Bibliothèque nationale de France

Les Archives de l'Internet

Une étude prospective sur les représentations et les attentes des utilisateurs potentiels

Bibliothèque nationale
de France
délégation à la Stratégie et à la Recherche

Auteurs : Philippe Chevallier et Gildas Illien
Contacts : philippe.chevallier@bnf.fr
gildas.illien@bnf.fr

TABLE DES MATIERES

1. Résumé exécutif

1.1. Principaux résultats

Une étude qualitative a été conduite fin 2010–début 2011 par la délégation à la Stratégie et à la recherche de la Bibliothèque nationale de France (BnF), en lien avec la direction des Collections et la direction des Services et des réseaux de la BnF, auprès de publics potentiels des archives de l'Internet afin d'explorer leurs besoins en termes de contenus et de services. Il s'agissait également d'analyser leurs représentations de ces archives pour identifier les moyens permettant d'accroître leur consultation. Quinze entretiens ont été réalisés au sein de trois populations : 1) chercheurs (histoire, philosophie, sociologie, sciences et techniques), 2) professionnels (avocat, consultant marketing, documentaliste, ingénieur brevet, journaliste), 3) le « tout venant » de la bibliothèque de Recherche sur le site François-Mitterrand.

Les chercheurs interrogés travaillent dans un univers web dont ils reconnaissent à la fois la richesse et la volatilité. Si l'intérêt d'une mémoire du web leur paraît évident, ils se heurtent à la difficulté de définir et circonscrire, dans un espace qui semble illimité, des corpus significatifs. Face à cette difficulté, la BnF est perçue comme un tiers de confiance capable de garantir au chercheur l'accès à des collections raisonnées et documentées. Les chercheurs ont également besoin que l'histoire du web, aujourd'hui disséminée dans les souvenirs de quelques spécialistes, soit reconnue, préservée et partagée.

Les archives soulèvent cependant pour les chercheurs des questions éthiques et méthodologiques :

- tout sur le web n'est pas de l'ordre d'une publication mise à disposition d'un public. En particulier, ce qui relève d'actions personnelles (discuter, acheter, participer à des réseaux sociaux, etc.) apparaît impropre à l'archivage, même s'il constitue un lieu d'observation particulièrement riche pour l'historien et le sociologue ;
- archiver un flux semble impossible, voire paradoxal. Les archives du web requièrent un effort de définition et de modélisation préalable, car elles ne peuvent être assimilées à une archive traditionnelle renvoyant à des unités documentaires stables : le « site », capturé isolément et ponctuellement, peut difficilement jouer ce rôle, car c'est son inscription dans un réseau et dans le temps qui intéresse d'abord les chercheurs. En outre, il n'existe pas encore de méthodologie bien définie pour analyser et utiliser les sources du web, si bien que les chercheurs hésitent à manipuler ce type de matériau.

Les entretiens avec **les professionnels** et **le « tout venant » de la bibliothèque de Recherche** ont permis de relever la présence de représentations concurrentes qui viennent brouiller la perception de ce que sont les archives du web. Les moteurs de recherche, mais aussi les archives en ligne proposées par certains sites comme les blogs donnent l'illusion à l'internaute que le web s'auto-archive. Pour ces deux catégories de public, les archives de l'Internet doivent préalablement démontrer leur pertinence face à un réseau qui semble déjà d'une profondeur et d'une mémoire infinie.

Au niveau des usages potentiels, **les chercheurs** reconnaissent le caractère encore spécialisé d'une recherche sur les archives de l'Internet ; mais leurs pratiques personnelles d'archivage ou de consultation du site de la fondation Internet Archive révèlent que ce type de recherche est appelé à se développer dans la durée et pour des besoins réguliers. **Les professionnels** interrogés ont en revanche manifesté un intérêt surtout ponctuel pour ces archives, celles-ci se situant à la marge de leurs besoins professionnels les plus essentiels. Certains se disent déjà dépassés par la masse d'informations qu'ils n'ont pas le temps de traiter. L'intérêt manifesté suffit difficilement à compenser le coût horaire que représenterait pour eux une visite à la BnF, institution dont ils ont une image à la fois floue et intimidante. Par conséquent, ils sont fortement demandeurs de services documentaires et de reproduction à distance. Enfin, le **« tout venant » de la bibliothèque de Recherche** semble très éloigné de tout usage potentiel ; mais cet éloignement est d'abord fonction d'une distance plus générale à l'égard du web comme lieu de recherche. Seuls les grands usagers du web, qui ont déjà une mémoire fine de cet univers, sont susceptibles d'être intéressés. L'équivalent pour le numérique de ce que sont le généalogiste ou le chercheur amateur pour les fonds patrimoniaux traditionnels est une cible à privilégier en termes de communication et de médiation.

Ces trois communautés s'accordent enfin sur un point : d'une part, il est impossible de préjuger des contenus qui auront à l'avenir un intérêt pour les chercheurs professionnels et amateurs. Mais

d'autre part, la sélection est légitime car elle est constitutive de tout acte d'archivage et les volumes existants la rendent inévitable. De ce point de vue, l'étude confirme que la stratégie retenue par la BnF, c'est-à-dire le « modèle intégré » associant des collectes de grande ampleur et des sélections humaines plus ciblées, apparaît comme le meilleur moyen de répondre à cette demande contradictoire.

1.2. Recommandations

Valorisation : adopter une démarche de communication plus adaptée et plus transparente

- justifier et documenter les principes de collecte et critères de sélection ; préciser et publier la politique documentaire de la BnF pour la collecte du web ;
- disposer d'une base d'exemples précis permettant de montrer les disparitions effectives de contenus en ligne, et donc l'intérêt des archives ;
- explorer et utiliser dans la communication les métaphores qui permettent à différents types de publics de se représenter le plus facilement possible ces archives.

Communautés : développer des partenariats et s'insérer en tant que tiers de confiance dans les réseaux de chercheurs universitaires ou amateurs où se trouvent les usagers potentiels

- engager des actions de dissémination ciblées auprès des chercheurs dans des conférences ou des colloques, participer à des instances et des projets *ad hoc* ;
- élaborer avec les communautés de chercheurs une méthodologie, voire un guide des sources visant à faciliter, promouvoir et légitimer l'usage des archives dans le cadre de travaux scientifiques ;
- rejoindre un public d'érudits amateurs du web, en partant des réseaux de chercheurs déjà connus et des contacts déjà établis par chaque département de collections de la BnF ;
- à ces fins, valoriser en interne et en externe la mission de veille des correspondants DLWeb dans les départements de collections, qui peuvent interagir directement avec des communautés scientifiques ciblées dans le but de promouvoir l'utilisation des collections voire de susciter des « dons » numériques.

Services : développer des services et des outils à distance, en particulier pour les professionnels

- donner aux internautes la possibilité de proposer en ligne leur site (ou d'autres sites ?) à archiver par la BnF au titre du dépôt légal ;
- mettre à disposition des outils pour savoir si un site est archivé et se repérer dans les archives, si possible à distance, même sans accès au document primaire ;
- développer des services de recherche documentaire à distance à destination des professionnels : recherche déléguée, authentification, datation, citation, reproduction, etc. ; explorer les possibilités de services payants et de ressources propres dans les limites du cadre juridique existant.

Contenus : maintenir la politique patrimoniale actuelle en renforçant une veille et des collectes plus sensibles à l'actualité du web

- développer des collectes qui permettent de garder la trace des nœuds importants et des réseaux ; archiver les sites les plus populaires, les pages de résultats de Google sur les recherches les plus fréquentes, etc. Plus largement : ne pas collecter seulement des éléments discrets ou isolés mais garder la trace des pratiques du web qui marquent « l'air du temps » et documentent les tendances, sociales, commerciales, etc., à grande échelle ;
- développer ou acquérir des outils permettant de travailler en ce sens.

2. Rappel du contexte et méthodologie de l'étude

2.1. Contexte et objectifs

Depuis la loi du 1er août 2006, la Bibliothèque nationale de France a en charge le dépôt légal de l'Internet français. Initiée en avril 2008, la consultation de ces archives, d'abord disponible sur une dizaine de postes informatiques, puis progressivement étendue à l'ensemble des postes des salles de lecture de la bibliothèque de Recherche, demeure expérimentale dans l'attente de la publication du décret d'application de la loi du 1er août 2006.

Les utilisateurs de ces archives disposent actuellement de trois modes de recherche ou de navigation :

- la recherche par adresse URL ;
- la recherche par mot (mais la plus grande partie des collectes n'est pas indexée) ;
- les quatre parcours thématiques réalisés par des bibliothécaires : le Web vert (01/2011), le Web militant (10/2009), S'écrire en ligne : journaux personnels et littéraires (02/2009), Cliquer, voter : l'Internet électoral (03/2008).

Compte tenu de l'état d'avancement de la réflexion de la communauté internationale et notamment de l'International Internet Preservation Consortium (IIPC) sur la question des fonctionnalités à mettre en œuvre pour la consultation des archives de l'Internet, il n'a pas semblé opportun de lancer une étude d'usages approfondie de l'interface BnF qui aboutirait sans doute aux mêmes résultats que ceux fournis par les précédentes études internationales depuis 2007 (British Library 2007, Library and Archives Canada 2007, Koninklijke Bibliotheek 2007, Bibliothèque nationale d'Israël 2009).

L'objectif de la présente étude était de mener une enquête auprès de publics potentiels des archives de l'Internet, afin d'explorer leurs besoins en termes de contenus (collections) et de services. Il s'agissait également d'analyser leurs pratiques de recherche sur le web et leurs représentations de ces archives afin d'identifier les moyens permettant d'accroître leur consultation.

Une cible de quinze entretiens qualitatifs a été déterminée, répartis en trois populations distinctes :

- chercheurs (sciences politiques et juridiques, sociologie des organisations, sociologie du militantisme et des mouvements sociaux, recherches sur l'autobiographie, histoire des pratiques et des traces d'écriture, histoire des arts et du spectacle) ;
- professionnels (juristes, journalistes, consultants en e-reputation, communicants d'entreprise ou d'administration publique) ;
- le « tout venant » de la bibliothèque de Recherche (site François-Mitterrand).

2.2. Constitution de l'échantillon

L'intérêt potentiel des publics-cibles a été défini en termes de métiers ou de disciplines, de manière intuitive ou en fonction des demandes d'informations parvenues au service du dépôt légal numérique. Dans un second temps, les métiers et disciplines retenus ont été croisés avec des thèmes de recherche ou d'activité plus précis, incluant par exemple les mots « numérique », « propriété intellectuelle », « net art », etc., afin de déterminer un échantillon d'usagers potentiels pertinent. Les contacts ont été pris sans présumer de l'intérêt ni d'une quelconque connaissance de l'archivage de l'Internet de la part des personnes interrogées. Afin de garantir la plus grande neutralité, les personnes interrogées ont été choisies hors des contacts développés par le service du dépôt légal numérique de la BnF.

Sur les quinze personnes interrogées, trois seulement avaient entendu parler des archives de l'Internet, mais sans avoir eu la curiosité de venir les consulter. L'éloignement des personnes interrogées par rapport à tout usage actuel, et parfois même potentiel, de ces archives a donné aux entretiens une teneur nécessairement plus générale, faisant remonter des questions préalables de pertinence de la collecte et de définition même de l'archive. Au lieu de les interroger plus avant sur l'offre actuelle des archives de l'Internet (périodicité et types de collectes, etc.), il a semblé plus intéressant de comprendre en profondeur leur propre pratique d'archivage et de recherche, afin que le service du dépôt légal numérique puisse lui-même réfléchir à la politique documentaire et à l'interface qu'il conviendrait de développer en regard.

Les entretiens, d'une heure environ, ont été menés à partir d'un « guide d'entretien » (voir Annexe 1). Ils ont été intégralement enregistrés et retranscrits pour être analysés.

2.3. Liste des personnes interrogées

Les chercheurs

- **Un historien**, chargé de recherche. Thèmes de recherche : écritures ordinaires, écriture de soi, fonction sociale de l'archive.
- **Un sociologue**, chercheur. Thèmes de recherche : corps, santé, usages informatiques, réseaux sociaux.
- **Un sociologue**, maître de conférences. Thèmes de recherche : pratiques d'écriture et de lecture dans le monde professionnel.
- **Un philosophe**, doctorant. Thèmes de recherche : maladie mentale, gestion du risque, politique pénale.
- **Un ingénieur**, chercheur. Thèmes de recherche : art et technique, net art.

Les professionnels

- **Un consultant** au sein de l'unité marketing d'un centre de formation professionnelle, directeur associé d'un cabinet de conseil en stratégie marketing et E.réputation.
- **Un grand reporter** dans un magazine, spécialisé dans l'actualité littéraire.
- **Un avocat** au Barreau, spécialiste en droit de la propriété intellectuelle.
- **Un ingénieur brevet** et **un responsable du service veille et documentation** dans un cabinet de conseil en propriété industrielle.
- **Un documentaliste**, responsable du centre de documentation d'une unité mixte de services (UMS).

Le « tout venant » de la bibliothèque de Recherche

- **Alain (France)**, 65 ans, chercheur amateur – histoire régionale
- **Amel (autre pays)**, 34 ans, doctorante en économie – économie du tourisme
- **Newman (autre pays)**, 28 ans, doctorant en philosophie – philosophie des sciences
- **Paul (autre pays)**, 60 ans, professeur d'histoire – histoire contemporaine
- **Pierre (France)**, 32 ans, doctorant en histoire – histoire de l'art

3. Résultats

3.1. Les chercheurs

3.1.1. Pratiques du Web

L'usage du Web est aujourd'hui omniprésent pour les chercheurs en sciences humaines, non seulement comme possibilité d'accéder à de la documentation scientifique (articles en ligne), mais également comme terrain de recherche, plus ou moins formalisé. Par effet de retour, le Web devient un lieu d'exposition du chercheur lui-même, désormais actif dans les réseaux sociaux et sur les blogs.

Malgré l'ouverture de ce terrain de recherche, la mention dans les travaux scientifiques de documents disponibles en ligne reste problématique à cause de la difficulté à constituer des corpus relativement maîtrisés et partagés par la communauté des chercheurs.

➔ A la recherche d'un « air du temps »

Indépendamment de programmes de recherche directement liés au Web ou à l'étude des représentations, le Web est cité comme un « *début de terrain* » : il fournit une première impression, sur le mode de la promenade (« *j'étais allé voir un peu ce qu'on racontait* »), sans méthode bien définie (« *sans en faire une analyse plus que ça* ») et ramenée à une pratique privée (« *c'est un peu empirique, car c'était plus pour ma propre information* »). Ce temps d'exploration vaut en particulier pour les chercheurs s'intéressant à des nouveaux objets, « *mouvants, très innovants* » (comme l'urbanisme, le design) ou des sujets qui font débats dans l'opinion publique (pédophilie, sécurité). Les mots utilisés pour décrire le contenu observé insistent à la fois sur sa nouveauté et sa futilité : « *dernier truc* », « *air du temps* », « *petits papiers* », « *petits objets* », opposée aux « *papiers académiques* ». Cette insignifiance est précisément ce qui intéresse le chercheur :

> « *C'est presque un début de terrain pour moi, une forme d'air du temps que j'observe, des petits objets à faire étudier à des élèves.* [Ce que je cherche,] *en fait, c'est rarement des papiers académiques, c'est souvent des petits papiers qui permettent de comprendre ce que ça veut dire de parler de ces choses-là en ce moment et quelle est la dernière innovation, le dernier truc.* »

Ces « *petits papiers* » se trouvent en particulier sur les blogs : ceux des chercheurs, mais également ceux tenus par des consultants, simples passionnés ou praticiens qui se « *fabriquent une expertise* ». C'est un ensemble de sources « *hétérogènes* », où les acteurs du terrain sont plus nombreux que les universitaires, où les informations objectives croisent les rêves et les utopies. « *Il y a beaucoup de blabla* », mais il faut justement capter ce « *blabla* », pour étudier les manières de parler. Ces manières de parler sont non seulement un objet en tant que tel, mais elles permettent au chercheur de positionner son propre discours : il s'agit d'étudier comment des acteurs de la société civile parlent de manière non académique des objets qui l'intéressent, pour « *essayer de comprendre quelle est la différence entre comment ils en parlent eux et comment j'en parlerais moi* ».

➔ La veille filtrante

Au niveau de leurs stratégies de recherche sur le Web, si certains continuent de recourir à Google, tout en reconnaissant les limites évidentes du moteur, d'autres chercheurs développent des pratiques de veille filtrante : ils ne font plus de recherche directement sur l'Internet, mais ils paramètrent leur « *écoute du Web* » en sélectionnant des sources de veille :

> « *Je cherche de moins en moins de choses sur l'Internet car je fais de la veille. Je suis dans un autre truc où je filtre un flux qui m'arrive comme ça* […], [d']*en haut. Il y a toujours une dimension d'aller chercher le bon flux qui est un peu loin, mais je suis rarement en position de chercher, mis à part des papiers, des choses très stables, des choses qui ne sont pas que sur l'Internet : un article.* […] *Sinon, j'essaye de calibrer mon écoute du Web.* »

Si cette veille permet de filtrer l'information, elle lui ajoute une dimension conversationnelle dans la mesure où elle passe aujourd'hui par les réseaux sociaux comme Twitter ou Facebook : « *Ma journée en ligne se fait pour commencer sur les réseaux sociaux, notamment Twitter qui, ces derniers mois, a pris une partie importante de mon activité. Parce que je trouve que c'est bien fait pour une veille*

scientifique. » Twitter est plébiscité au détriment de la veille standard par flux RSS : avec un flux RSS, « *faut que j'y aille, que je regarde, je fouille* », alors qu'avec Twitter « [les blogueurs] *m'interpellent pour me dire qu'ils ont mis ça sur leur site* ». La rapidité et la brièveté des *tweets* favorisent l'éclatement des usages, retrouvant les réflexes de la discussion informelle. Même s'il peut se composer des identités distinctes (via les pseudonymes), qui discriminent le type d'informations échangées et le type d'interlocuteurs, il devient difficile de distinguer les pratiques professionnelles d'un chercheur sur le Web de ses pratiques privées : « *C'est difficile de dire à un certain moment si je fais de la veille scientifique ou alors tout simplement si je suis en train de twitter avec mes amis* ».

Par effet de retour, cette veille en réseau devient un lieu où le chercheur prend l'habitude de s'exposer lui-même, à l'intérieur de cercles plus ou moins larges et contrôlés : page Facebook, compte Twitter, blog personnel (individuel ou collectif). Un autre visage du chercheur se révèle sur l'Internet, une autre manière de se valoriser : en partageant les résultats de sa propre veille (tout en dissimulant ses sources, pour ne pas perdre l'avantage) ou en assurant la promotion de ses publications non académiques (blogs). D'autres codes ou règles de reconnaissance se développent. Tout d'abord, parce que le chercheur-blogueur garde un côté contrebandier, exerçant souvent « *en cachette* », à l'insu de certains de ses collègues qui ne considèrent pas ce type d'activité comme sérieuse : « *C'est un type d'écriture qui est dévalorisé ; dans notre équipe* [de recherche]*, on ne peut pas en parler. Si on en parle, ça fait toute une histoire* ». Ensuite, parce que les blogs sont des créations souvent très personnelles ou des relevés d'expérience (« *comme un carnet de recherche* »), où le chercheur expérimente de nouvelles manières d'écrire, bien distinctes d'une publication papier : « *On n'écrit pas de la même manière* […]. *Pour moi c'est impossible de faire un livre avec ça, parce que d'abord le blog c'est une forme de parution et d'exposition* […] *qui est particulière. Quand les gens vont* [sur le blog]*, c'est une apparition sur l'écran.* »

➔ Les données collectées sur le Web

Dans le cadre de programmes de recherche portant directement sur le Web, un chercheur distingue trois types de données collectées, en indiquant au passage que ces données n'ont pas vocation à être conservées « *ad eternam* » :

- Les données qualitatives : « *Tout contenu en ligne que vous analysez dans une approche épistémique : comprendre le sens, analyser d'un point de vue lexicographique. Ce peut être des images, des textes.* »
- Les données quantitatives : « *Des données calculées : le temps de connexion, le nombre de contacts, le volume de lecteurs d'un certain site* ».
- Les données relationnelles : « *Troisième grande famille, qui commence à prendre une envergure considérable et qui en plus posent un tas de problème et ouvrent un ensemble de pistes de réflexion ; mais aussi nous met face à des enjeux qui n'existaient pas auparavant* […]*, c'est : qui parle avec qui ? Qui est connecté avec qui ?* ». Pour cela, il faut « *clusteriser, analyser les réseaux, voir quels sont les degrés de densité ou le type de maillage qu'on arrive à créer dans ces réseaux d'amis ».* Mais une telle analyse pose aussitôt des problèmes éthiques : « *Parce que c'est vraiment difficile d'anonymiser des données relationnelles.* »

Indépendamment ou en marge de programmes de recherche où le stockage des données est organisé par le partenaire industriel ou institutionnel, certains chercheurs se sont constitué leurs propres archives de sites Internet. Cet archivage privé se fait par impression papier ou capture d'écran, dans des conditions matérielles souvent difficiles : « *J'ai un bureau envahi par les dossiers d'impressions* », « [j'ai] *pas mal de disques durs, disques externes que je stocke surtout chez moi* ». D'autres chercheurs ont effectué des dépôts dans des fonds conservés par des institutions : un fonds comme Sida-Mémoires (déposé à l'IMEC) inclut ainsi des copies d'écran de sites ou de « pages personnelles » (l'Internet ayant joué un rôle central dans l'histoire de la lutte contre le Sida).

Ces impressions papier et ces copies d'écran sont effectuées dans un sentiment d'urgence et de déperdition, en réaction à la volatilité du Web : « *Vous êtes là en train de capturer des données qui disparaissent, un terrain qui fane, qui s'évapore* ». Mais il s'agit d'un pis-aller. Les chercheurs ayant pratiqué un archivage privé reconnaissent son absence de principe bien établi (« *Je n'ai pas toujours été rigoureux* ») et son peu de validité scientifique : « *Problème : bien sûr, ces pages se mettent à jour. Donc le fait d'avoir imprimé, du point de vue authenticité de la source, ne signifiait strictement rien. Soit avec l'imprimé, soit avec les captures d'écran, on se retrouve avec des contenus statiques, alors que les sites sont censés être mis à jour et avoir des contenus dynamiques.* »

A côté de ces pratiques privées, que l'on pourrait qualifier d'artisanales, certains chercheurs travaillant directement et presque exclusivement sur le Web « *se créent leurs propres archives en pdf et stockent des gigaoctets de données.* » Ces chercheurs « *hyperconnectés* », qui « *attrapent tout le*

temps », ont développé leurs propres outils d'analyse textuelle. Par conséquent, ne les intéresse du Web ou de son archive que ce qu'ils peuvent télécharger sur leurs ordinateurs, pour *« faire mouliner la machine »*. Mais c'est un investissement jugé très coûteux en temps de veille et en matériel.

Pour ceux qui n'ont pas de pratique systématique d'archivage, le recours à Internet Archive est aujourd'hui courant. Si l'archivage du domaine français par la BnF n'était connu que de deux chercheurs, quatre connaissaient le site de la fondation américaine : http://www.archive.org. Certains y ont recours, mais de manière seulement ponctuelle (« *deux ou trois fois l'an. Après, ça peut être intense pour quelques jours* »), en particulier après avoir constaté la disparition d'un site utilisé dans une précédente recherche : « *Il y a des années* […] *j'avais travaillé sur jennycam.com, la fameuse…*[1] […] *l'autre jour j'avais besoin de remobiliser ces matériaux, je suis allé sur l'Internet et il n'y avait plus rien. Donc, voilà, je suis allé en fait sur un site qui fait de l'archivage* ». Un chercheur reconnaît quant à lui avoir abandonné la consultation d'Internet Archive depuis plusieurs années, à cause de la perte trop fréquente des « *contenus dynamiques* » et des « *liens* » dans les archives mises en ligne.

> « *Au début, j'avoue, jusqu'à 2003-2004, la Wayback machine semblait fonctionner assez bien, ou alors peut-être que les sites que je ciblais étaient beaucoup trop simples et n'avaient pas le type de complexité qui s'est manifestée par la suite, quand j'ai même décidé d'abandonner l'usage de Wayback machine, parce que, en gros, j'ai commencé à constater que la plupart des sites avaient beaucoup trop de broken links ou alors de liens morts qui renvoyaient à des entités, à des images, ou alors à d'autres sites qui n'existaient plus.* »

Les chercheurs ayant eu connaissance de l'existence d'une collecte de sites français par la BnF (par le biais d'une information émanant de la BnF ou d'un message relayé au sein de leur établissement de recherche) ne sont pas venus en bibliothèque de Recherche pour les consulter. Pour le moment, la plupart n'en ressentent pas le besoin, même s'ils travaillent sur des domaines incluant une forte dimension Web. La démarche de venir en bibliothèque de Recherche pour consulter ces archives est liée dans leur esprit à des projets de recherche plus spécialisés, liés en particulier à l'histoire du Web : « *Usage ? Pas forcément moi personnellement, mais des étudiants… C'est pas mes sujets de travail. Plus tard, dans une perspective de recherche du genre "Comment les sciences sociales ont investi le Web", potentiellement oui.* » Un seul chercheur interrogé se dit prêt à venir à la Bibliothèque, poussé par la curiosité : « *Je suis très intrigué, donc j'irai voir, ça c'est clair.* ».

➔ Le Web dans les travaux scientifiques : un usage problématique

L'usage ici interrogé ne renvoie pas à la citation d'un texte publié en ligne (pratique aujourd'hui courante chez les chercheurs), mais à la convocation dans un travail scientifique d'un site Internet comme archive : preuve ou illustration d'un phénomène sociologique ou historique.

Un chercheur ayant consulté d'une manière encore informelle des forums de discussion touchant à son objet de recherche reconnaît, modestement, ne pas avoir su ou pu utiliser ces matériaux dans son travail :

> « *Je suis incapable d'avoir la méthodologie qui permettrait d'utiliser de manière scientifiquement valide ce type d'information. Tu as des procédures pour des enquêtes de terrain, t'as les procédures pour des recherches en bibliothèque, mais on n'a pas, en tout cas pour l'instant, et à ma connaissance, de procédure pour savoir quoi faire de ces matériaux. Ce qui serait à mon avis quelque chose à travailler parce que… Comment est-ce que tu sélectionnes et que tu peux justifier de la sélection des forums sur lesquels tu vas te situer ? Ça c'est le problème assez général de l'Internet : il y a tellement d'informations que trier c'est la grande difficulté.* »

Plusieurs chercheurs avouent en effet leur perplexité devant la mention d'un site Internet dans un travail scientifique. Deux exemples sont cités dans des travaux récents d'historiens : Thomas Laqueur, *Le Sexe solitaire. Contribution à l'histoire culturelle de la sexualité* (Gallimard, 2005) et Nicole Eidelman, *Histoire de la voyance et du paranormal, du XVIIIème siècle à nos jours* (Seuil, 2006). Pour actualiser leur recherche, le premier fait référence à des sites de masturbateurs, et la seconde à des sites de voyants. Deux problèmes se posent alors : d'abord le référencement, c'est-à-

[1] 1996, Jenny, une étudiante américaine installe une webcam dans sa chambre et la fait tourner 24/24h sur Internet. L'image en noir et blanc ne se rafraîchit que toutes les 3 minutes mais le spectacle est total. Au plus fort de son succès, Jenny attire entre 3 et 4 millions de personnes par jour. Après 7 ans de direct, jennycam.com s'arrête le 31 décembre 2003. Il n'en reste aujourd'hui plus rien, à part quelques captures.

dire la possibilité de renvoyer à une source vérifiable, surtout si le site a disparu ou changé d'adresse ; ensuite, le principe de sélection : pour que celui-ci soit justifié, il faudrait que l'archive citée puisse être inscrite dans un corpus aux contours maîtrisés et partagés par une communauté de chercheurs.

> « *Là, finalement, les matériaux* [que Thomas Laqueur et Nicole Eidelman] *mobilisaient, il y avait un vrai problème de référencement, un vrai problème aussi de dire : Qu'est-ce que c'est ? Le principe pour un historien, c'est quand même de travailler sur des sources partagées. Le mec qui cite des choses que personne n'a vu et que personne ne peut voir c'est rien, c'est du roman.* […] *Aujourd'hui, une collecte sur l'Internet, elle doit être documentée et elle doit être argumentée, parce que autrement, c'est ce que j'ai ressenti à la lecture du travail de ces collègues historiens qui se mettent à mobiliser des ressources sur le net, tu te dis : "C'est quoi la justification ? Pourquoi vous avez pris ce site et pas cet autre ?"* »

Or, nulle maîtrise actuelle du Web, personnelle ou collective, ne semble permettre cette justification. Les chercheurs habitués de l'Internet reconnaissent que sa bonne connaissance est toujours un leurre. Le bon connaisseur est celui qui reconnaît l'énormité de ses lacunes, du fait d'une organisation des sites en réseaux : « *Ce sont des réseaux souvent très egocentrés.* [Un chercheur] *qui travaille sur les blogs musicaux croyait qu'il avait tout, et on s'est rendu compte qu'on avait 10% de trucs en commun ! Car il y a vraiment un biais : de plus en plus, avec la veille, tu te construis, tu te fabriques ton propre Web, un monde, et puis tout d'un coup tu tombes sur un truc et tu te dis : c'est incroyable que je ne le sache pas ! Quand on n'est pas dans certains cercles on peut complètement ignorer des sites importants.* »

Dans ces conditions, les contours des corpus étudiés demeurent incertains. Pour la connaissance du net art par exemple, il n'y a pas d'autres solutions que de chercher par noms d'artistes sur le Web avec un moteur de recherche traditionnel ; avec la difficulté supplémentaire que les artistes du net art ne sont pas connus et diffusés par les mêmes canaux que l'art traditionnel (les musées, galeries et expositions qui « *assurent une certaine circulation entre des groupes hétérogènes* »). Pour le net art, « *ça tourne quand même sur des cercles fermés* ». Les découvertes se font le plus souvent « *au hasard* » (« *c'est du pif !* ») et la collecte ne peut prétendre à aucune exhaustivité : « *Vous êtes vraiment obligés de tomber sur un nom pratiquement au hasard et de vous dire "Tiens, ce que fait ce gars-là a l'air intéressant, il cite machin, j'y vais, je regarde". Donc, ça veut dire que vous n'êtes jamais sûr d'une certaine exhaustivité.*. »

Si une exhaustivité, même relative, n'est pas imaginable, toute analyse historique en termes de nouveauté ou d'antériorité est du même coup relativisée : difficile de savoir si tel site est véritablement novateur, ou si tel autre n'a pas présenté la même nouveauté trois ans ou trois heures auparavant. Dans ces conditions, la base documentaire n'est pas jugée assez solide pour permettre un travail académiquement contrôlable, car elle repose sur un savoir insuffisamment partagé, encore disséminé dans les souvenirs de quelques personnes. Au point que le Web est considéré par certains enseignants comme peu recommandable, en l'état, pour des doctorants. Il s'agit d'un domaine « *très très dangereux pour des thésards : le risque de passer à côté d'un gros truc que le gars du jury connaît.* […] *Donc c'est pas des territoires que je conseillerais aux jeunes thésards, car c'est quand même des trucs à haut risque* ».

Pour les travaux sur le Web actuel, les chercheurs mettent cependant en place des protocoles pour constituer des corpus aussi représentatifs que possibles :

- Définition d'un périmètre : par exemple, la détermination d'une ère culturelle de recherche (« *une nation, ou alors une communauté linguistique, je pense au Web francophone* »).
- Exploration et observation : « *on a établi un protocole d'observation qui passe par des points d'entrée qui sont évidemment les points d'entrée de tout un chacun : les moteurs de recherche.* » L'enjeu est de déterminer les mots-clés les plus efficaces, qui permettent de déterminer « *un premier noyau de sites* ».
- Sélection et description qualitative : « *en gros, on commence à voir si ces sites sont connectés avec d'autres sites, quel type de lien il y a entre eux* ». Ce lien doit être qualifié, car le renvoi à un site peut être simplement polémique. « *C'est un travail à faire à la main* ». Ce travail permet d'établir une « *cartographie de comment la blogosphère est structurée* », et de comparer entre elles des blogosphères de différentes ères culturelles, nationales.

Malgré ce protocole, dans le cas de communautés volatiles (« *avec des personnes qui cherchent toujours à être à jour, donc suivre la dernière mode, et si la dernière mode est de passer tous sur Tumblr alors on passe tous sur Tumblr* »), les données collectées sont jugées encore « *peu fiables* » et

surtout « *extrêmement éphémères* ». Les chercheurs eux-mêmes reconnaissent la nécessité d'effectuer des collectes rapprochées des sites qui les intéressent : « *tous les trois mois* », propose l'un d'eux.

3.1.2. Perception des archives de l'Internet

Avant de présenter les recommandations particulières des chercheurs en termes de collecte, il nous a semblé important de rassembler dans cette partie leur perception générale des archives de l'Internet : les jugements qu'ils ont exprimés *a priori* sur l'intérêt, la légitimité éthique et la manière dont il conviendrait d'approcher cette question. En ressort la nécessité de faire un effort réflexif préalable pour trouver un modèle adéquat permettant de se représenter rigoureusement cette archive singulière : à quoi pourrait-on la comparer ?

➔ Intérêt d'une mémoire du Web

Les chercheurs interrogés s'accordent tous sur l'intérêt de conserver une mémoire du Web. Comme nous l'avons vu, ils ont pris conscience de la volatilité de l'Internet et se sont souvent constitués leurs propres archives. Il y a une « *vraie fragilité numérique* » : la plupart des chercheurs interrogés ont déjà eu l'expérience d'un site dont la disparition ou la modification a porté préjudice à leur recherche. Les premières manifestations de l'art en réseau (au tout début de l'Internet) semblent aujourd'hui irrémédiablement perdues : n'en subsistent que des descriptions dans des livres, sans possibilité de remonter à la source pour vérifier la pertinence de ce qui est décrit. Beaucoup de sites d'artistes ont disparu, d'autant plus facilement que ces créations avaient la performance ou le happening pour principe :

> « *En gros, ce qui reste, c'est le témoignage des gens qui les ont faits. C'était une forme assez éclatée car le début de l'art en réseau c'était… Parmi les débuts de l'art en réseau, les traces sont très ténues. Il y a eu le fax, des gens qui se faxaient des messages à travers la planète, chacun se rajoutait quelque chose sur un fax. Y'a des artistes qui ont fait faire plusieurs fois le tour du monde à des fax. Et les premiers travaux Internet n'étaient pas des sites Web, puisque l'Internet est quand même beaucoup plus ancien que le Web, et donc c'était sous des formes proches de l'ASCII art*[2]*, en utilisant simplement des imprimantes. Il y avait aussi des forums, des news groups, qui regroupaient des gens pour faire des choses en réseau. Je pense que personne n'a pris la peine d'archiver ces newsgroups, qui ont vraiment disparu* ».

Cette volatilité native est renforcé pour certaines zones du Web : communautés jugées indésirables par les fournisseurs d'accès (sites pro-aneroxiques, qui disparaissent et réapparaissent tous les mois), communautés très jeunes (promptes à suivre la dernière mode, à migrer très facilement de « WebRing » à « Tumblr », etc.) ou encore activités liées à une certaine forme de marginalité ou d'illégalisme (les sites artistiques de l'Undernet). Les chercheurs-blogueurs partagent cette inquiétude pour leur propre production : malgré les sauvegardes, l'achat de nom de domaine, etc., « *Il y a quand même une espèce de crainte de la disparition ; on voit des choses disparaître* ». Le monde des blogs est considéré comme particulièrement fragile, étant souvent le fait d'amateurs, de passionnés qui n'ont pas un temps indéfini ni des moyens conséquents à consacrer à leur passion : « *Il y a un site qui est un super site de veille un peu rigolo : "Bienbienbien.net"*[3]*. C'est des gens très jeunes qui font ça, donc à un moment ils n'en peuvent plus, donc ils arrêtent. Cet arrêt-là…, est-ce que les gens laissent* [le site en ligne] *? Et jusqu'à quand ? Est-ce qu'à un moment ils arrêtent de payer ? Là, que ce soit archivé c'est typiquement très très bien* ».

La disparition de contenus en ligne peut être également voulue par leurs auteurs eux-mêmes, en particulier pour les sites institutionnels souhaitant effacer les traces de leurs évolutions, pour afficher un discours immuable. Cette fragilité du Web donne à son archive une « *force politique* », avec exemple à l'appui d'une agence nationale française épinglée par un chercheur grâce à Internet Archive :

> « [Un chercheur m'a parlé d'Internet Archive] *comme d'un truc complètement fou, qui donnait à voir aussi, notamment dans des pages qui seraient censées être stabilisées, officielles, des mouvements, des recouvrements de textes, de mots qui étaient transformés. Je vais prendre un exemple très simple, c'est* [une agence nationale] *qui du jour au lendemain a changé*

[2] L'ASCII art (années 1960-1980) consiste à réaliser des images uniquement à l'aide des lettres et caractères spéciaux contenus dans le code ASCII.
[3] http://bienbienbien.net, site toujours en ligne, mais annoncé comme étant « en hibernation ».

complètement son vocabulaire. Et donc [les chercheurs] *ont pu remonter comme cela, pour vraiment faire l'étude d'un dispositif* […] *qui a changé d'un coup, qui n'a pas prévenu qu'il changeait, qui fait comme si rien ne s'était passé. Eux, ça les a sidérés. Là je me suis rendu compte de la force, en fait de la fragilité du Web et donc de la force d'une archive, la force éminemment politique d'une archive, dans un domaine où tu pourrais n'avoir accès qu'à la dernière version tout le temps.* »

L'archive permet de relativiser cette « *dernière version* » en ligne, que les étudiants prennent encore trop souvent pour une version pérenne : « *L'autre utilité de l'archive du Web, c'est de relativiser ce que tu as sous les yeux à l'instant* t. *Il y a deux ans c'était complètement différent, car c'est des objets super mouvants, qui se déplacent sans arrêt.* » L'archive est un « *moyen de mettre de la distance, comme toute profondeur historique* ».

Cette notion d'archive du Web est cependant troublée par d'autres représentations concurrentes et trompeuses : « *Sur l'Internet, il y a une confusion très forte sur les sites, puisqu'on a toujours la rubrique "archives" dans un site,* [c'est-à-dire] *l'idée qu'on peut aller voir l'archive, alors que l'archive ce n'est que ce qui n'est plus en première page* ». Comme le vérifient les entretiens avec des professionnels, plus éloignés que les chercheurs de la problématique de l'archive : le Web invisible et l'historique des discussions (Wikipédia) sont d'autres représentations erronées d'une mémoire du Web. L'Internet donne à l'internaute l'illusion d'être sa propre archive. A l'opposé, aux yeux de l'historien, il s'agit de conserver la trace d'un état antérieur où les contenus sont inséparables de leur « *surface d'inscription* » : « *le contenu est très lié à l'architecture* [du site] ». L'archive d'un site est donc aussi son « *plan* », qui doit être considéré en lui-même comme « *une forme d'écriture* » à part entière, conditionnant le reste.

➔ Questions éthiques posées par l'archivage du Web

La légitimité de l'archivage est jugée différemment selon le type de pages archivées, amenant très vite des questions sur le respect de la vie privée et la signification du terme « public ».

L'archivage est jugé évident, et même nécessaire d'un point de vue citoyen, pour le **Web institutionnel**, regroupant tous les sites qui « *s'appuient sur une forme de stabilité apparente* » et ont eux-mêmes une activité de publication en ligne (sites des ministères, des administrations et établissements publics, du CNRS, etc.). Au niveau du vocabulaire employé, il est intéressant de noter que les sites d'institutions dépendant d'un ministère (le site du CNRS pour le ministère de l'Enseignement supérieur et de la recherche) sont considérés comme des « *publications* » de ce Ministère, donc des « *publications de l'Etat* », alors que d'autres domaines sur l'Internet se verront plus loin contester cette appellation de « *publications* ». Le Web institutionnel apparaît en tête de toutes les propositions de collecte, sans doute parce qu'il résiste le mieux aux aspects les plus mouvants du Web (il semble donc plus facile à capter, à délimiter), mais aussi parce que cette « *stabilité apparente* » est justement ce que l'archivage doit permettre de critiquer : montrer comment les discours officiels ont évolué, se sont contredits ou ont été corrigés sans prévenir. La notion de « *droit à l'oubli* » ne vaut pas lorsqu'il y a un débat public (un « *sujet politiquement en plein débat* »), et l'archivage des sites publics permet d'exercer une surveillance « *citoyenne* ».

Les **blogs** peuvent également être archivés car ils sont créés pour laisser une trace et produisent déjà leurs propres archives (via l'entrée calendaire) : « *Le blog, c'est un flux pensé avec la possibilité de revenir en arrière* ». Les chercheurs-blogueurs ne voient eux-mêmes aucune difficulté à être archivés, bien au contraire, puisqu'ils ont eux-mêmes ce souci de redonner de la visibilité à leurs billets les plus anciens : « [Notre blog] *est déjà pensé comme un carnet de recherche, comme un truc qui s'accumule, dont on se dit : "C'est dommage que les archives ne remontent pas". Alors on les fait remonter sur Facebook, on met le premier post, etc. Moi je trouve ça très bien que ce soit archivé.* » Sur le plan de l'intérêt documentaire, les blogs sont aujourd'hui considérés comme une forme d'écriture singulière méritant dans bien des cas l'archivage (« *ça peut être très intéressant* »). Non seulement les chercheurs ont tous une sélection de blogs qu'ils suivent régulièrement, mais quatre sur cinq sont eux-mêmes des blogueurs actifs, réguliers.

Par contre, l'archivage d'un outil comme **Twitter** pose question, car il fait quitter le domaine de la publication pérenne pour celui informel de la conversation : « *Est-ce que ça doit valoir comme trace ? Je connais des gens qui sur Twitter effacent systématiquement ; il y a des outils pour ça. Ils restent à 100 tweets, alors qu'ils en ont produit 5000. Parce qu'ils investissent ce discours-là non pas comme un discours mais comme un mode de conversation, et il est hors de question de laisser des traces. Dans des mondes comme les nôtres si dans cinq ans* [tel chercheur qui twitte] *est président d'université ! Si on ressort un truc qu'il a dit comme une vanne, hors contexte, c'est très risqué.* » La

notion de conversation renvoie nécessairement à un cercle restreint, relativement contrôlé par le locuteur, même si ceux qui passent par là peuvent tendre l'oreille. C'est toute la difficulté de qualifier cet « *espèce d'espace public restreint* » que dessinent Facebook et Twitter, « *où on dit des trucs en public, on écrit des trucs en public, mais avec un public de gens qu'on contrôle un peu et avec une confiance* [en] *cet espèce d'espace public restreint. Là, s'il y a des traces de ça, c'est compliqué.* »

Lorsqu'il est investi dans ces réseaux sociaux, le chercheur peut d'ailleurs avoir plusieurs identités (pseudonymes) qu'il ne souhaite pas voir croiser ou fusionner : « *J'ai vraiment deux blogs, deux identités Twitter, complètement différentes. Certains savent que je suis sociologue, mais les sociologues ne savent pas que je fais le reste et je vois la force de l'Internet pour ça. Au nom de quoi on archiverait ça pour le consolider ?* » La force des réseaux sociaux sur l'Internet est justement de reproduire, grâce à l'usage des pseudonymes, ce que permet la vie réelle : « *On a plusieurs identités projetées et on ne se comporte pas du tout de la même manière en famille qu'auprès des amis proches, qui ne savent pas trop ce qu'on fait au boulot, et les gens du boulot.* »

Plus généralement, ce qui semble impropre à l'archivage sur l'Internet renvoie toujours à des actions individuelles : converser, se promener, acheter quelque chose. Aujourd'hui, ces actions se font sur l'Internet comme elles se faisaient auparavant dans la rue, dans un magasin, etc. L'Internet n'est plus seulement un lieu de publication, « *c'est devenu quasiment la réalité* » (au moins un « *degré de réalité* ») ; et l'utopie de son enregistrement automatique fournirait « *une archive de l'ensemble du réel* ». D'où l'assimilation de la recherche dans les archives de l'Internet à une forme d'archéologie : « *Ce sera passionnant, c'est évident. Un archéologue du temps futur, au lieu d'aller travailler dans des strates de terre, il aura intérêt à aller travailler sur Internet et sur les archives Internet, car il y trouvera tout ce qu'un archéologue peut y chercher dans la terre : des petits bouts, des petites choses, ça c'est très bien* ». Mais le même interlocuteur ajoute en contrepoint que ce serait « *terrifiant, tout à fait terrifiant* », « *un projet presque "borgésien"* ». Même s'ils sont en accès libre, des pans entiers de l'Internet ne relèvent pas d'ensembles documentaires mis à la disposition d'un public, mais de traces laissées par des individus venus « *faire un certain nombre de choses* » en ligne :

> « *L'Internet est-ce que c'est vraiment du domaine public*[4] *au sens où l'on entend qu'un ouvrage, une œuvre est dans le domaine public, j'ai du mal à accepter ce principe. Lorsque tu te balades dans la rue, est-ce que tu te publies ? La réponse est, me semble-t-il : non. D'ailleurs tu as un droit à ton image. Quelqu'un vient te prendre en photo et archiver les photos, il pourrait à mon avis avoir des problèmes. Et il me semble que l'Internet ça ressemble plus à ça.* […] *Est-ce qu'il y a une si grande différence de ce point de vue-là : je dois aller à la boulangerie pour acheter une baguette de pain, je dois aller sur l'Internet pour faire un certain nombre de choses. Une fois de plus, l'Internet, c'est aujourd'hui une réalité du même ordre. Comme tu laisses des traces dans la rue, ou dans un magasin, est-ce qu'on va imaginer que demain on va archiver l'ensemble des données des caméras de surveillance des magasins ? L'argument est le même : on rentre dans un lieu, pour faire un certain nombre de choses.* »

Inquiet de la réutilisation malveillante des données stockées, un chercheur insiste pour que soit mise en place à la Bibliothèque « *une accréditation ciblée, et surtout qu'on ait un contrôle sur la réutilisation de ce qu'on va faire de ça, parce que l'accréditation est seulement en entrée* ». Le modèle cité en exemple est celui des bases de données d'enquêtes dans certains laboratoires ou institutions scientifiques qui font l'objet de traçage systématique, en taguant toute exportation : « *Si je télécharge un ensemble de données, il y aura la possibilité d'introduire des métadonnées qui peuvent permettre de suivre où ils ont été copiés-collés, remis en ligne ou exploités dans des tables, des graphes. Techniquement c'est possible* ». Au-delà du contexte juridique particulier du dépôt légal, est également évoqué un devoir, à plus ou moins long terme, de « *restitution des données* » à ceux qui les ont produites, en particulier face au risque de remise dans le domaine public des informations stockées par les institutions (prônée aujourd'hui par des mouvements d'opinion).

➔ Manière de concevoir cette archive et d'en définir (ou non) les unités

La notion d'archive est mise en regard et parfois opposée à la notion de « *flux* » : « *Comment est-ce qu'on archive le flux du temps ?* » se demande un chercheur. Par définition, ce qui se donne comme flux continu ne connaît pas d'unité, et seul un geste extérieur peut y découper des « *séquences* », des « *instantanés* », des « *photogrammes* » :

[4] L'article L131-2 du code du patrimoine parle de « mise à la disposition d'un public ». Le terme « domaine public » est ici impropre.

« *L'Internet, d'une manière ou d'une autre, ça fonctionne par flux continu. Donc là, on se retrouve devant un autre problème, philosophiquement très intéressant, c'est qu'on ne peut pas archiver du flux. On n'archive que des séquences d'une manière ou d'une autre, et je ne vois pas comment on peut envisager comme résultat d'archiver ce qui par essence est fluctuant, flux, ça c'est quelque chose qui me paraît difficile à imaginer.* »

Ce caractère fluctuant, encore exceptionnel il y a quelques années (avec des contenus statiques, mis à jour à intervalles espacés), devient aujourd'hui la règle : les sites de rencontres, par exemple, offrent aux internautes la possibilité de s'exposer en temps réel, « *par flux* ». Le phénomène récent du « Chatroulette »[5] laisse présager ce que pourrait être le Web de demain et la difficulté de mettre en place une stratégie d'archivage qui ne soit pas prise de vitesse par ces mutations. « *Désormais, une page Web est une partie infime de contenus statiques : c'est peut-être seulement le titre, et sinon pour le reste vous avez des images qui défilent, des galeries qui passent, des RSS qui déroulent tout le temps.* » A l'heure du *cloud computing*, il est possible qu'« *on puisse être vite dépassé par le type de nouveaux services et le type de phénomène d'émergence : soit sociale, phénomènes sociaux, soit émergence de contenus* ».

L'archive traditionnelle papier, dont on peut définir de manière stable les unités, est mentionnée à plusieurs reprises comme modèle repoussoir pour décrire l'archive de l'Internet. Il ne faut pas considérer cette archive « *comme une archive du XIX^e^ siècle, avec un corpus qui est stable, un instantané, figé dans le temps. Cette archive ne peut pas vivre si elle n'est pas une archive vivante, dynamique* ». Même le « site » n'est pas forcément l'unité documentaire pertinente (« *ce n'est pas en termes de sites que je pensais* »), étant donné qu'un site se définit également par le réseau dans lequel il s'inscrit.

Il ne faut donc pas tenter de transposer d'anciens modèles d'archivage. Il faut plutôt aller chercher du côté d'autres pratiques scientifiques des modèles nouveaux permettant de parler de cette archive absolument singulière. Ainsi, par exemple, de l'étude des traditions orales, qui remet en question la notion même d'archive :

« *La grosse erreur, ce serait de considérer que le Web c'est l'équivalent des paroisses du Moyen-Âge et que je vais récolter tout ça. C'est un processus dynamique qui continue toujours. Or, en essayant d'éliminer la dimension temporelle, on va être largement à côté. Avec une vision qui change beaucoup de ce qu'est l'archive. Le vrai parallèle du Web… Le Web est de l'ordre de l'oralité, on est dans la tradition orale. Il faut que les gens du livre oublient un peu leur tradition documentaire et se disent : "On est dans les sociétés de tradition orale". Le vrai parallèle pour moi, il est là. Y'a pas de surface d'inscription définitive. Parce que la surface d'inscription n'existe que dans la mesure où l'appareil pour l'utiliser existe – donc, en ça, on est très proche du fonctionnement de la mémoire humaine –, ce qui se perd paraît échapper à toute raison, etc. Il faut se ressaisir de ce qu'était les traditions orales pour les chercheurs en sciences humaines, pour mieux comprendre comment fonctionne cette mémoire. Donc on est loin de l'archive de ce côté-là. Est-ce que c'est transformable en archive ? Mon opinion empirique tout à fait personnelle : je dirais "non". Ceux qui archivent doivent repenser l'archive sur ce territoire-là.* »

3.1.3. Contenus

➔ Politique documentaire : ne pas redouter la sélection

Malgré l'importance de conserver les liens d'un site vers d'autres sites, viendra un moment où l'archiviste devra « couper » dans la toile : « *On est dans la toile infinie, il y a un moment où il faudra couper. Il suffit juste de décrire les critères qui ont donné lieu à coupure.* » Une telle coupure n'est pas seulement une contrainte de fait (parce qu'on ne peut pas tout prendre), mais pour certains chercheurs, ce geste négatif a aussi une fonction positive. Par définition, le non-archivage est constitutif de l'archive :

« *D'une manière générale, je trouve que l'Internet ne doit pas neutraliser l'acte d'archivage qui est d'abord un acte de non-archivage, puisque que fait l'archiviste ? D'abord il détruit, d'abord il trie et il détruit* […]*, je trouve qu'on a tort de pas le dire. À un moment donné, il y a deux aspects :*

[5] Site où les internautes s'exposent publiquement (webcam) dans des fenêtres qui peuvent apparaître de manière complètement aléatoire : « *tu cliques, tu zappes, tu cliques, tu zappes, tu tombes sur des gens différents tout le temps* ».

l'archiviste, oui il conserve, mais il détruit d'abord ; il suffit de voir les bennes et le matériau détruit, de voir les dispositifs de destruction extrêmement intéressant des archives municipales. »

Ne pas choisir serait d'autant plus impensable que l'Internet met l'archiviste face à une situation inédite : un ensemble qui se donne comme une « *nappe absolument horizontale* », indifférenciée, abolissant les hiérarchies. Auparavant, même le dépôt légal des imprimés ne fonctionnait que sur fond d'un ensemble de sélections préalables, que l'automatisme apparent du dépôt nous faisait oublier : sélection par les éditeurs, par les moyens matériels nécessaires à l'édition, par les validations multiples, etc. Or, « *la particularité de l'Internet, ça a été d'effacer la plupart des procédures de sélection* ». Le témoignage d'un chercheur-blogueur parlant du lancement d'un blog collectif vérifie cette immédiateté : « *Quand* [X] *et* [Y] *m'ont parlé d'un blog, j'ai envoyé le premier papier. J'ai pas répondu "oui", j'ai envoyé un premier papier et "tac" !* [c'était en ligne] ». L'Internet met ainsi à l'épreuve notre difficulté actuelle à poser des choix en termes de valeur, et nous contraint à nous réinterroger sur ce qu'il faut garder et ce qu'il ne faut pas garder.

Nulle crainte, donc, exprimée par les chercheurs de ne pas « tout » avoir, à partir du moment où les choix faits par l'archiviste sont clairs et accessibles : « *La question que pose l'Internet et que les archives nous posent, c'est à un moment de dire :* […] *on le prend où on le prend pas ? On le prend pas, y'a pas de problème, c'est pas grave* […] *mais au moins on le dit* ». L'exhaustivité de l'archive n'existe en effet que très rarement pour l'historien, qui s'en accommode souvent très bien. Travaillant sur une activité commerciale au XX^e^ siècle, un chercheur raconte ainsi sa visite aux Archives nationales du monde du travail : « *Là, l'intérêt des Archives du travail, c'est qu'il y avait les archives d'une entreprise. Je n'en ai pas besoin de cinquante. Mais en même temps, si j'en avais cinquante, ça permettrait de faire du travail. En même temps, bon, est-ce que c'est… ? Pour moi, ça m'intéresse d'avoir une entreprise pour voir comment ça fonctionne.* »

Les chercheurs restent en fait partagés entre la reconnaissance que tout peut être *a priori* intéressant pour les historiens futurs, même les objets les plus futiles ou les plus ténus, et la conscience du rôle structurant pour l'archive des choix opérés : il faut qu'il y ait du tri et il faut qu'il y ait de l'oubli. Avant même toute référence à un « *droit à l'oubli* », il y a un « *oubli* de facto », qui est qualifié d'« *oubli heureux* » : des choses se perdent, disparaissent, etc.

La légitimité d'opérer des choix clairement formulés rejoint par ailleurs les craintes exprimées par un chercheur d'une « *archive complètement automatisée* ». Le mot « *automatique* » est banni, car il ouvrirait la possibilité d'un traitement pareillement automatique de l'archive, en particulier des données personnelles qu'elle inclut. Une seule personne interrogée souligne l'intérêt d'opérer, en plus d'une collecte ciblée, « *une sorte de sélection relativement indifférenciée à l'instant t* », « *un panorama relativement transversal* », insistant au passage sur l'importance de faire durer l'instant t pour avoir un peu d'épaisseur temporelle (et capter des phénomènes comme l'exposition en flux).

➔ Quelques pistes : par nœuds, par thème, par rupture, par pratique

Si le contenu des collectes a été peu abordé dans le détail, faute d'une connaissance de son cadre technique actuel, quelques pistes ont été naturellement mentionnées :

- Une recherche préalable devrait permettre de déterminer « *où sont les nœuds du réseau ?* », afin de déterminer les points nodaux de la toile, si nécessaire à l'aide des modèles mathématiques adéquats, bien connus des Télécoms. Comme on ne peut pas tout prendre, l'audience est considérée comme un mode de sélection parmi d'autres, afin d'« *avoir l'essentiel de ce qui préoccupait à l'époque au temps t* », « *les sites les plus visités, les plus en réseau* ».
- Même si tout est susceptible d'intéresser les générations à venir d'historiens, un « *échantillon* », avec une « *entrée thématique* », « *par types d'activité* », par « *types de sites* » peut suffire (cf. l'exemple cité plus haut des archives d'une entreprise commerciale).
- Une collecte ciblée devrait également se concentrer sur ce qui fait vraiment rupture, avec « des *programmes de recherche qui se posent la question de la nouveauté* » : les « *pratiques qui renouvellent* [un] *genre* » sur l'Internet. Prenant l'exemple des sites pornographiques, un enquêté note en effet : « *ça sert à rien de donner des descriptions précises de choses qui finalement sont assez anciennes, et quand on a fait l'état des lieux une fois, ça suffit* ». Par contre, « *la grande innovation sur l'Internet aujourd'hui* », ce sont les sites comme « *Chatroulette* » : « *S'ils sont dans le domaine français, il faut les archiver. Ça, c'est pas possible de pas les archiver. Parce que ce sont vraiment des sites fondamentaux.* ». Mais prendre une séquence (un jour, ou plusieurs moments de la journée) sur quelques sites devrait suffire : « *Honnêtement, c'est tellement répétitif, qu'il n'y a pas besoin que ça dure très longtemps* », « *à*

partir du moment où tu as archivé quelques sites comme ça, tu peux pratiquement t'arrêter ». Il suffit de « *mettre à la disposition des gens le fait que ces choses-là existent* ».

Devant la singularité du Web et la difficulté de lui appliquer un modèle traditionnel de collecte, un chercheur propose de considérer l'Internet également comme une pratique, et d'archiver des traces de cette pratique considérée comme un certain « *ordre du discours* », dicté par les moteurs de recherche. Par exemple, les listes de résultats obtenus sur Google en tapant un certain nombre de mots-clés intéresseraient le chercheur :

> « *Parce qu'au fond qu'est-ce que c'est que l'Internet ? L'Internet c'est quelque chose qui n'existe qu'à partir du moment où tu mets sur Google quelque chose. Tu rentres par les moteurs de recherche.* [C'est] *en tout cas l'usage le plus courant ; plus personne, franchement, n'écrit une adresse www. Donc,* […] *ce qui m'intéresserait du point de vue vraiment du chercheur, ce serait de connaître au fond – pour reprendre un terme foucaldien – un "ordre du discours". Quand on tape "Foucault", qu'est-ce qui sort ?* […] *Mais il faudrait pas prendre "Foucault", il faudrait prendre les termes "Démocratie", "Écriture". Vous avez des thésaurus ? Prendre le thésaurus, et voir ce qui apparaît à un moment ; et donc archiver le* [résultat]. […] *Ce qui m'intéresse là, c'est l'entrée "pratique" en fait. L'Internet, c'est d'abord une pratique. Donc : archivons la pratique, parce que sinon on va perdre la pratique.* »

Enfin, une question a émergé sur l'articulation entre politique documentaire et dépôt légal. En effet, le dépôt légal traditionnel, qui est d'abord une collecte non sélective, ne connaît pas ce geste de destruction, constitutif de l'archive. A ce titre, l'expression « archives de l'Internet » sous l'égide du dépôt légal surprend : « *Vous n'êtes pas les archives nationales, et c'est là qu'il y a une vraie question. Le terme d'"archive" n'est pas tout à fait juste, pour moi.* » Mais si le cadre demeure le dépôt légal, certains demandent alors pourquoi la Bibliothèque cherche par elle-même à capturer les sites au lieu d'inciter leurs créateurs à les déposer au même titre que les éditeurs : « *essayer de faire que des gens qui considèrent avoir créé quelque chose d'intéressant sur le net aient une certaine propension à déposer par eux-mêmes* ». Il s'agirait non seulement d'un moyen de collecte parmi d'autres, mais également d'un principe de sélection : l'acte de dépôt, dans le domaine Internet, pourrait fournir des résultats « *en soi* » intéressants. « *Je pense qu'il y a un intérêt en soi à la démarche du dépôt. Les gens qui l'ont fait : c'est une attitude par rapport à leur propre travail qui, dans les champs qui m'intéressent, est intéressante.* ».

3.1.4. Outils et services

➔ Documenter la collecte

Compte tenu des problèmes posés par l'usage du Web dans un travail scientifique (cf. 3.2.1.), le fait qu'un établissement public comme la BnF assure l'archivage de l'Internet est jugé positivement. Le statut de la Bibliothèque, la transparence de ses procédures et l'accessibilité de ses collections offrent au chercheur une garantie minimale contre une collecte anarchique, dont les modalités sont invérifiables : « *Le fait de s'abriter derrière une institution, type BnF, qui ferait ce travail, me permettrait de renvoyer à une cote, ou à quelque chose* ». La demande d'une « *côte* » exprime ce besoin d'authentification par un tiers reconnu, au-delà de la mention aujourd'hui habituelle d'une URL et d'une date de consultation dans les articles scientifiques.

Au-delà de la côte, il faut pouvoir répondre aux questions : « *Comment caractériser ces collectes ? Qui a archivé ? Quel est cet état* [de l'archive] ? ». Parce qu'elles ne donnent pas de réponse à ces questions, les collectes privées demeurent problématiques pour le chercheur. Elles induisent des biais personnels considérables que le recours à des instances publiques permet de réduire. Sont décisifs pour le chercheur la définition et l'affichage d'une politique claire de collecte : « *Il doit y avoir une note afin de définir la collecte dans laquelle ça s'est fait, c'est-à-dire pourquoi vous avez collecté "Jenny"*[6] *et dans quel cadre. »*. Les critères de sélection doivent être eux-mêmes archivés, car ils sont appelés à évoluer.

L'« *État* », les « *bibliothèques publiques* », les bibliothèques nationales ou cinémathèques sont ici considérés positivement comme des instances de préservation, avec une valeur ajoutée de transparence. Ironie de l'histoire, c'est par exemple l'Etat qui préserve les productions audiovisuelles gauchistes et anarchistes des années 1960-1970, réalisés par des militants qui ne se souciaient pas de la question de la mémoire : « *Donc, pour moi, c'est très nécessaire qu'il y ait des trucs publics* [qui

[6] Voir note 5.

archivent] *parce que pour moi les trucs un peu créatifs, hors normes, n'ont pas tendance à développer des archives.* »

➔ Décrire l'archive

Cette requête de description révèle chez le chercheur un besoin de « *hiérarchisation* », de « *discriminations* » entre les sites collectés : « *Le problème de l'archive du Web serait de rendre tout à plat* » ; « *Si t'essayes d'établir une archive indifférenciée de l'ensemble de l'Internet, en fait tu ne fais que reproduire une forme d'idéologie typique de l'Internet qui laisse entendre qu'il y a une forme d'indifférenciation totale, alors que dans la réalité, il est bien évident que chaque site, chaque institution qui a un site opère des sélections, fait des choix* » (les sites vers lesquels il pointe, les sites qu'il abrite, etc.). Il convient de nommer de grandes ensembles au sein de l'archive de l'Internet.

Une typologie documentaire est cependant difficile à déterminer : un chercheur qui avait commencé à classer les sites par grandes catégories (les institutions gouvernementales, les institutions non-gouvernementales, les entreprises, etc.), s'est rendu compte que la liste était « *infinie* » et qu'une telle énumération n'était pas pertinente. Un autre propose « *de décrire des archives du quotidien, puis les archives de diverses institutions, etc.* » ; mais le problème est qu'une archive du quotidien (un blog) peut être abritée par un site institutionnel, etc. Le risque est d'appliquer une pratique documentaire traditionnelle à un ensemble d'objets dont il convient d'abord de décrire « *les lignées* », « *les maillons* », les « *hiérarchies* », qui déjouent les partages traditionnels. D'où le parallélisme proposé par un chercheur avec l'histoire naturelle. Ce recours récurrent à des champs lexicaux éloignés du vocabulaire de la bibliothéconomie (cf. la mention des traditions orales en 3.2.2.) indique la nécessité de renouveler non seulement les manières d'archiver, mais les manières de parler de l'archive :

> « *On rejoint l'un des plus vieux débats de l'histoire naturelle et des classifications.* […] *Et au fond finalement, je me demande si, plutôt que de comparer ça à des archives, je me demande si la meilleure analogie ce ne serait pas précisément celle-là : c'est-à-dire celle des classifications des êtres naturels, telles qu'elles ont pu se mettre en place. Parce que je pense que l'Internet ça relève plus, à mon sens, beaucoup moins de la publication et des sources, que finalement d'un degré de réalité ; et que finalement on se confronte plus à un problème assez équivalent à celui des naturalistes confrontés à la volonté d'établir des classements, de hiérarchiser, et de conserver, et de nommer, situer l'ensemble des êtres naturels. Et alors tu te retrouves évidemment face au débat le plus classique : est-ce que tu fais le choix d'un système artificiel, évidemment avec son artificialité, mais qui permet un classement commode qui permettra à tous et aux chercheurs de se repérer, même si elles sont artificielles ("archives du quotidien", "archives institutionnelles", c'est particulièrement artificiel), ou bien tu fais le choix, un peu à la Buffon*[7]*, d'établir des généalogies, des liens réels entre des lignées de sites et des rapports d'engendrements quasiment entre sites. Des dérivations, etc.* »

Si l'on entre dans le détail des « données », sont demandés par un chercheur :

- L'URL et la date de capture ;
- La place de la page capturée dans le site et l'arborescence ;
- Les sites qui pointaient vers cette adresse et les sites pointés par cette adresse (information qui n'est pas nécessairement associée à la possibilité de *surfer* sur une archive). Ce qui intéresse le chercheur n'est pas tant le site, que son inscription dans un ensemble de sites : les « liens » (qui sont justement les données jugées défectueuses sur Internet Archive) : « *Il est bien évident qu'une chose est très importante, c'est la possibilité de rendre compte non pas de l'existence du site mais de l'existence du site dans un réseau de sites, et des renvois entre un site et un autre site, et tout ce système là, qui est un système très particulier parce qu'il opère un certain nombre de distinctions. Ça n'est évidemment pas la même chose d'avoir un blog qui est rattaché à Libération que d'avoir un blog qui est rattaché à La Croix.* […] *les blogs, les différents sites se rattachent à une nébuleuse de sites, et cette nébuleuse il faut en rendre compte, car c'est celle-là qui est en réalité la plus intéressante* » ;
- Des statistiques de vues, « *une forme d'audience : est-ce que ce truc était important ou pas, ça arrivait en combientième position de Google, etc. ?* ». Une bonne description doit en effet rendre utilisable la collecte en évitant de tout mettre sur le même plan, en particulier au niveau de la notoriété et de la fréquentation des sites. Avoir une idée de l'audience est revenu plusieurs fois : « *Les sites sont souvent des microcosmes et si tu mets pas les statistiques et le genre de*

[7] Buffon (1807-1888), naturaliste français.

public qui va avec ce serait vraiment biaisé ». Autant un historien sait faire la différence entre « *Le Monde* » et « *le petit journal du coin* », même cinquante ans après, autant cette différence est difficile à établir sur l'Internet, car il s'agit d'un univers fait de « *niches* » et de « *microcosmes* », où les gens « *s'inter-citent* ». Ainsi du blog http://bienbienbien.net : « *C'est pas connu partout et, par contre, pour ceux qui s'intéressent au Web et à la culture Web, c'était incontournable* ».

➔ Collaborations et communautés

La toile est « *infinie* » et nul ne peut prétendre en faire le tour. Ce truisme n'est pourtant pas la chose la mieux partagée du monde, et l'illusion de bien connaître le Web dans son domaine prévaut encore chez certains étudiants ou chercheurs. Même les pratiques de veille décrites plus haut développent des réseaux très « *égocentrés* » qui deviennent autant de filtres à la connaissance du Web :

> « *Quand on n'est pas dans certains cercles, on peut complètement ignorer des sites importants* […] *Car il y a vraiment un biais : de plus en plus, avec la veille, tu te construis, tu te fabriques ton propre Web, un monde, et puis tout d'un coup, tu tombes sur un truc et tu te dis : "C'est incroyable que je le sache pas !"* »

A cela s'ajoutent les zones obscures ou cachées de l'Internet : les sites qui ne sont volontairement pas référencés de manière classique pour rester discrets (en particulier dans le domaine de l'art underground), l'Undernet ou encore la « *zone grise* » :

> « *Si vous commencez à vous balader sur le net, il y a des îlots perdus. Vous pouvez surfer tant que vous voulez dans ce monde-là, si vous restez dans ce monde-là, vous n'accéderez jamais à celui-là. Ça, c'est le net dominant : si vous êtes dans ce monde-là, il faut que quelqu'un vous donne une première référence pour pénétrer dans l'îlot. C'est ce qu'ils appellent la zone grise, c'est des modèles théoriques qui ont été développés dans les années 90, pour essayer de décrire l'Internet.* »

Pour cette raison, un groupe d'experts qui recenserait les sites importants dans la plupart des domaines laisse dubitatif. Il est conseillé une démarche plus simple et plus ouverte : « *Plutôt que d'avoir un ou deux conseillers, avoir plus de personnes et faire un truc très simple : envoyer un questionnaire rapide à plus de monde. "Les cinq flux que vous suivez le plus en ce moment". Un questionnaire tous les six mois.* » Il ne s'agit donc pas de sélectionner, de porter un jugement (les meilleurs sites, les sites les plus importants), mais d'enregistrer simplement des pratiques : « *Qu'est-ce que tu utilises ?* »

Pour aider aux collectes ciblées, un chercheur parle aussitôt de « *programmes de recherche* ». Cette mention souligne le fait qu'il n'y a pas de collecte simple sur l'Internet, que toute collecte requiert de l'exploration et de l'observation, menées sur la durée et collectivement.

En aval de la collecte, la collaboration avec des chercheurs doit permettre la mise en place d'outils de travail leur permettant de se repérer au sein des sites collectés. L'expression proposée à un chercheur de « Guide des sources » est reconnue comme étant pertinente. Un chercheur évoque la formation de « *communautés de recherche* », la mise en place de « *portails spécialisés* », « *où les gens se passent des tuyaux* », pour que le savoir puisse être partagé. C'est en comparant leurs sources sur un domaine précis du Web dont ils sont spécialistes que deux chercheurs ont découvert qu'ils avaient « *10% de trucs en commun* ». « *Votre fonds sera réellement exploité si émergent des portails de gens qui partagent les mêmes préoccupations* ». Un chercheur s'intéressant au net art souligne le fait que les sites d'artistes sont une catégorie « *très difficile à traquer* », et les moteurs de recherche ne sont dans ce domaine que de peu d'utilité car ils ne permettent pas de qualifier les résultats qui remontent. Il faut donc « *un travail d'indexation à la main* », pour rendre un fonds d'archive utilisable dans le domaine artistique – un travail nécessairement qualitatif et collectif de description, et donc de hiérarchisation (« *ça, c'est intéressant* », « *le plus intéressant* »).

> « *C'est-à-dire : que des gens repèrent les sites artistiques, qu'un portail se développe, et que les gens disent : "Ça, faut voir, c'est intéressant". Je vois pas d'autre solution, car c'est pas écrit dessus. Soit vous partez de la notoriété des artistes ; si vous tapez "Fred Forest" vous aurez des choses, mais c'est pas forcément là que le travail est le plus intéressant. Je pense qu'il y a des gens peut-être moins connus qui ont fait des trucs plus intéressants* »

Ce « *travail cumulatif* », qui ne peut être fait par aucun automate, dans un domaine où le savoir reste d'ordre privé et diffus, pourrait être assurée par la BnF :

> « *Deuxième chose, c'est vraiment d'arriver – la BnF a plus de moyens que d'autres lieux pour ça –, c'est que le travail des chercheurs soit largement accessible, car je crois beaucoup au travail des chercheurs sur les fonds. En gros, que la BnF essaie de capter autant que possible le travail fait par les chercheurs sur son propre fonds. Ce qui fait complètement défaut chez Internet Archive : de savoir que monsieur untel, un fouilleur de grimoires, a travaillé pendant deux ans sur le truc, et puis a laissé des traces de son travail. Ça c'est intéressant, car quand on connaît la préoccupation d'un chercheur, ça donne aussi beaucoup de pistes* […] »

Malgré l'appel à des pratiques académiques traditionnelles (programmes de recherche, communautés de recherche), un chercheur insiste sur le savoir contenu également dans les « *pratiques amateurs* » : celles des amateurs et des passionnés, proches de la retraite, « *les gens qui ont une mémoire de ça* », et prennent le temps de « *remettre leurs souvenirs en ordre* ».

3.2. Les professionnels

3.2.1. Pratiques du Web

Les professionnels interrogés ont des usages multiples du Web. La grande variété de ces usages amène les professionnels à les distinguer en fonction de temporalités (« *ça, c'est quotidien* »[8]), de spatialités (« *en amont* », « *en aval* ») ou de contenus (« *ça, c'est de l'information* »). En croisant ces catégories, on peut dessiner trois ensembles d'usages :

- Des recherches ciblées, indispensables pour leur activité quotidienne et situées le plus souvent « *en amont* » de celle-ci : accès aux publications officielles, recherche d'information sur un client, etc. ;
- Une veille plus fluctuante, qui vise à « *sentir un peu l'air du temps* », « *donner la température* » d'un domaine, « *suivre l'actualité* ». Cette veille se concentre autour des blogs, si l'on entend ce terme en un sens large qui les distingue d'abord des sites institutionnels et professionnels[9]. Le mot « blog » renvoie à une parole personnelle sur le Web, d'autant plus intéressante qu'elle est spécialisée. Chaque professionnel cite naturellement deux ou trois blogs qu'il suit quasi-quotidiennement et dont il souligne la valeur en termes de contenu : tel blog « *très étonnant* », d'une « *valeur incontestable* ». Dans cet univers des blogs, sont cités aussi bien les blogs de praticiens et de créateurs (juristes, écrivains), que les blogs de fans (« *qui sont souvent des choses très pointues, de très bonne qualité* »). Tous les domaines sont concernés : même dans le domaine du droit, où le commentaire juridique était auparavant détenu par l'université, avec une parole de praticiens limitée à des sujets techniques et pratiques, on assiste à la « *libération d'un espace non contrôlé pour les professionnels du droit* ». Si un professionnel interrogé est plus réticent à parler positivement de ce domaine, il reconnaît le surveiller. Il a soin d'opposer les verbes « *surveiller* » et « *regarder* » à « *consulter* » : « *Je le surveille plus que je consulte… Je regarde* » ;
- Un usage d'enrichissement et de contextualisation de contenus qui se situe « *en aval* » de l'activité habituelle : Google Earth, Street View ou Flickr. En apparence plus anecdotique, cet usage n'en dit pas moins un nouveau rapport à l'Internet depuis quelques années (les sites et logiciels cités ont tous été lancés dans les années 2004-2007) : donner à voir le monde à distance. Au point que Flickr fait partie des sites qu'un professionnel recommande d'archiver.

➔ Un cas singulier : la recherche préalable à une demande de brevet

De par ses missions spécifiques, le cabinet de conseil en propriété industrielle a un usage du Web différent de ceux précités : son service de veille et de documentation croise la consultation de serveurs payants dédiés au domaine des brevets avec une recherche transversale sur la totalité du Web, effectuée avec Google. En effet, dans le cadre du dépôt d'une demande de brevet, il s'agit de vérifier que l'invention n'a pas été préalablement mise à disposition du public. Comme n'importe quelle page du Web peut être un lieu de divulgation, le service « *cherche tous azimuts. On a beaucoup de choses sur les blogs, les forums…* » Les métaphores utilisées, empruntées aussi bien à la mythologie grecque qu'aux jeux vidéos, insistent sur le caractère démesuré et indéfini de cette recherche : « *Je pense qu'on a ouvert la boîte de Pandore* » ; « *On est des "Pacman", des "Pacwoman"* »[10]. Dans une telle recherche, où il s'agit d'établir ou de contester des antériorités, l'information collectée ne sera utilisable par les ingénieurs brevets que dans la mesure où elle sera datable (jours calendaires). Sur le Web, la date de mise en ligne est donc primordiale : « *Toute information mise sur le Web aujourd'hui, pour nous, dans nos métiers, il faudrait qu'on soit capable de pouvoir dater à quel moment elle est arrivée sur la page et toutes les modifications ultérieures.* »

Pour ses recherches larges sur le Web, le service de veille et de documentation a recours également au site de la fondation américaine Internet Archive (http://www.archive.org/), même si la collecte est jugée souvent insuffisante en regard des besoins : « *C'est pas très fiable parce que, bien entendu, tous les sites Internet ne sont pas passés dans cette moulinette* ». Est appréciée la « *présentation en elle-même* » du site, en particulier la « *sorte de petit tableur Excel* » qui indique les dates de collecte

[8] Les citations entre guillemets et en italique renvoient à des verbatim collectés durant les entretiens.

[9] La description du Web qui ressort des entretiens se limite souvent à une partition simple : 1) un domaine officiel (« ce qui sort des universités », « institutions publiques », domaine « scientifique et technique ») ; 2) les blogs (mis en parallèle avec les expressions : « des sites qui m'intéressent », « sites personnels », « sites spécialisés »). Auxquels peuvent s'ajouter : 3) les sites de journaux et d'entreprises ; 4) les forums.

[10] Nous retrouverons un même usage des métaphores lorsqu'il s'agira d'évoquer, avec d'autres catégories d'usagers potentiels, l'archivage du Web : « *Pyramide de Kheops* », « *projet borgésien* », etc.

successives d'un même site. Mais si le dépôt de demande de brevet a eu lieu entre deux collectes trop éloignées, les chances de prouver une éventuelle divulgation sont limitées. « *Entre le 10 janvier et le 25 mai, je ne sais pas si la page du 10 janvier est restée du 10 janvier au 25 mai dans le même état* ». Notons enfin que cette recherche sur Internet Archive vient seulement dans un temps second ; elle est effectuée à partir d'URL, préalablement repérées sur le Web actuel :

> *« Je fonctionne uniquement par URL. Parce qu'en général, ce que je fais : je commence une recherche au jour d'aujourd'hui sur le moteur de recherche. Je vois qu'il y a un site qui parle de ça, mais qui s'est un peu modifié ou – beaucoup dans les médicaments –, ce que je vais faire, c'est que je vais essayer de trouver les pages du titulaire du brevet, et d'essayer de remonter le plus possible vers le moment où la molécule est sortie, pour voir ce qui a été écrit et à quel moment. »*

Le site d'Internet Archive est décrit uniquement par son origine géographique : « *C'est un site américain* », sans autre indication de son statut légal. La confiance dans le contenu présenté semble naturelle à son utilisateur : alors que le Web est reconnu comme un domaine d'information aujourd'hui peu fiable (« *pendant des années, on prenait ça pour argent comptant, même au niveau scientifique. Alors que maintenant, même les pages Wikipédia, on se méfie du contenu* »), aucune question particulière n'émerge sur les principes ou les garanties des collectes de la fondation : « *Je fais confiance au site* » est répété deux fois par l'utilisateur.

➔ Un rapport méfiant aux contenus numériques

Malgré les outils informatiques disponibles pour conserver sous format numérique un contenu Web et pouvoir y accéder de n'importe où, l'impression papier est citée comme une pratique courante : des examinateurs de l'Office européen des brevets peuvent opposer des pages Web imprimées à une demande de brevet (dans certains domaines : informatique en particulier) et le journaliste interrogé reconnaît une pratique d'impression papier systématique des pages qui l'intéressent sur le Web :

> « *Sachant que ces sites – surtout sur les sujets un tout petit peu sensibles – bougent, sont modifiés, moi, systématiquement, j'imprime. Depuis qu'il m'est arrivé de trouver un truc très intéressant sur le net et d'y retourner le lendemain et de ne plus le voir, j'imprime. J'imprime beaucoup. D'abord dans un souci d'organisation personnelle, de clarté, pour avoir tout sous les yeux… Et parce que je sais que parfois ça disparaît.* »

Le journaliste joint ces impressions papier aux dossiers relatifs à chaque affaire, qu'il conserve ensuite dans son bureau. Cette pratique obéit pour lui à une double nécessité : non seulement conserver des pages qui peuvent disparaître, mais également conserver des pages dont le chemin d'accès pourrait être difficile à retrouver sur le Web.

Par contre, la recevabilité juridique d'une page Web imprimée est diversement appréciée selon les instances : un tribunal français a rejeté l'impression d'une capture d'écran utilisée comme preuve, alors que les examinateurs de l'Office européen des brevets ont recours à ce type de document pour s'opposer à une demande de brevet (sous la forme également d'une simple copie papier avec indication de l'URL, pour une page toujours en ligne) : « *Je crois, il me semble* [qu']*on n'est pas très d'accord en ce moment sur l'utilisation…* » [11]

---

[11] Il n'y a, à ce jour, qu'une jurisprudence dans ce domaine. Les juristes s'accordent sur le fait qu'un constat d'huissier sur Internet, établi en respectant certaines règles, constituera devant un tribunal une preuve difficilement contestable. Par contre, dans une affaire récente de contrefaçon, trois captures d'écran récupérées devant huissier à partir d'Internet Archive ont été refusées en Cour d'appel (arrêt de la Cour d'appel de Paris du 2 juillet 2010) :
« Que s'il n'est pas contesté que les pages en question n'ont pu faire l'objet de falsification postérieure, il convient, toutefois, d'observer que l'indication des dates précitées sur la page de résultat des recherches relatives aux pages archivées du site M6 Boutique au cours des années 1996 à 2009 du site The Wayback Machine, et en bas des tirages des pages écran n'établit pas avec certitude qu'à chacune de ces dates, s'affichait, dans la configuration imprimée, la page écran correspondant ;
Que, comme le relève justement la société HSS, le constat a été effectué à partir d'un service d'archivage exploité par un tiers à la procédure, qui est une personne privée sans autorité légale, dont les conditions de fonctionnement sont ignorées ; qu'il ressort de l'extrait des questions posées sur son fonctionnement, communiqué en pièce n°12 par l'intimée, que cet outil de recherches n'est pas conçu pour une utilisation légale ;
Que pour obtenir depuis ce site tiers un relevé des pages archivées provenant de M6 Boutique, l'huissier indique avoir tapé dans le champ de recherche l'adresse Url www.m6boutique.com/M6C/FR/Fi... ;
Que la page incriminée présumée datée du 15 septembre 2007 porte la référence http://web.archive.org/web/20070915... les références '20071107072924' et '20080114034935' apparaissant respectivement sur les deux autres pages ;

Devant cette incertitude actuelle, le cabinet de conseil évoque un retour possible aux preuves traditionnelles, jugées plus tangibles. Les publications ayant originellement un support papier sont associées à une présence physique indubitable (« *le papier, il est "là"…* »), dépendant d'un processus de validation (« *avant, quelque part, un livre était validé* »), alors que les contenus disponibles en ligne ne sont pas forcément validés et sont plus facilement falsifiables. Le cabinet mentionne ainsi une affaire récente où la preuve d'antériorité a été trouvée dans un Littré du XIX[e] siècle :

> « *Finalement, on retourne… Regarde les dernières affaires qu'on a eues : on a tendance à vouloir chercher des preuves dans des vieux livres, on cherche des vieux documents.* […] *Maintenant, je me dis que c'est très beau tout ça* [ndr : les documents sur le Web], *mais on va quand même… Mais est-ce que, humainement, on ne va pas rester dans l'envie d'avoir quand même une preuve papier.* […] *C'est assez paradoxal, même si on manipule toutes ces données informatiques, on a tendance à vouloir du papier. Et je pense qu'on va être de plus en plus méfiant sur les contenus, ça c'est clair.* [Un document sur le Web]*, c'est plus facile de falsifier* ».

3.2.2. Intérêt pour les archives de l'Internet

➔ Des représentations concurrentes

Un certain nombre de représentations du Web viennent freiner l'intérêt des professionnels pour les archives de l'Internet :

- Un usage essentiellement tourné vers l'instant présent : « *L'Internet a une espèce d'instantanéité, donc il va nous falloir beaucoup de temps pour associer ça à l'idée d'archive, d'archivage* » ; « *c'est la pertinence du moment où tu tapes qui t'intéresse* ».
- Une masse d'informations tellement énorme qu'elle semble se suffire à elle-même : « *Effectivement, on croit que Google nous donne tout* » ; « *Le Web nous a rendus paresseux. Tu tapes cinq mots et tu vois quand même ce qui tombe…* ». Il y a déjà suffisamment d'informations à traiter sur le Web actuel.
- Une masse d'information tellement énorme qu'elle semble constituer sa propre archive. Cette idée est renforcée par la place incontournable du moteur de recherche (Google), présentant dans sa recherche avancée une sélection par date de mise en ligne (« dernières 24h », « sept derniers jours », etc.) :

> « *Parti comme c'est parti, on a l'impression que tout reste en stock sur Internet : quelqu'un qui a écrit un truc en 2004 sur Internet... J'ai l'impression que mon stock d'archive, il est là ! J'ai qu'à ouvrir Google et à taper des mots-clés. Que va m'apporter de plus – enfin je parle très naïvement – que va m'apporter de plus un archivage à la BnF ? J'en sais rien. Parce que moi, j'ai l'impression que la grande originalité de ce système, c'est… comme tu n'as pas de problème de stockage, au sens spatial du terme, tout est encore là. Et si tu sais le chercher, il est là. Alors qu'un vieux livre, du XVII[e] ou du XVIII[e] siècle, j'aurais beaucoup de difficultés à le trouver par ailleurs. Donc, là, ok, je vais me déplacer. Mais je ne vois pas exactement ce que vous vous offrez en plus.* »

Enfin, à titre personnel, un professionnel interroge la volonté actuelle de tout collecter. Celle-ci nous empêche d'être vraiment créatifs en nous arrimant au passé et à son ressassement : « *L'archivage systématique tient d'une névrose* » ; « *Trop c'est trop. Suffit !* ». Cette critique ne porte pas d'abord sur la qualité de ce qui est collecté (même si celle-ci est évoquée dans un second temps), mais sur un rapport maladif au passé. L'archivage systématique va contre une certaine naturalité de l'oubli : « *Il y a tout de même un tri un peu naturel* », « *faut accepter que notre passage et que notre époque soient temporaires, évanescents, ne soient pas sacrés* ».

➔ Venir à la BnF : un coût humain important

A ces représentations concurrentes, vient s'ajouter un frein important pour les professionnels : le coût humain d'un déplacement rapporté à l'incertitude de trouver ce que l'on cherche. Le déplacement physique est « *une barrière importante* » dans des emplois du temps surchargés. Cette incertitude concerne à la fois à l'accessibilité du document (dans un cas mentionné par un professionnel, le

---

Que l'absence de toute interférence dans le cheminement donnant accès aux pages incriminées n'est donc pas garantie ; que pas davantage n'est-il démontré de façon incontestable à quelle opération précise –affichage, modification, retrait, archivage ou autre- correspond la date mentionnée dans la référence de ce cheminement ;
Qu'il s'ensuit que le constat dressé le 2 mars 2010 est dépourvu de toute force probante quant au contenu […] »

document demandé avait disparu des magasins) et la complexité administrative de la Bibliothèque. La Bibliothèque est mal connue dans son fonctionnement et ses missions, et cette méconnaissance suscite des craintes qui sont à la hauteur du caractère intimidant de l'institution. Le simple présentatif « *C'est "la BnF"* » semble résumer à lui seul un imaginaire qui la maintient à distance. Dans cet imaginaire, aller à la BnF s'apparente à un parcours du combattant, coûteux en temps :

> « *L'image que j'ai de la BnF,* [c'est] *de me dire : "On va arriver, on va pas trouver* [le document]*, on va passer trois plombes à trouver …" Alors c'est peut-être faux, mais l'image que j'en ai, c'est ça : on va passer de services en services, on va perdre du temps pour y aller et pour y revenir ; sur place on va avoir un mal de chien, et finalement…* »

L'accent négatif est mis également sur les procédures : « *Il faut faire une demande* » ; « *On aurait dû justifier le besoin de consulter le livre* ». Un journaliste, usager ponctuel de la BnF, confirme cette impression : aller à la BnF, « *c'est quand même une démarche, en énergie, en accréditation, en autorisation, etc.* » Un ingénieur brevet va se sentir plus à l'aise dans une bibliothèque universitaire, qui lui semble mieux correspondre à son élément naturel, car « *c'est un endroit où tous les livres qui traitent de science sont* ».

Face à ce coût humain, le professionnel ne viendra en bibliothèque de Recherche que s'il est sûr d'y trouver « *les choses que je ne peux pas trouver en faisant une recherche* [avec Google] » ; « *ce que je ne peux pas consulter directement sur mon ordinateur* ». La définition peut sembler évidente, mais elle permet de comprendre la manière dont le professionnel se représente son intérêt potentiel à venir : il réfléchit en termes négatifs (« *ce que je ne peux pas trouver* » ; « *ce que je ne peux pas consulter* » ailleurs qu'à la BnF), et non pas en fonction d'une valeur ajoutée en termes de services, d'aide à la recherche ou de cadre de travail. Il demande à être convaincu de l'exclusivité des informations que vont lui apporter les fonds BnF. Or, les entretiens montrent qu'il ne l'est pas pour le moment. Plusieurs professionnels ont eu l'expérience, ou ont l'impression, de pouvoir trouver par d'autres moyens ce qu'ils cherchent : « *Plusieurs fois, trois ou quatre fois, je me suis dit : "Tiens, il faudrait que j'aille à la BnF", et je n'ai jamais eu à sauter le pas, car j'ai trouvé d'autres moyens* ». Autre exemple mentionné : le cabinet de conseil a trouvé sur Google Books le Littré qu'il avait besoin de consulter (via l'Angleterre, car sa consultation en ligne n'était pas possible à partir de la France). Le journaliste interrogé partage cette même impression, alors qu'il a recours par ailleurs à la BnF de manière ponctuelle (MSS et ASP) : « *Ce que nous donneraient les archives de l'Internet, on va l'avoir indirectement, par un autre biais, un témoignage, une vidéo* ».

La frontière entre les contenus disparus et les contenus en ligne est elle-même questionnée, tant le Web semble aujourd'hui profond, en partie caché. Entendant que certains sites n'existent plus, un professionnel s'étonne : « *ils ne sont plus en ligne ? Même pas en page cachée ?* […] *Il me semblait qu'il y avait moyen de retrouver des états antérieurs* ». Pour faire apparaître clairement cette ligne qui sépare le Web actuel de son archive, et susciter l'envie de venir, un professionnel propose de mettre en avant le « *truc* » qu'ici on ne va pas trouver là-bas :

> « *Car, à mon avis, il va falloir, si j'ose dire, vraiment motiver les gens. Car encore une fois, la personne qui veut voir l'édition originale du* Misanthrope [de Molière]*, elle n'a pas trente-six solutions : elle est un peu obligée de se déplacer ou de consulter Gallica. La difficulté avec l'Internet, c'est que le média lui-même est un moteur de recherche. Donc, par définition, tu as le sentiment que tout est là. Donc il faut vraiment, si j'ose dire, motiver les gens, pour leur dire : "Attention, vous allez trouver là-bas un truc que vous n'allez pas trouver ici."* »

➔ En conclusion : des besoins surtout ponctuels

- Le conseiller en E.reputation reconnaît l'intérêt qu'il y aurait à donner une « *profondeur historique* » aux analyses de présence d'une marque sur le Web, mais également à l'évolution dans le temps de la prise de parole de tel ou tel concurrent. Les archives de l'Internet apporteraient un « *socle de réflexion* ». Le besoin d'une telle recherche est avéré pour l'avenir : « *On en aura de plus en plus besoin* », mais il n'est pas encore formulé comme tel, à ce jour, par les clients. Tout d'abord parce que les analyses marketing du Web n'en sont qu'à leurs balbutiements, même si « *des approches commencent aujourd'hui à se structurer avec des logiques de tracking* » et d'analyse du Web à l'instant *t* (est citée ici la société Louis Harris). Mais également parce que la valeur ajoutée d'une telle recherche n'est pas évidente à démontrer pour des clients qui attendent des résultats rapides et au meilleur prix.

- Le journaliste précise pour sa part qu'il est rarement dans la recherche de « *la preuve ADN* », dans « *la preuve de justice* » : « *Je suis dans l'information. Moi ça me suffit* ». Une information importante qui aurait disparu en ligne et aurait été archivée à la BnF a des chances d'être retrouvée par d'autres moyens (témoignages). Jugées comme une proposition « *hyper pointue* », les archives de l'Internet représentent « *la marge de la marge* » des lieux où un journaliste est susceptible de faire des recherches. Un besoin ponctuel n'est cependant pas exclu, mais seulement dans le cas d'une enquête « *lourde, longue, où vraiment il y a un vrai… une vraie question importante qui est très déterminante pour l'article, où quelqu'un nie farouchement, conteste : là, oui.* »
- L'avocat quant à lui ne mène pas d'enquête ; il est saisi par un client qui a déjà constaté sur tel site en ligne que ses droits étaient bafoués. Il s'agit donc non pas de chercher, mais de « *solidifier* » des éléments de preuve déjà fournis en amont par le client. Par exemple : faire constater par un huissier que tel produit est vendu en ligne. Que la preuve se trouve sur une page Web disparue et demande d'effectuer une constatation à la BnF n'est bien entendu pas exclu, mais l'avocat interrogé rappelle que l'infraction constatée doit être suffisamment importante pour motiver de la part d'un particulier ou d'une société une procédure longue et coûteuse. La plupart du temps, la preuve s'obtient directement « *sur le marché* ». La contestation de brevets lui semble le seul domaine où les archives auraient une utilité immédiate, ce que confirme l'entretien avec le cabinet de conseil en propriété industrielle.

3.2.3. Recommandations et attentes

➔ Contenus (collections)

Pour la collecte large, le domaine .fr est jugé très insuffisant : « *Le .fr est aujourd'hui un peu dépassé. Nous sommes*, [dans notre cabinet], *devenus .eu. Aujourd'hui, vous ne crawleriez pas du tout notre site Internet.* »

Pour une collecte ciblée, sont indiqués :

- Un ensemble institutionnel ou officiel (« *public* »), souvent cité en premier, et décliné en fonction des catégories professionnelles interrogées : « *Journal officiel* », archives du *Monde*, « *ce qui est scientifique et technique* […] *tout ce qui sort des universités* », « *les colloques, les consortiums* », « *les pages d'entreprises* ».
- Les blogs : la valeur ajoutée de cette forme d'expression est aujourd'hui largement reconnue, surtout quand elle est spécialisée dans un domaine. Avec une précision importante : il faut archiver les blogs « *et leurs commentaires* », car « *sociologiquement, sur l'état de l'opinion, de l'expression – pas toujours très heureuse d'ailleurs – je trouve ça assez intéressant* ».
- Un exemple singulier, Flickr : « *c'est très intéressant sur ce que ça dit des gens, de leurs intérêts, de leur pratique photographique* ».

Par contre, dans le cas précis de recherches portant sur la divulgation d'information, le cabinet de conseil en propriété industrielle ne voit pas « *de périmètre limite* » : pas de profondeur limite (« *Ça peut aller très profond dans le site, puisque ça peut aller à un power-point, c'est pas seulement la page* ») ; pas non plus de catégorie particulière de sites ou de documents à archiver (« *Nous on balaye un petit peu tout en zappeur : des images, ça peut nous être utiles, des vidéos, quand on est sur des problématiques en mécanique* »).

Notons que ces contenus illimités sont autant demandés que redoutés. Le service de veille et de documentation du cabinet de conseil reconnaît être déjà submergé par les informations non traitées, malgré l'augmentation constante en moyens humains, l'ouverture d'antennes en province. Cet engorgement est causé par des données collectées automatiquement (« *crawler* », « *alertes automatiques* »), signe que les moyens techniques vont aujourd'hui plus vite que le temps humain nécessaire au traitement de l'information :

> *« On n'a pas installé de crawler spécifique qui nous ramène de l'info, parce que déjà, rien qu'en faisant des veilles vraiment basiques, on est déjà sous l'eau. Le problème auquel on se confronte aujourd'hui, c'est qu'on a plein d'outils qui peuvent nous ramener de l'info, mais au bout d'un moment, on peut pas les mettre en place. A la limite, on veut pas les mettre en place* [rires] *! Parce que ça va nous crawler un tas de trucs, et on n'aura pas le temps de les lire. Par exemple, quand on fait des libertés d'exploitation* […]*. Rien que ça, l'OEB* [Office européen des brevets] *a mis des alertes automatiques. Donc nous, on a rentré les numéros de brevet des ingénieurs qu'ils veulent suivre, et on est bombardé tous les jours, et… c'est bien gentil, on est bombardé, mais nous on les*

met à côté, on fait notre travail quotidien, et, du coup, les alertes elles s'empilent, et le jour où on veut faire quelque chose, on recommence à zéro, car il y a des piles. On est un peu… le serpent se mord la queue : on est alerté, mais on n'a pas le temps de le faire. On arrive vraiment à un point de saturation où on se dit : "comment on va faire ?". »

La conclusion sur l'intérêt pour le cabinet de consulter les archives de l'Internet est donc ambivalente : « *Oui, ce sera intéressant, parce que je récupérerais des choses, mais je n'aurai pas le temps de les digérer. Ça m'angoisse, car je vois les choses arriver et je n'ai pas le temps de le faire !* ».

➔ Outils et métadonnées

Le professionnel veut pouvoir télécharger la page archivée et l'imprimer, étant donné qu'il est dans une perspective d'usage, et non d'érudition ou de curiosité (il cherche une archive pour en faire quelque chose).

Cet usage potentiel étant majoritairement celui de la preuve (avocat, ingénieur brevet, journaliste), la date et le contenu du document doivent être authentifiés : « *A la limite, on a démarré sur l'authentification de la date ; mais le problème ça va être ce qu'il y a dedans : est-ce qu'on peut être sûr que personne n'ait modifié la page, n'ait changé quoi que ce soit ?* ». Dans le cadre d'un archivage par la BnF, la simple mention de ces éléments suffit, compte tenu de la confiance dont bénéficie naturellement l'établissement. Deux professionnels interrogés insistent sur ce point : « *Je ne suis pas méfiant quand je vais à la BnF, pas du tout !* » ; « *J'irai les yeux fermés* » ; « *Vous êtes des gens sérieux tout de même* ». Il ne viendrait pas à l'idée des professionnels interrogés de mettre en doute l'authenticité d'une archive consultée à la BnF.

Par contre, dans ses relations avec un tiers, le professionnel peut avoir à certifier que le document est bien issu des collections de la BnF :

- Dans le cas d'une impression papier, un signe distinctif doit indiquer que cette impression a été faite à la Bibliothèque ; le signe de reconnaissance qui vient naturellement à l'esprit est celui de la République, comme un document administratif : « *avec une Marianne. Voilà, dès qu'il y a une Marianne, c'est bon* ».
- Dans le cas d'une archive en version électronique, un professionnel demande pareillement « *une certification, une sorte de signature électronique* ».
- Si l'archive devait donner lieu à constat par huissier[12], l'avocat demande à ce que soient accessibles facilement les éléments dont l'huissier a besoin en temps normal pour faire un constat en ligne (moyen informatique et type de connexion utilisés, état de la mémoire cache de l'ordinateur, etc.).

Seul le journaliste interrogé demande que les sites archivés aient également une notice bibliographique : « *Une petite biographie du site, si j'ose dire. Au moins : sa date de naissance, son acte décès, ses animateurs. Une petite fiche signalétique du site. Que l'on sache un peu : qui parle ? Quel est le champ d'action ?* »

➔ Une possibilité de sous-traiter la recherche

Le déplacement étant à plusieurs reprises mentionné comme rédhibitoire, deux professionnels émettent le souhait que ces archives soient directement accessibles en ligne, hors les murs. En effet, si le chercheur a l'habitude des longs trajets pour consulter un fonds d'archive, le professionnel est dans une autre temporalité. En ne permettant pas l'accès hors les murs aux professionnels, « *vous restreignez automatiquement à la région parisienne* ».

Malgré le coût humain d'un déplacement à la BnF rapporté à l'incertitude d'y trouver ce que l'on cherche, l'intérêt pour les archives de l'Internet est confirmé par le cabinet en propriété industrielle (« Êtes-vous intéressés ? » : « *Oh oui !* »). Dans le cas d'un dossier important, le cabinet ne regarde pas aux moyens engagés : « *En fait, quand on a vraiment besoin, on a la carte Gold illimitée…* ». La seule contrainte incompressible est alors celle du temps : pour une recherche récente qui demandait la consultation d'un ouvrage en bibliothèque de Recherche, « *on l'a pas fait parce qu'on était débordé* ». Dans ces conditions, le cabinet demande si un service ne pourrait pas être dédié aux professionnels, suggéré ici par le terme de « *sous-traitance* » : « *Je sais pas si par exemple on aurait pu sous-traiter, s'il*

[12] L'avocat interrogé indique cependant qu'il ne sait pas comment ni sous quelle forme un document issu du dépôt légal peut être utilisé à titre de preuve. Il précise donc qu'il transpose simplement la procédure habituelle du procès-verbal par huissier sur Internet : « *Lorsque* [l'huissier] *fait son constat sur un site Internet actuel, actuellement en ligne, il va faire toute une série de démarches préalables à ces constatations pour authentifier son procès-verbal : vider la mémoire-cache de l'ordinateur pour vérifier que cette connexion est bien contemporaine, qu'il se connecte pas à un site qui était antérieurement exploité, il va exposer qu'elle est sa connexion et quel moyen informatique il utilise.* ».

y a des départements sur lesquels on pourrait s'appuyer en sous-traitant, en leur disant : "On cherche tel type d'information ?". Est-ce qu'il y a des documentalistes ? »

3.3. Le « tout venant » de la bibliothèque de Recherche

3.3.1. Pratiques du Web

Pour le « tout venant » de la bibliothèque de Recherche (trois étudiants, un universitaire et un chercheur-amateur sur le site François-Mitterrand), les usages décrits sont à la fois spécifiques et éclatés. L'âge n'a ici aucune influence : tous n'ont pas la passion du Web et n'en font pas un outil primordial pour leur recherche.

Le Web comme objet d'étude à part entière est uniquement cité par Newman : doctorant en philosophie, il utilise les moteurs de recherche classiques pour voir si les concepts sur lesquels il travaille sont « *toujours actifs aujourd'hui* », si « *des communautés sur le Web s'y réfèrent* » et de quelle manière. Ainsi, une recherche sur la notion foucaldienne de « technique de soi », utilisée en mots-clés, fait remonter sur Google des résultats aussi variés que la gym orientale ou des sites d'associations d'autodéfense. Cette recherche sur l'actualité des concepts l'oriente exclusivement vers les contenus discursifs disponibles sur le Web, qu'il utilise et cite dans ses travaux comme n'importe quel autre support de discours : « *Comme dans les livres, je relève des citations* ». Newman évalue à 10 % la part des citations dans ses travaux scientifiques renvoyant aujourd'hui à des documents Web ; usage qu'il qualifie lui-même de « *fréquent* ». Notons que cet usage du Web n'est pas accepté par toutes les instances académiques : à l'opposé, un autre enquêté indique que son directeur de recherche refuse systématiquement les références à des sites Internet dans les travaux rendus (histoire de l'art).

Alain, chercheur-amateur à la retraite, passionné par l'histoire de sa région natale (l'Anjou), ne consulte l'Internet qu'en complément d'une recherche préalable sur les archives papiers. Il vient d'abord consulter à la BnF les anciens journaux de sa région et teste ensuite ses découvertes sur le Web : « *Si je trouve quelque chose dans le journal que je dépouille, je vais voir sur l'Internet s'il y a un blog, un site, etc.* ». Ayant ainsi appris l'existence jusqu'en 1948 d'un chemin de fer régional en Anjou (le « Petit Anjou »), Alain est tombé le site Web de l'association des Amis du Petit Anjou http://www.petit-anjou.org. Il reconnaît par ailleurs faire régulièrement des impressions d'écran de ce qui l'intéresse sur le Web.

Deux étudiants interrogés, Amel et Pierre, n'ont pas de recherche sur le Web à proprement parler. Ils consultent un nombre limité de sites, uniquement professionnels ou scientifiques, qui leur fournissent des informations factuelles (chiffres, résultats d'études), toujours en lien direct avec leur recherche. Pour Amel, qui travaille sur l'économie du tourisme : le site de l'INSEE et des revues scientifiques en ligne. Pour Pierre, qui travaille sur l'histoire des ventes publiques, les sites actuels de vente : http://www.gazette-drouot.com, http://www.auction.fr, http://www.interencheres.com. Ces deux étudiants ne s'aventurent pas hors de cet ensemble bien délimité. Consulter des blogs pour leur recherche, « *Je n'ai jamais pensé à ça !* » (Amel). Même écho chez Pierre, qui n'a jamais fait le test d'entrer ses thèmes de recherche (« Estampe », « Gravure ») sur Google, dont il reconnaît par ailleurs « *ne pas être un grand consommateur* ». Comme Paul, universitaire américain, historien de la Première guerre mondiale, qui avoue ne pas aller « *fréquemment* » sur l'Internet pour sa recherche, même s'il reconnaît qu'il aura « *de plus en plus intérêt à le faire* » compte tenu de la mise en ligne de fonds d'archives.

De toutes les personnes interrogées, c'est Pierre qui est le seul à avoir entendu parler des archives de l'Internet, par l'intermédiaire d'un tract, mais sans avoir eu la curiosité d'aller voir.

3.3.2. Contenus

Après une brève présentation du principe des archives de l'Internet, plusieurs lecteurs interrogés ont spontanément cité des sites dont ils ont constaté la disparition ou la transformation. Sont ainsi cités : le site des Amis des frères Goncourt, disparu puis recréé sous une autre forme : http://www.goncourt.org ; le site consacré à Blake & Mortimer (anciennement « La Marque jaune »), disparu et relancé sous un nouveau nom par les mêmes auteurs : http://www.centaurclub.com ; ou encore un site dédié au concept créé par Deleuze et Guattari de « schizo-analyse », définitivement fermé.

Malgré la conscience de cette volatilité, l'intérêt de cet archivage est perçu de manière très différente. « *Oui, bien sûr, c'est une très bonne idée !* », s'écrie immédiatement Newman, afin de « *voir ce qu'il y avait à disposition à un moment donné* ». Le terme « *à disposition* » est éclairé dans la suite de la discussion : ce qui est à disposition, c'est ce qui remonte en tête des résultats des moteurs de recherche, « *les sites les plus visités* ».

Bon connaisseur du Web, Alain hésite dans un premier temps entre l'intérêt et l'incrédulité devant une telle entreprise, à grand renfort de métaphores : « *Mais ça doit être un travail apocalyptique !* » ; « *Un archivage systématique, mais ça paraît démesuré !* ». « *C'est un peu comme la pyramide de Kheops !* ». Mais dans un second temps, appliquant le principe que « *tout est susceptible de devenir important* » pour les générations futures, il propose de relever le défi de l'exhaustivité : « *A partir du moment où on a ce genre d'ambition, faut tout prendre !* », « *Faut tout rafler, c'est un pari sur l'avenir* », « *On ne peut pas décider pour le futur. Tout est intéressant* ». L'exemple du site http://www.petit-anjou.org, qui rassemble à peine plus d'une centaine d'internautes amateurs, est cité à l'appui : s'il venait à disparaître, « *Ce serait très ennuyeux* ». Si Alain exclut toute sélection, l'entretien se resserre cependant autour du domaine très riche des blogs amateurs, des sites de « *fans* », en particulier dans le domaine de la BD : ces sites amateurs « *peuvent être extrêmement bien faits. Y'a des pros ! Et ce sont souvent des gens de mon âge. Ce sont des chercheurs du troisième âge, des passionnés, qui travaillent comme des malades et qui ont des moyens.* » Certains de ces chercheurs amateurs ressuscitent des héros ou des domaines peu connus de l'histoire de la BD : ainsi d'une bande dessinée de chevalerie publiée par Bayard entre 1953 et 1959 (*Les Aventures de Thierry de Royaumont* : http://royaumont.free.fr) ou d'une maison d'édition de comics français des années 1950 (« Artima »), à laquelle un « *vieux militaire en retraite* » consacre un site de qualité. La nécessité de conserver ces sites est une évidence : « *Ce sont des sites très très bien faits, mais quel avenir ? Ça, ça va exister tant qu'il y a des gens vivants* [pour les maintenir] ».

Pierre, par contre, a du mal à percevoir l'intérêt d'un archivage pour les sites qu'il consulte (sites d'enchère et de ventes publiques, site de l'INHA) : « *Intérêt ? Je n'en vois pas !* ». Il lui suffit que ces sites aient « *leurs propres archives en ligne* ». Il est intéressant de noter dans sa réponse que l'expression « *archives en ligne* » recouvre en fait des réalités très diverses : Pierre cite aussi bien l'historique des ventes sur le site http://www.gazette-drouot.com que les conférences en ligne sur le site de l'INHA (dans une rubrique qui s'appelle effectivement « archives audiovisuelles »). Une archive, au sens d'un état antérieur du site, ne le concerne pas : étant donné qu'il ne cherche que des informations précises, peu importe que celles-ci soient accessibles ou non dans leur environnement Web d'origine.

Après une explication illustrée par quelques exemples, Amel comprend l'intérêt de l'archivage, afin de « *voir ce qui a été dit sur ces sujets-là* » et les « *évolutions dans le temps* » des discours sur le Web. Mais elle reconnaît ne pas avoir pensé à ce type d'analyse avant l'entretien.

➔ **Des représentations difficiles**

Au cours de l'entretien, plusieurs manifestent une difficulté à se représenter ce qu'est une archive de l'Internet. « *C'est un peu abstrait pour moi* » reconnaît Pierre. Pierre et Paul ont en fait tous deux d'autres représentations concurrentes dans la tête. Celles-ci ont nécessité à plusieurs reprises de réorienter l'entretien et d'expliquer à partir d'exemples précis la notion d'« état antérieur » d'un site. Pierre pense d'abord à l'archivage papier de documents relatifs à l'Internet. Il évoque ensuite les sites ayant « *leurs propres archives en ligne* » : pour lui, le glissement de textes ou d'informations de la première page vers d'autres pages dédiées constitue « *une archive du site* » (génitif ici subjectif et non pas objectif, qui dit l'ambiguïté de l'expression « archives de l'Internet »). Paul pense quant à lui à un répertoire des archives disponibles sur l'Internet. Il y revient à plusieurs reprises, même après explication. Quand on lui demande ce qu'il faudrait collecter, il répond exclusivement en termes d'inventaire de ressources actuellement en ligne : « *Par exemple, quels catalogues sont consultables en ligne et ceux qui ne le sont pas ; quelles sont les archives en ligne et celles qui ne le sont pas* ».

A la question « Irez-vous consulter les archives ? », Amel répond : « *Oui, normalement oui. Venir une fois, pour voir si c'est intéressant, si c'est bien classé* ». L'intérêt exprimé est encore occasionnel, ne dépassant pas le stade de la curiosité.

3.3.3. Outils

Amel trouve qu'une recherche par mots-clés serait « *le plus intéressant* » et cite comme exemples de ce qu'elle aimerait pouvoir taper : « *développement durable* », « *tourisme* », « *consommation d'eau* », « *économie présentielle* ».

Alain est prêt à venir consulter les archives de l'Internet, mais il veut savoir à l'avance ce qu'il va chercher et propose de mettre à disposition des lecteurs un « *fichier* » des sites archivés : « *Je ne sais pas… Oui bien sûr je viendrai ; mais il faut savoir que tel site a existé. Si je sais qu'il a existé, j'aurais la curiosité de venir ici.* »

ANNEXE 1 : GUIDE D'ENTRETIEN

Pratique du web, connaissance et perception des archives de l'Internet :

- Quel est votre usage du Web dans votre pratique de recherche/professionnelle ? Ce que vous y cherchez, comment vous le cherchez et comment vous l'exploitez ? Ce que vous produisez/publiez et comment vous le publiez [ne pas négliger l'aspect utilisateur « producteur » de contenus, en particulier dans le contexte du droit à l'oubli numérique] ?
- Connaissez-vous les archives de l'Internet ?
- Intérêt que ces archives pourraient avoir pour votre propre recherche : dans quels domaines, dans quelles perspectives (type d'exploitation/d'analyse, type de production : thèses, ouvrages, articles, calculs sur les données, etc.) ?

Les contenus souhaités :

- « Politique documentaire » à engager (autre formulation : que pensez-vous qu'il faille archiver en priorité ou au contraire abandonner ?) : choix des sites/choix des données [n.b. : pour certains chercheurs le « site » n'est pas l'unité pertinente], périodicité des collectes, profondeur ou périmètre de la collection, intérêt et sujets des « parcours guidés » [expliquer le terme si nécessaire] : comment les constituer ?

  [Point de vigilance : être concret, procéder par des exemples]
- Publics cibles et types d'utilisation à développer (utilisation pédagogique, recherche personnelle, etc.) ? Eventuellement : noms de personnes à contacter, utilisateurs actuels ou potentiels.
- Informations fournies dans l'affichage des résultats, dans l'affichage d'une page Web ?
- Seriez-vous prêts à collaborer aux collectes ciblées/thématiques ?

  [Point de vigilance : ici encore, être concret, citer au besoin les collaborations existantes]

En option :

- freins éventuels à votre utilisation en l'état actuel ;
- 10 minutes de découverte d'un site mettant à disposition des archives de sites Internet « en ligne » ; observation ethnographique, enregistrement des réactions. Le site pourra être celui de la BnF si l'entretien a lieu en bibliothèque de Recherche ; sinon, l'un des trois sites suivants :

http://www.archive.org

http://www.webarchive.org.uk/ukwa/

http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

ANNEXE 2 : LISTE DES SITES CITÉS DANS LES ENTRETIENS

Administrations publiques, établissements publics, gouvernance :
http://www.aeres-evaluation.fr
http://www.rand.org (Think Tank américain)
http://www.inha.fr
http://www.insee.fr

Art, net art :
http://www.cultdeadcow.com (réminiscences de l'undernet)
http://www.etoy.com (collectif d'artistes suisses)
http://www.fredforest.com (site de Fred Forest, artiste, théoricien et enseignant)
http://www.donforesta.net (site de Don Foresta, artiste et théoricien)

Design :
http://www.hyperbate.com/dernier/?cat=3 (blog d'un enseignant en design)
http://liftlab.com/think/nova (blog d'un consultant, Nicolas Nova)

Droit, justice :
http://www.maitre-eolas.fr (blog d'un avocat)
http://www.breese.blogs.com (blog d'un juriste sur la propriété intellectuelle)
http://www.laurent-mucchielli.org (blog d'un sociologue sur la justice et la délinquance)
http://www.artaas.org (Association de recherche sur les traitements des auteurs d'agressions sexuelles)
http://www.criminocorpus.cnrs.fr (plateforme de publications scientifiques sur l'histoire de la justice)

Internet :
http://bienbienbien.net (site de veille, en hibernation depuis 2010)

Littérature et bande dessinée :
http://passouline.blog.lemonde.fr (blog Pierre Assouline)
http://lettres.blogs.liberation.fr (blog Raphaël Sorin)
http://renaud.camus.pagesperso-orange.fr (blog Renaud Camus)
http://www.goncourt.org (site de la société des Amis des frères Goncourt)
http://royaumont.free.fr (site amateur consacré à l'œuvre du dessinateur Pierre Forget)
http://www.centaurclub.com (forum sur Blake & Mortimer)
http://meteor.proftnj.com/artima.htm (site amateur sur les éditions Artima)

Marketing :
http://www.docnews.fr
http://www.influencia.net
http://loiclemeur.com/france/ (blog de Loïc Le Meur)

Rencontres (sites à usage pornographique) :
http://bazoocam.org
http://www.chatroulette.com

Sociologie, histoire contemporaine :
http://anthem-group.net (The Actor network theory – Heidegger meeting)
http://coulmont.com (blog de Baptiste Coulmont, l'un des premiers sociologues à avoir eu un blog).
http://www.scriptopolis.fr (blog de trois chercheurs (sociologie et histoire) autour de « l'écrit et ses mondes »)

Urbanisme et innovation :
http://www.innovcity.fr (site d'information du laboratoire « Paris Région Innovation » sur les innovations dans les municipalités du monde entier)
http://www.urbain-trop-urbain.fr (urbanisme et architecture, blog animé par deux consultants)
http://leblogdelaville.canalblog.com (urbanisme, blog animé par un consultant)
http://www.groupechronos.org (cabinet d'étude sur l'innovation)
http://www.pop-up-urbain.com (blog animé par Philippe Gargof, géographe)
http://www.pasdetransportsansdesign.fr (agence pour la promotion de la création industrielle)
http://urbanites.rsr.ch/laboratoire-des-villes-invisibles/ (blog, un géographe et un ingénieur chercheur)