique de l'*influential citation measure*, laquelle a vu le jour à l'initiative de l'*Institute of Scientific Information* (ISI), avant d'être implémentée dans le *Science Citation Index*, auquel succède bientôt le *Web of Science*. Malheureusement, et c'est ce que dénonce JACKSÓ en 2005[20], Google ne fournit aucune documentation qui puisse assurer la transparence du service; impossible d'évaluer la couverture documentaire exacte ou le degré d'actualisation proposés par les services de Google, ajoute le chercheur WALTER, dans l'étude empirique dédiée qu'il publie à la même époque.

De son coté, *Scirus* apparaît comme un dispositif hybride combinant contenu du Web visible et information collectée à partir du W. I. A, puisqu'il s'agit d'un moteur de recherche scientifique indexant à la fois le Web académique visible, mais également quelques compilations de données issues de sources en accès libre ou bien à mettre sur le compte du *Science Direct* d'Elsevier. Avec environ 50 millions de documents indexés à partir du Web visible, *Scirus* est de loin le plus imposant des moteurs de recherche scientifiques développés avec la technologie norvégienne *FAST* (MC KIERNAN, 2005)[21].

À ces deux premiers systèmes que sont *Google Scholar* et *Scirus*, viennent s'ajouter le *Bielefeld Academic Search Engine (BASE)*[22] et *Vascoda*[23] ; ce sont tous deux des projets universitaires dans lesquels s'investissent aussi bien des bibliothécaires que des fournisseurs d'information, qui ont choisi d'ouvrir leurs collections. Ces dernières sont principalement constituées de bases de données académiques faisant autorité, de catalogues et bibliothèques virtuelles, ainsi que de documents additionnels libres de droits (que l'on peut notamment trouver sur le Web visible). Il va sans dire que chacun des systèmes en question a, ou aura tôt ou tard, recours à la technologie propre aux moteurs de recherche. Ces modèles s'enrichissent, en outre, de leurs propres outils (ex. : banque de citations, filtrage spécifique ou traitement spécifique de l'hétérogénéité).

Si l'on reprend la définition énoncée par LOSSAU en 2004 [24], *BASE* est pour sa part un moteur de recherche qui intègre une information mêlant les données issues des catalogues de la Bibliothèque de l'Université de Bielefeld à celles obtenues à partir d'environ 160 sources en libre accès, soit l'équivalent de 2 millions de documents. Le système fait lui aussi appel à la technologie des outils de requête *FAST*. À ses côtés, il y a *Vascoda*, le prototype du portail scientifique interdisciplinaire, lequel intègre les collections érigées par les bibliothèques, des bases de données relatives aux ouvrages scientifiques et quelques éléments de contenu académiques en surplus. *Vascoda* fonctionne comme un "méta-portail" : il délègue les requêtes elles-mêmes, c'est-à-dire qu'en aval, il les confie à certains instruments d'orientation spécifique ou bien encore à des sous-communautés dédiées. Dans la pratique, chacun des domaines composant le système *Vascoda* est responsable de son propre portail disciplinaire, dont la structure peut reposer sur de multiples technologies. C'est pourquoi *Vascoda* peut être considéré, en toute légitimité, comme *le* modèle alternatif visant à combler la sorte d'espace interstitiel que constitue le W. I. A. Sont à l'origine du dispositif les professionnels des bibliothèques et centres de documentation allemands, qui lanceront bientôt la dernière version de *Vascoda*, supportée elle aussi par la technologie *FAST*.

La conception et le développement de ces différents systèmes d'accès et/ou référencement des divers contenus Web donne un sentiment de foisonnement, lequel montre encore, s'il en était besoin, qu'une indexation réelle et bien menée du Web académique repose définitivement sur la concertation des acteurs en présence… Car toute approche unilatérale présente des forces et faiblesses bien spécifiques, que seuls des efforts conjoints sont à même d'effacer. La situation qui en découle actuellement est tout à fait éloquente : d'un côté, il y a persistance d'une couverture documentaire étendue mais largement polarisée puisque confinée aux résultats d'ordre commercial, avec une incapacité notable à exclure des résultats des requêtes les enregistrements qui ne relèvent pas du secteur universitaire ; de l'autre résiste une vision du W.I.A. encore très succincte, alimentée par des ressources en texte intégral passablement minoritaires.

4. La bibliothéconomie : un levier de développement pour le Web académique

Au regard des fonds documentaires qui se destinent plus spécifiquement à la recherche, Il importe de garder à l'esprit que le Web Invisible Académique présente une véritable valeur ajoutée aux yeux des universitaires, professionnels des bibliothèques et autres chercheurs du secteur académique. On sait aujourd'hui que le W. I. A. est en mesure de fournir tout ce qui a trait aux procédures scientifiques. C'est pourquoi la taille et le format des données mises en jeu s'avèrent une question centrale. Dans les faits, l'ensemble de cette information scientifique – des données de référence – comporte aussi bien des études écrites (articles ou ouvrages universitaires) que des données brutes (ex. : informations tirées d'un sondage) et du contenu purement numérique, c'est à dire exclusivement accessible en ligne (ex. : documents libres de droits).

Ce constat fait, il nous appartient de rappeler qui sont les principaux fournisseurs de contenu internet dévolu au W. I. A ; car sur un plan plus orienté bibliothéconomie, c'est une véritable dimension pivot que constitue l'appréhension globale du Web académique. Il s'agit d'abord des fournisseurs de Bases de données documentaires, qui compilent un ensemble de métadonnées bibliographiques enrichies par l'indexation manuelle (thesaurii, classifications et autres systèmes d'organisation des connaissances) et délivrent en outre certains services complémentaires tels que la livraison de documents. A leurs côtés, se trouvent les bibliothèques, qui produisent elles aussi des compilations d'ouvrages à travers un système de gestion en libre accès : en l'occurrence, ce sont ce qu'on nomme OPACs ou *Online Public Access Catalogues*. Elles proposent elles aussi leurs données en les enrichissant *via* indexation manuelle ou en y ajoutant des services complémentaires. On trouve

ensuite les éditeurs commerciaux, lesquels fournissent essentiellement du contenu en texte intégral, mais aussi certains répertoires de sociétés et associations, ainsi que des répertoires en accès libre tels que *Citebase* ou *OpenROAR*.

Or, nombre de ces matériaux ne sont pas forcément directement issus du W.A. I. Néanmoins, ils sont régulièrement mis au jour par les principaux outils de recherche internet. Pour les utilisateurs confrontés à ces différents systèmes et structures d'information (qui varient selon la source), la démarche à adopter renvoie à un nécessaire processus d'adaptabilité. C'est ainsi que la plupart des fournisseurs de fonds documentaires universitaires conservent leurs propres accès par domaine et modèles de structuration de l'information, en raison de leurs coutumes historiques et du type de contenu indexé. Les bibliothèques, plus précisément, indexent surtout les ouvrages et "compilations" de données au moyen de fichiers normalisés qui font aujourd'hui autorité, tandis que les fournisseurs plus orientés bases de données ont recours à des *thesaurii* thématiques (*domain specific*) et classifications au format propriétaire pour indexer les articles de presse. Face à eux, les éditeurs combinent indexation manuelle et indexation automatique au regard des documents qu'ils traitent, des textes intégraux ou documents en *full text*.

C'est l'occasion pour le scientifique KRAUSE de parler de "*polycentral information provision*" et de conclure en 2003 à la complexité de la situation de recherche de l'utilisateur final, un internaute qui peut se retrouver bien dérouté face au *cross data-base searching*. En l'occurrence, il s'agit là d'une recherche de données transversales ou multi-bases qui peut mettre une requête à mal… En d'autres mots, on peut parler de fracture informationnelle, dans la mesure où l'on passe du système de publication imprimée traditionnel à celui de l'édition numérique : les moteurs de recherche affluent, tandis que l'on voit se décentraliser les différents fournisseurs d'information.

Face à ce tableau des acteurs pivots de l'indexation documentaire qui se veut assez exhaustif, les chercheurs allemands LEWANDOWSKI et MAYR [25] tentent de définir le Web Invisible Académique. Pour ce faire, ils reprennent à leur compte le point de vue du chercheur LOSSAU[26], lequel prône une ouverture accrue des bibliothèques sur un Web académique qui demeure pour ces dernières encore bien obscur. Car, dans la pratique, un certain pourcentage des données entreposées au sein des bases universitaires pâtit de cette indexation défaillante qui se manifeste à travers les efforts conduits en matière de bibliothéconomie ou sciences documentaires. En bref, il s'agirait pour cet acteur clé qu'est la bibliothèque d'orienter beaucoup plus ces efforts sur un développement des technologies liées aux moteurs de recherche auxquels elle recourt.

Pour y remédier, c'est donc une démarche collaborative que le secteur de la bibliothéconomie aurait tout intérêt à mettre en place, et ce afin de favoriser la "mise en visibilité" d'un contenu qui reste pour beaucoup quasi inexistant. Tout en gardant toutefois à l'esprit que le W. I. A. est loin de centraliser tous les fonds documentaires se rapportant aux bibliothèques et catalogues online, dans la mesure où il ne représente qu'une partie du Web seulement, et que nombreux sont les documents tout à fait pertinents que l'on peut également trouver sur le Web de surface ou Web visible.

5. Le Web académique maghrébin

Pourquoi un observatoire maghrébin du Web académique ?

Un examen approfondi du monde de la recherche documentaire et de l'Information en général vient mettre en exergue un fait de premier plan : s'est imposé au travers de mutations d'ordres technologique et informationnel, aussi bien en Europe et aux États-Unis que dans les états du Maghreb, un modèle d'accès aux ressources académiques en ligne quasi standardisé… Dans les faits, la situation dénote une sous-utilisation des contenus du Web académique, parmi lesquels on trouve notamment les bases de données et revues spécialisées. Cette "sous-utilisation" est à déplorer, et ce d'autant plus que des dépenses considérables sont régulièrement consenties par les universités maghrébines en vue de s'abonner à des sources électroniques telles que bases de données ou revues scientifiques sur internet. A cet effet, il faut souligner que le système national de documentation en ligne (SNDL) dépouille chaque année le budget du ministère de l'enseignement supérieur de pas moins de 170 millions de dinars algériens pour un retour sur investissement non évalué à ce jour.

Or, la volonté d'établir un observatoire maghrébin du Web académique découle de cette velléité de mieux cerner les modalités d'une recherche documentaire en milieu universitaire et les traits d'un environnement technique et technologique spécifiques au Maghreb : en l'occurrence, il s'agit d'une tentative de diagnostic visant à suivre l'évolution du Web universitaire en vue de mesurer et analyser en son sein les divers usages à l'œuvre au sein de la communauté académique. C'est aussi l'occasion de porter un regard critique sur l'orientation de la Recherche et Développement au Maghreb, dans laquelle le secteur académique joue un rôle des plus actifs.

Dans la pratique, établir un observatoire du Web académique maghrébin, c'est pour la communauté scientifique le moyen d'atteindre un certain nombre d'objectifs fondamentaux, parmi lesquels on trouvera (la liste n'est pas exhaustive) la volonté de :

- Pratiquer une veille informationnelle sur l'évolution du Web académique et les comportements de recherche d'information de la communauté universitaire maghrébine ;
- Définir les différents impacts, enjeux et défis du Web 2.0 dédié au secteur académique maghrébin, ce qui constitue également une problématique de taille.
- Pousser l'acteur humain à mieux maîtriser les compétences informationnelles et se servir au mieux des fonctions et services du Web, en vue d'améliorer son niveau éducatif, social et économique
- Instaurer une culture informationnelle autour du rôle et de l'impact du Web sur les activités de recherche et développement au Maghreb ;
- Capitaliser les savoirs et/savoir-faire des chercheurs maghrébins en les encourageant à se manifester davantage sur le Web, de manière à refléter leurs activités aussi bien quantitativement que qualitativement ;
- Créer un réseau maghrébin de spécialistes de l'information et de la documentation, et organiser des ateliers de formation thématiques sur le Web en fonction des différents profils rencontrés.

***Quel impact sur la Recherche et le Développement* ?**

En toute société, mais encore plus au Maghreb – comme dans les autres pays du Sud –, le développement économique est fortement subordonné à la capacité à produire d'une part, et à celle de transmettre connaissances et information de l'autre. C'est en ce sens que la visibilité, qui prend une dimension particulière au 21^{e} siècle en acquérant une dimension digitale toujours plus grande, est à prendre en considération. Pour ce faire, les promoteurs de cette visibilité doivent s'attacher à développer :

- Une stratégie d'édition - diffusion collant aux réalités économiques…
- Les subventions allouées à l'accessibilité des banques de données scientifiques – sachant que la plupart des données de ce type sont encore l'apanage des pays plus développés…
- Des infrastructures performantes susceptibles de supporter l'expansion fabuleuse des nouvelles technologies de l'information.

En d'autres mots, une Recherche scientifique bien menée repose forcément sur une accessibilité à des résultats de recherches, qui n'est pas bridée et qui n'a de cesse d'être améliorée. À cet égard, le phénomène des archives ouvertes qui a pris son envol au sein du monde académique, dans les pays du Sud, est éloquent : au service de la liberté et de la gratuité d'accès aux ouvrages scientifiques à l'international, il symbolise les premiers soubresauts de l'*Open* Access, ce mouvement qui s'est développé autour des ressources numériques, qu'un bon nombre d'experts jugent indissociable d'une ouverture sur les contenus *onlines*. C'est un point qui risque de poser problème, sachant que les chercheurs arabes éprouvent encore quelques réticences à l'idée de faire

procéder à l'auto-archivage de leurs productions... Pourtant, l'*Open Access* constitue la voie royale pour qui cherche à intensifier les échanges d'information et à développer, par conséquent, son économie numérique.

6. Pour conclure:

Un Web académique maghrébin en devenir : la promesse d'un fonds documentaire enrichi

Désormais, nous voici plus à même de cerner, au Maghreb, les enjeux de la transmission de l'Information en général, et de la gestion de la Recherche documentaire plus précisément. Car nous l'avons vu, certains acteurs de premier plan, ainsi que certains concepts viennent nécessairement prendre part à la mise en place d'une stratégie de développement économique qui doit aussi se baser sur un traitement optimal de l'information. En effet, il s'agit de garder à l'esprit que Recherche scientifique et développement économique ont partie liée.

À cet effet, c'est avec un nouveau secteur, celui de la bibliothéconomie, que les promoteurs du progrès (politiques et institutions) se doivent de composer. C'est donc une action concertée des différents acteurs du secteur documentaire – bibliothèques, centres de recherche et éditeurs numériques, notamment, - qu'il faut viser. Bien évidemment, il ne s'agit pas pour autant de réduire la recherche universitaire à une simple histoire de profit puisqu'il s'agit aussi d'Information au sens large… Néanmoins, l'émergence des principes bibliothéconomiques souligne qu'à l'instar des autres pans de l'économie, l'Information est un matériau, mais également un produit. Et comme tout produit, il nécessite de la part des acteurs (aussi bien dans les pays très industrialisés qu'au sein des états moins développés) qui le prennent en considération de définir et mettre en œuvre les outils de gestion adéquats.

Dans cette perspective, les acteurs de la bibliothéconomie maghrébine se doivent, eux aussi, de se donner les moyens, de découvrir le Web Invisible Académique, cette partie de "la Toile" qui recèle des ressources en information inexplorées, et parfois insoupçonnées. Partant du lien étroit existant entre secteur académique et bibliothèques – dans le sens où la gestion documentaire, au cœur des préoccupations de ces dernières, est indissociable des avancées de la Recherche universitaire – c'est bien la virtualisation de la gestion documentaire qui est en jeu : c'est pourquoi les différentes institutions doivent se donner pour mission d'adopter une démarche collaborative susceptible d'améliorer non seulement la rapidité des outils de requête (technologie(s) propre(s) aux moteurs de recherche), mais également la capacité de couverture (*broad coverage)* des dispositifs de recherche scientifique mis en place par les acteurs du secteur, et en particulier les Bibliothèques.

Dans ce cadre, la vision étriquée consistant à considérer qu'une bibliothèque virtuelle unique est à même d'indexer l'ensemble du Web Académique maghrébin est à proscrire. C'est pourquoi les fournisseurs commerciaux de moteurs de recherche – dotés d'une certaine supériorité technologique et financière – se doivent de travailler de concert avec les bibliothèques, ces dernières disposant d'une expérience aboutie de la collecte d'information, de modèles rodés d'accès aux disciplines et de structuration des données (à travers l'établissement d'ontologies et de taxonomies notamment), en mêlant leurs effort à ceux des éditeurs et fournisseurs de bases de données susceptibles d'apporter leur contribution en "ouvrant" encore plus qu'il ne le font leur collections.

Au final, on retiendra d'experts comme LEWANDOWSKI quelques recommandations des plus pertinentes : concentrer ses efforts sur la *répartition infométrique* (principes statistiques appliqués à la gestion de l'information) dans le but de mieux évaluer la taille du Web Invisible Académique; classifier les différents contenus dans le but de savoir dans quelle mesure chaque discipline contribue à la teneur du W. A. I. et bâtir des moteurs de recherche spécialisés ; faire en sorte, enfin, que les bibliothèques suivent l'exemple des fournisseurs de bases de données qui rendent leurs collections disponibles sur le Web visible, à l'image de Google, afin d'optimiser les processus d'indexation documentaire. Appliquées à la réalité maghrébine, il est bien certain que ces préconisations constituent quelques clés de développement technologique pour un Maghreb soucieux de modernité.

Notes :

1. Sherman, C. and Price, G. (2001), The Invisible Web: Uncovering Information Sources Search Engines Can't See, Information Today, Medford, NJ.
2. Bergman, M.K. (2001), "The Deep Web: surfacing hidden value", Journal of Electronic Publishing, Vol. 7, No. 1, En ligne: www.press.umich.edu/jep/07-01/bergman.html (Consulté le 30/05/2014).
3. Op. Cit
4. LEWANDOWSKI, D. (2001, 2005) : Web Information Retrieval (DGI, Franckfort)
5. BERGMAN, M. K. (2001) : "The Deep Web: surfacing hidden value" in Journal of Electronic Publishing (vol. 7)
6. Lyman, P., Varian, H.R., Swearingen, K., Charles, P., Good, N., Jordan, L.L., et al. (2003), "How much information 2003?" En ligne: [www.sims.berkeley.edu/research/projects/how-much-info-2003] (Consulté le 18/12/2014).
7. LAWRENCE, S. et GILES, C. L. (1999) : "Acessibility of information on the Web" in Nature (vol. 40)
8. Op.Cit

9. Stock, W.G. (2003), "Weltregionen des Internet: Digitale Informationen im WWW und via WWW", Password, Vol. 18, No. 2, pp. 26-28.
10. Lewandowski, Dirk et Mayr, Philipp. Exploring the Academic Invisible Web. En ligne: [http://www.ib.hu-berlin.de/~mayr/arbeiten/lewandowski-mayr_LHT06.pdf],(Consulté le 03/01/2015)
11. Williams, M.E. (2005), "The state of databases today: 2005", in Gale Directory of Data-bases, Vol. 2, pp. XV-XXV, Gale Group, Detroit, MI.
12. Op. Cit
13. Op. Cit
14. WILLIAMS, M. E. (2005), "The state of databases today: 2005" in Gale Directory of Data-bases (vol. 2)
15. Op. Cit
16. Op. Cit
17. http://scholar.google.com
18. (Il est regrettable de signaler la fermeture fin janvier 2014, du moteur de recherche SCIRUS consacré exclusivement à la recherche scientifique) http://corist-shs.cnrs.fr/fermeture_scirus
19. Op. Cit
20. Op. Cit
21. JACSÓ, P. (2005) "Google Scholar: the pros and the con", in Online Information Review, vol. 29
22. MC KIERNAN, G. (2005) : "E-profile: Scirus… for scientific information only" in High Tech News, vol. 22
23. http://www.base-search.net/
24. http://www.vascoda.de
25. LOSSAU, N. (2004) : "Search engines technology and digital libraries: libraries need to discover the academic internet", in D-Lib Magazine (vol. 10)
26. Op. Cit