

Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire ?

Olivier Ertzscheid.

Maître de Conférences.

Université de Nantes. IUT de la Roche sur Yon. Equipe CREC.

Laboratoire DOCSI (Université de Lyon).

olivier.ertzscheid@univ-nantes.fr

Gabriel Gallezot

Maître de Conférences

Université de Nice - Sophia Antipolis

Urfist PacaC

Laboratoire I3M (EA 3820)

gallezot@unice.fr

Introduction

a) - Problématique

Le document est plus que le texte Dans ses déclinaisons matérielles et communicationnelles tout d'abord (multimédia). Dans ses résonnances sociales ensuite. Il semble aujourd'hui que « ceci » (le document) ait non pas « tué » mais remplacé « cela » (le texte).

Le phénomène que nous voulons ici analyser concerne le document à l'échelle de son déploiement sur les réseaux : en changeant d'échelle, il change nécessairement de nature. Le néologisme « docuvers » (contraction de document et d'univers) sur lequel nous reviendrons en est l'incarnation la plus claire. Or du point de vue de l'accès et de l'organisation des connaissances dans une perspective bibliothéconomique, l'arrivée des moteurs de recherche, la modification des acteurs de la chaîne documentaire, la possibilité offerte à chacun d'être auteur, éditeur et/ou diffuseur, l'atomisation et la fractalisation à laquelle sont soumis des corpus et/ou des documents autrefois perçus comme autant d'objets homogènes, cet ensemble de paramètres bouleverse et conditionne les usages associés au document. Cet article se propose de réfléchir à ces usages en adoptant la perspective d'analyse décrite ci-après.

La problématique documentaire à l'échelle des micro-réseaux (Intranets) comme des macro-réseaux (Internet) s'est aujourd'hui presque entièrement focalisée sur les deux questions que sont d'une part, le traitement de masses et de corpus documentaires inédits dans leur forme (formats) et dans leur taille (volumétrie), et d'autre part la question de la recherche et de l'accès pertinent et/ou raisonné auxdits documents.

De nouvelles logiques documentaires sont donc apparues marquant l'inscription de l'actuelle médiasphère dans un paradigme qui, plus que jamais, est celui décrit par la notion d'archive foucaldienne (FOU 94)¹. A l'appui de ces logiques, deux phénomènes interrogent particulièrement la question des pratiques et des usages du document.

b) - Hypothèses

Il s'agit, premièrement de l'arrivée et de la généralisation des pratiques d'indexation sociales, baptisées « folksonomies », dont l'horizon définitoire hérite de problématiques anciennes (organisation et classification des connaissances) mais également ancrées dans des pratiques

¹ Cf. infra

métier d'ordre techniciste (conservation et accès aux documents, métadonnées, Dublin Core), et enfin augurant des développements actuels sur le web de prochaine génération (web 2.0, ontologies sociales, web sémantique et socio-sémantique). Ces pratiques (folksonomies) disposent aujourd'hui d'outils et d'interfaces permettant un « balisage »² entièrement subjectivé et à vocation non pérenne de l'information et plus globalement des objets de connaissance.³

L'autre phénomène qui découle directement du premier concerne cette fois la renégociation, pour la sphère documentaire, de l'habituelle dialectique entre « carte » et « territoire », notamment observable au travers de ce symptôme que constituent les pratiques de « geotagging » (localisation et indexation géographique) mais également le constat, nouveau à cette échelle d'observation, d'une harmonisation et d'une auto-régulation spontanée de cette description pourtant aléatoire et non-raisonnée des contenus documentaires de toute nature (photos, textes, vidéos) avec la mise en place de motifs (patterns) tendant à démontrer qu'il pourrait exister pour tout ensemble ou unité documentaire donné, une série de termes (descripteurs) constituant le plus petit lexique commun permettant de la décrire pour optimiser son partage, son repérage et sa localisation. La préemption par un public non expert de techniques d'analyse et d'outils s'inscrivant habituellement dans l'héritage de la linguistique de corpus et plus globalement des sciences du document pose également question sur la nature de ce processus de description, sur les objets qu'il prétend embrasser, sur les processus de validation afférents et enfin, sur sa pérennité à l'échelle de la sphère publique connectée.

c) - Méthodes

De cet ensemble de faits observables notre analyse portera essentiellement sur les questions du repérage et de « l'accès » vécues au travers des exemples déjà cités que nous reprendrons et détaillerons, comme les fonctions du document les plus en lutte avec un environnement bouleversé par l'arrivée de nouveaux entrants, tels les moteurs de recherche et les « masses » d'utilisateurs anonymes assumant les fonctions allant de l'édition à la diffusion de contenus.

d) – Discussions

Nous reviendrons, en conclusion de notre étude, sur le fait que, continûment aux travaux et réflexions engagés par le groupe RTP-DOC (RTP 03) il nous apparaît que de toutes les fonctions liées au document et à sa sphère de socialisation et de médiatisation ainsi qu'aux enjeux d'engrammation des savoirs dont il est le dépositaire, de toutes ces fonctions, c'est celle de l'accès qui est la plus bouleversée. Ce qui implique que tout un ensemble de problématiques documentaires changent d'axe. Jusqu'ici, cet accès était globalement subordonné au classement (modèle bibliothéconomique). Or c'est désormais le classement qui pourrait être subordonné à l'accès, du fait, notamment des nouveaux modèles (économiques cette fois), dont sont porteurs les nouveaux entrants : nous reviendrons ici sur le problème que posent les politiques de numérisation massives de biens culturels par des sociétés commerciales, telle l'initiative de Google Print. Avec comme résultat le fait que l'on peut désormais craindre (et d'ores et déjà observer dans certains cas précis) que les documents qui seront « classés » seront prioritairement ceux qui seront « accédés », et donc accessibles. Ce genre de glissement paradigmatique (possible et non encore avéré) serait en tout état de cause le second "big bang" documentaire après celui d'Otlet.

² On parle de « tags » pour les balises et de « taguer » (tagging) pour le balisage de l'information

³ A ce sujet nous ne reviendrons pas sur les discussions entre folksonomie et ontologie, elles représentent chacune des manières d'organiser l'information qu'il conviendrait effectivement de détailler par un texte entier. Notons simplement pour notre propos que l'ontologie propose une modélisation d'un monde (d'un domaine) a priori alors que la folksonomie laisse les usagers modéliser leur vision d'un domaine, « sans a priori ».

1 : Nouvelles logiques documentaires ou théorie de la dérive ?

1.1 : Continents documentaires.

Le web naît officiellement en 1989 avec la publication d'un article de Tim Berners Lee « L'hypertexte et le CERN » (BER 89). Pour que le web existe comme « continent » documentaire, il faut la conjonction de trois éléments distincts : des adresses (URL) permettant de localiser l'information⁴, des navigateurs (permettant d'y accéder) et un format d'encodage (HTML) permettant d'afficher l'information récupérée. Sans oublier les protocoles d'échange de données (http et autres TCP-IP).

1.2 : Première période : « Quoi » indexer

Vers la fin des années 90 nous disposons ainsi d'un web public (le « www ») indexé par les moteurs et contenant différents types d'informations se déclinant elles-mêmes sous différentes formes documentaires : les articles scientifiques y côtoient les pages personnelles, les sites de presse et les documents factuels. A ses côtés, un web « opaque » se constitue via des organisations qui déploient à partir de base de données antérieurement constituées des pages web dynamiques, invisibles à l'œil des moteurs. Il s'agit dans ce dernier cas de documents générés à la volée, dynamiquement, à partir de requêtes déposées sur les sites par les utilisateurs. Ces « contenus documentaires » sont donc purement virtuels et n'ont pas d'inscription physique stable fût-elle numérique, sauf à considérer les bases de données comme des documents granulaires organisables à souhait.

Du point de vue de la recherche documentaire en particulier et de la recherche d'information en général, les premiers documents (web public) sont librement consultables et accessibles via les index des moteurs de recherche alors que ces mêmes moteurs peinent encore (pour des raisons techniques) à indexer le contenu des seconds, justifiant ainsi l'expression d'un « web invisible ». En termes de logiques documentaires présidant à l'indexation et au stockage des contenus, seuls les contenus du web public sont ainsi repérables. En parallèle, les pratiques informationnelles consistant à échanger des courriers électroniques ou à stocker des documents de travail sur son disque dur personnel échappent à ce mouvement. La question qui permet alors de scinder la masse documentaire en « visible / invisible », « indexée / non-indexée » est alors encore celle de la nature des contenus informationnels : « Quoi » indexer ? Avec une première évolution qui stigmatise un changement notable dans ces nouveaux lieux de stockage, d'accès et d'indexation que sont les moteurs et annuaires de recherche, puisque ce ne sont plus seulement des contenus validés par un processus éditorial (scientifique ou commercial) qui sont indexés et accessibles.

1.3 : Deuxième période : « Qui » indexe

Un pas est franchi à l'heure actuelle avec « l'indexabilité » des quatre types de documents évoqués ci-dessus : en sus du web public et du web encore il y a peu « invisible », ces continents documentaires que sont l'ensemble de nos correspondances électroniques personnelles ainsi que les fichiers et documents stockés sur nos ordinateurs personnels, sont désormais accessibles aux moteurs, lesquels les indexent aussitôt par le biais d'outils dédiés (Google Mail, Google Desktop)⁶. Ce bouleversement dans la perception documentaire (un document étant perçu comme appartenant à un espace public, si restreint soit-il) place entre

⁴ Avec notamment le DNS, puisqu'on passe d'un adressage IP en chiffre à une « étiquette textuelle »

⁶ Et ce de manière consciente ou méconnue par les utilisateurs de ses outils.

les mains de quelques acteurs marchands l'ensemble du matériau documentaire⁷ qui définit notre rapport à la l'information, et dans certains cas, à la connaissance : courriers privés, fichiers personnels, pages web publiques, pages web d'entreprises, publication savantes, ouvrages imprimés et fonds numérisés de bibliothèques. Un seul et même outil – ce qui constitue un gain – mais surtout, une seule et même société commerciale⁸ – ce qui constitue un risque - garantit l'indexation et l'accès à cet ensemble. L'objectif de cet article n'étant pas de déterminer si cela est une bonne ou une mauvaise chose, nous nous bornerons ici à indiquer qu'en termes d'accès et de droit à l'information, l'extrême mouvement de concentration qui touche ici la médiasphère est à tout le moins problématique. Ajoutons à cela qu'en sus de ces nouveaux documents, qui se définissent stricto sensu par leur capacité à être indexés, laquelle capacité les constitue de facto comme autant d'unités documentaires, de nouveaux usages informationnels voient le jour, ajoutant à cette masse déjà considérable une dimension relevant de « l'extime » (TIS 01), au travers du phénomène des blogs.

A l'échelle de ce nouveau continent documentaire réunifié, la question permettant de mesurer le bouleversement en profondeur des enjeux documentaires n'est plus celle de savoir « quoi indexer ? » mais bien « Qui ? » indexe. D'autant que toute indexation à un coût et la gratuité de celle-ci n'est qu'apparente : l'ensemble des contenus ainsi indexés est soumis à une analyse visant à rentabiliser les routines d'indexation par la diffusion massive de publicité contextuelle sur tout type de contenu documentaire (courriels, ouvrages, etc.).

Ce qui se dessine ici ressemble à l'incarnation de l'archive telle que décrite par Foucauld :
« Par archive, j'entends d'abord la masse des choses dites dans une culture, conservées, valorisées, réutilisées, répétées et transformées. Bref toute cette masse verbale qui a été fabriquée par les hommes, investie dans leurs techniques et leurs institutions, et qui est tissée avec leur existence et leur histoire. Cette masse de choses dites, je l'envisage non pas du côté de la langue, du système linguistique qu'elles mettent en œuvre, mais du côté des opérations qui lui donnent naissance. (...) C'est, en un mot, (...) l'analyse des conditions historiques qui rendent compte de ce qu'on dit ou de ce qu'on rejette, ou de ce qu'on transforme dans la masse des choses dites. » (FOU 94 p.786)

1.4 : Vers des bases de données « intentionnelles » ?

A la lumière de ce rapprochement entre des univers informationnels totalement distincts qui voit se heurter deux modèles antagonistes dans leurs fondements (celui, bibliothéconomique, de la bibliothèque avec son accès raisonné aux documents et celui, marchand, des moteurs plaidant pour la marchandisation de tout contenu documentaire), et à l'aube d'une troisième période dont certains analystes⁹ relèvent qu'elle se caractérisera par le recoupement systématique des données collectées en lien avec nos usages informationnels et documentaires privés à des fins de monétisation de services publicitaires, l'arrivée de pratiques « d'indexation sociale » peut être lue comme un réponse et une alternative possible à la situation monopolistique décrite jusqu'ici. Dans le même temps, ces pratiques interrogent à leur tour les usages du document et les modes de représentation et de navigation qu'une collectivité en réseau est spontanément capable de s'approprier.

⁷ Il faut indiquer, pour l'heure, l'exception notable de réseaux P2P qui représente encore un continent indexé et non marchand.

⁸ Google (www.google.com) dispose ici d'un leadership incontestable, lequel ne peut être élargi au delà des deux sociétés concurrentes que sont Yahoo et Microsoft. Les récentes négociations commerciales entre Dell et Google et Yahoo et Ebay viennent renforcer ce propos en terme d'accès unifié.

⁹ Voir à ce sujet le billet de Francis Pisani : http://pisani.blog.lemonde.fr/pisani/2005/10/the_search_2_no.html

2 : L'indexation sociale.

2.1 : Indexation normée et indexation grand public.

Depuis l'avènement d'Internet et l'arrivée massive sur le réseau d'informations relevant des sciences et techniques, la question du classement, de la préservation et de l'archivage documentaire n'appartiennent plus aux seuls moteurs de recherche. Le monde des bibliothèques et de la documentation a mis en œuvre des méthodologies (métadonnées), des normes et des spécifications permettant d'appliquer à ces contenus un ensemble de principes de classement de nature bibliothéconomiques (ex :Dublin Core¹¹), se réappropriant ainsi les problématiques afférentes et déployant pour y répondre de nouvelles compétences. L'exemple actuel de déploiement des archives ouvertes¹² et institutionnelles rendues interopérables grâce au protocole OAI-PMH en est l'exemple le plus frappant. Pour autant ces normes présentent un certain nombre de limites et ne s'appliquent, pour une très grande majorité, qu'à des contenus institutionnels ou scientifiques validés et ne concernent donc pas l'ensemble des documents constituant le web.

Une première possibilité de remettre les routines d'indexation aux mains des utilisateurs apparût par le biais de l'usage des métadonnées publiques, ne renvoyant pas, à l'inverse de celles du Dublin Core, à un mode d'organisation bibliothéconomique. Ces métadonnées permettent à chacun d'insérer dans le code HTML des documents publiés un certain nombre d'informations annexes (titre, description, mots-clés). Cette première tentative se solda pourtant par un échec et ce pour deux raisons principales :

- premièrement les pré-requis techniques étaient bien trop lourds pour des utilisateurs n'ayant aucune connaissance du balisage HTML ;
- deuxièmement, des professionnels du référencement s'emparèrent de ces balises afin de positionner au mieux des sites d'entreprises sur les moteurs de recherche et l'on vît à cette occasion apparaître du même coup toute une série d'usages déviants d'indexation (Spamdexing) consistant, par exemple, à positionner un site sur de faux mots-clés ou sur les mots-clés d'un concurrent. Ces pratiques se généralisèrent et furent reprises par certains sites de particuliers, faussant du même coup les techniques de classement statistiques des moteurs et obligeant ceux-ci à ne plus les prendre en compte, ou à très fortement minorer cette prise en compte dans leurs algorithmes.

2.2 : Indexation sociale et folksonomies.

Les folksonomies désignent « *un processus de classification collaborative par des mots-clés librement choisis, ou le résultat de cette classification*¹³ ». Elles puisent leur origine dans le croisement de deux phénomènes renvoyant à des techniques de recherche et de partage de documents. Tout d'abord des techniques de recherche et de filtrage. Grâce à des services comme Del.icio.us¹⁴, de nouvelles plateformes d'échange de signets ('social bookmarking') virent le jour. Sur ces plateformes en ligne, chaque utilisateur dispose d'un compte grâce auquel il peut déposer ses signets et les assortir de mots-clés. Il est ensuite possible de laisser un accès et une visibilité complète aux listes ainsi constituées pour les partager avec d'autres utilisateurs. En termes de recherche d'information et pour permettre la navigation dans des

¹¹ <http://dublincore.org/>

¹² Voir par exemple l'archive ouverte en sciences de l'information et de la communication « Archivesic » : <http://archivesic.ccsd.cnrs.fr/>

¹³ Définition extraite du site Wikipedia (<http://www.wikipedia.org>)

¹⁴ <http://del.icio.us>

listes ainsi constituées, ces services autorisèrent la mise en place de mots-clés choisis par les utilisateurs pour « décrire » tel site, tel signet.

2.3 : Re-documentarisation.

Deux tendances de fond vont ensuite venir appuyer une renégociation documentaire sans précédent à cette échelle. Premièrement, l'avènement d'une informatique répondant à des besoins nomades qui fait que nombre d'utilisateurs font dans leur routines quotidiennes de travail le choix d'outils distants. Une masse considérable d'informations bascule alors sur le réseau (textes et index). Ce changement d'échelle marque également un changement de statut. Une fois mises en ligne, ces informations se constituent comme autant d'unités documentaires. Ainsi les listes de signets utilisés par chacun d'entre nous dans le cadre de nos navigateurs, et à l'appui de nos navigations, se caractérisent d'abord par des fonctionnalités documentaires de nature technique. En franchissant la barrière du réseau elles s'affranchissent de ce statut de simple fonctionnalité pour devenir autant d'unités documentaires « stables », c'est à dire pouvant être recherchées et accédées en tant que telles dans le cadre des services les hébergeant ou par des moteurs de recherche classiques.

2.4 : Communautarisation.

L'autre tendance de fond est liée à l'explosion documentaire qui marque chaque évolution du web, la dernière en date étant celle liée au phénomène des weblogs. Devant cet accroissement, devant la multiplicité des outils de recherche permettant d'y accéder et renvoyant (dans le cadre d'une utilisation optimale) à une expertise et à des typologies documentaires bien trop complexes pour la majorité des utilisateurs, des communautés d'expertise se constituent pour faciliter l'accès à des contenus documentaires bien ciblés. Ces communautés d'intérêt ont comme premier objectif le partage de liens. Progressivement et en lien avec les logiques économiques de concentration d'offre de services des trois grands acteurs dominant ce marché¹⁵, d'autres services essaient et permettent à leur tour la multiplication d'unités documentaires en ligne¹⁶.

A l'échelle des weblogs, quelques outils de recherche dédiés¹⁷ ont rapidement mis en place des systèmes de mots-clés là encore librement choisis et en dehors de toute optique, considération ou compétence de nature documentaire. Au regard de la difficulté de mise en œuvre de normes idoines sur des contenus et des corpus documentaires pourtant déjà homogènes, au regard ensuite de l'usage perverti des métadonnées grand public, une analyse sommaire pourrait porter à croire qu'un tel système non-contraint et appliqué à des corpus documentaires totalement hétérogènes (dans leur formats comme dans leurs finalité) s'avérerait rapidement être un échec. Or il n'en est rien à ce jour et nous proposons ci-après quelques éléments d'analyse qui permettent de comprendre les raisons de cet engouement puis de ce succès et d'anticiper ses probabilités de constituer une modalité documentaire autonome et pérenne à l'échelle du réseau.

3 : Indexation sociale : les raisons du succès

¹⁵ Google, Yahoo et Microsoft.

¹⁶ Le service Flickr (www.flickr.com) permet ainsi de partager ses photos, chacune pouvant être assortie d'une série de « Tags ».

¹⁷ <http://www.technorati.com>

S'il est raisonnablement permis d'affirmer que ces folksonomies et autres techniques d'indexation sociale seront pérennes c'est principalement pour trois raisons.

Une raison historique tout d'abord. Comme nous l'avons indiqué plus haut, ces folksonomies s'inscrivent dans un héritage historique déjà ancien concernant la description et l'accès à des contenus documentaires. Elles permettent également d'ouvrir la voie aux ontologies sociales qui elles-mêmes constituent la base des futurs développements du web sémantique.

Raison économique ensuite : les outils de recherche, et particulièrement les trois grands, développent une offre de portails personnalisés dans laquelle sont intégrées de manière plus ou moins naturelle et intuitive, la gestion de ces communautés d'intérêt et la possibilité de soumettre tout type de document (depuis le mail jusqu'au photos de vacances en passant par des documents de travail) au partage et à l'indexation ('tagging').

Des raisons techniques enfin puisque voient le jour des interfaces¹⁸ souvent très intuitives qui même si elles ne font parfois que singer grossièrement les outils de la linguistique de corpus, ont le mérite de les rendre accessibles à tous en ajoutant une dimension de partage, de capitalisation et de diffusion en temps réel.

3.1 : Un faible coût cognitif.

L'activité de catégorisation est une activité cognitivement complexe qui s'incarne dans nombre de compétences des métiers de l'information de la documentation et des bibliothèques. Or de prime abord ces indexations sociales et le fait qu'elles réclament la mise en place de descripteurs en font, stricto sensu, des activités de catégorisation. On pourrait donc s'étonner de l'engouement des utilisateurs à leur sujet. Pour autant et en accord avec l'hypothèse avancée par (SIN 05), cette activité de tagging, à la différence d'une activité de catégorisation ou d'indexation dans un contexte documentaire professionnel, ne nécessite qu'un très faible coût cognitif :

« le marquage ("tagging") élimine la phase de décision (choisir la bonne catégorie) et dissipe la phase de paralysie d'analyse ("the analysis-paralysis stage") pour la plupart des gens. (...) Il offre un retour social et sur soi-même immédiat. Chaque "tag" traduit un peu de vos centres d'intérêts et les ancrent dans un contexte social immédiat. La beauté du marquage ("tagging") est qu'il est inscrit dans un processus cognitif déjà existant sans lui ajouter de coût cognitif supplémentaire. »

3.2 : La renégociation des espaces documentaires ou le passage de la carte au territoire par le « Geotagging »

Les rapports entre carte et territoire en terme de « représentations » sont déjà anciens. La carte en tant que support documentaire communicationnel et transactionnel occupa très tôt un rôle déterminant puisque dès le XVIème siècle elle était utilisée comme support du commerce. A l'heure actuelle, et suite au lancement par Google des applications Google Maps puis Google Earth¹⁹ lesquelles permettent de « naviguer » au sens propre sur des représentations en deux et trois dimensions de la terre (et ce avec des niveaux de détail et de résolution proprement impressionnants) la « carte » documentaire prend un nouveau sens. Avant de préciser lequel, rappelons que sur la base de ces cartes une série d'API²⁰ permettent de transformer lesdites cartes en autant de documents virtuels sur lesquels s'appuient des services innovants²¹.

¹⁸ <http://www.ivy.fr/revealicious/>, <http://www.tagclouds.com>

¹⁹ <http://maps.google.com> et <http://earth.google.com>

²⁰ applications mises à disposition gratuitement

²¹ Des sites (<http://65.39.85.13/google/default.htm>) permettent par exemple de coupler les fonctions de Google Maps avec des données démographiques et immobilières : le principe est celui des API couplées à Google Maps

Mais ces cartes ont aussi permis de redéployer une catégorie particulière de requêtes documentaires adressées aux moteurs de recherche. Il s'agit des requêtes transactionnelles : il est désormais possible, grâce aux techniques de géolocalisation et à la possibilité offerte aux utilisateurs d'interroger les moteurs depuis des terminaux mobiles (téléphones portables) de trouver « la pharmacie ouverte la plus proche », de « réserver dans un restaurant », de « commander une pizza », etc.

Au delà de l'aspect pratique induit par ces techniques de géolocalisation, ces dernières font désormais de chacun d'entre nous des citoyens non plus simplement « connectés » mais « situés ». La frontière entre d'un côté, un espace documentaire factuel (adresses, heures d'ouverture des restaurants par exemple) et de l'autre un espace documentaire incarné et totalement subjectivé, tombe définitivement.

Nous prenons non plus simplement symboliquement mais « physiquement » place au sein du document que nous parcourons. Ainsi le paradigme Batesonien selon lequel « *Le langage entretient avec les objets qu'il désigne le même rapport que la carte entretient avec le territoire.* » [BAT 77 p.212] apparaît remodelé conformément à la vision Deleuzienne²² des mécanismes de « dé- » puis de « re-territorialisation ». Car la navigation sur ce support documentaire que constitue Google Earth est de nature schizophrénique. Les espaces de la carte et du territoire semblent se confondre au nom des deux prémisses suivantes : nous sommes sur la carte, et la carte est (tout au moins numériquement) « à l'échelle » du territoire. C'est cette inscription documentaire subjectivée (et qui peut être déclinée à l'infini selon la localisation géographique depuis laquelle sont envoyées les requêtes et le type de service qu'elles visent) que stigmatise et illustre l'essor des techniques de Geotagging. Celles-ci désignent :

« *le processus d'ajout de métadonnées d'identification de nature géographique sur différents médias comme les sites webs, les fils RSS ou les images. Ces données sont habituellement composées d'une latitude et d'une longitude mais peuvent également prendre en compte des noms de lieux, des altitudes, etc ...* »²³

Là encore, et de la même manière que les représentations graphiques des systèmes de type folksonomies s'inscrivent dans l'héritage des outils de la linguistique de corpus, le geotagging se revendique comme l'application grand public des SIG (Systèmes d'Information Géographiques)²⁴.

3.3 : L'observation de motifs récurrents dans l'usage.

Comme en écho à cette renégociation des espaces documentaires dans laquelle les discours se dissipent au profit des dispositifs géospatiaux les mettant en scène, dans laquelle seule la cartographie fait – ou en tout cas – semble faire sens, la question doit être posée de savoir ce qui dans cette activité démultipliée d'indexation sociale peut être considéré comme garant d'une stabilité dans l'accès et la description des documents concernés. Pour pouvoir faire sens

: on navigue sur la carte et l'on affiche pour telle ou telle zone géographique, les informations démographiques et immobilières du type : nombre de biens immobiliers, prix de vente moyen, population, répartition homme/femme, répartition ethnique (sic ...), etc. En fonction de la zone ou ville choisie, les données sont disponibles avec un focus sur 3 couronnes : 1, 3 et 5 Miles.

²² « Nos sociétés ne fonctionnent plus à base de codes et de territorialités, mais au contraire sur fond d'un décodage et d'une déterritorialisation massive. »

²³ <http://www.wikipedia.fr>

²⁴ Notons enfin à titre anecdotique qu'un nouveau vocable fit une fugitive apparition sur le web pour désigner la prise en main par chacun de ces données de nature géographiques : folksonomologies : "ad hoc structures that can be built and span across different locations".

(<http://tecfu.unige.ch/perso/staf/nova/blog/2005/10/05/folksonomologies-and-spatial-technologies/>)

à l'échelle des problématiques du document, cette activité doit faire la preuve d'une régularité observable et quantifiable.

C'est dans cette optique que s'inscrit l'étude de (GOL 05) qui analyse deux corpus de données issues précisément du système de signets collaboratifs annotés baptisé Del.icio.us. Le premier de ces corpus analyse les tags déposés sur un ensemble (set) de 212 URL eux-mêmes partagés dans 19422 signets. Le second ensemble comprend cette fois une liste aléatoire de 68 668 signets. D'après ses résultats, les apparentes et effectives faiblesses documentaires des systèmes de « tags » (non prise en compte des relations sémantiques, aspect non hiérarchique) leur permettent néanmoins de mettre en place des motifs récurrents dans la manière dont les usagers de tels systèmes décrivent à l'aide des mêmes « tags » des sites donnés. Les régularités apparaissent indépendamment du type de tag « employé », que celui-ci soit descriptif, orienté action (« to read »), ou relatif à une perception personnelle (« funny »). Plus exactement les conclusions attestent de :

« (...) regularities in user activity, tag frequencies, kinds of tags used, bursts of popularity in bookmarking and a remarkable stability in the relative proportions of tags within a given URL » (GOL 05).

Ainsi donc, dans une perspective cybernétique, c'est l'usage et la mise en partage qui font fonction de régulation (feedback) et permettent d'atténuer les lacunes de ces « tags » (polysémie, synonymie, niveaux de discours différents, etc...) jusqu'à dans certains cas pouvoir proposer des descriptions homogènes, ou en tout cas reposant sur un ensemble de tags composant eux-même une matrice homogène. Observation et mesure scientifique par ailleurs confirmée dans une entrevue d'Adam Bosworth, Vice-Président en charge de l'ingénierie de la firme Google et dans laquelle il indique :

« La sagesse des foules (sic) fonctionne étonnamment bien. Les systèmes qui marchent sur le web fonctionnent du bas vers le haut ('bottom-up'). (...) Par exemple Flickr ne dit pas à ses utilisateurs quel tag utiliser pour leurs photos. Loin de là. N'importe qui peut déposer n'importe quel tag sur n'importe quelle photo. Mais - et c'est la clé - Flickr offre un retour sur les tags les plus utilisés et les plus populaires, et les gens souhaitant attirer l'attention sur leurs photos (...) apprennent rapidement à utiliser ce lexique si celui-ci fait sens. Cela rend le système étonnamment stable. Del.icio.us fait la même chose. Le succès de Google pour rendre les recherches plus pertinentes reposait sur la puissance de cette sagesse populaire (PageRank)... »²⁵

C'est donc ici un principe d'auto-régulation de nos perceptions documentaires qui semble se mettre en place. Ledit principe s'ancre lui-même dans ce qui fût décrit par Lévy comme le troisième des 6 principes de l'hypertexte :

« (...) multiplicité, emboîtement des échelles (« l'hypertexte s'organise sur un mode « fractal », c'est-à-dire que n'importe quel nœud ou n'importe quel lien, à l'analyse, peut lui-même se révéler composé de tout un réseau (...)) » (LEV 90 p.30-31)

A l'échelle de l'hypertexte planétaire (« docuverse ») comme à l'échelle des processus de nano-publication ('nanopublishing') qui caractérisent les documents supportant ces systèmes de « tags »²⁶, les modes documentaires dominants relèvent d'un principe d'organisation

²⁵ <http://acmqueue.com/modules.php?name=Content&pa=showpage&pid=337>

²⁶ Un document peut ainsi désigner les « unités documentaires » suivantes : une photo sur le service Flickr (<http://www.flickr.com>), un signet ou une liste de signets sur le service del.icio.us (<http://www.del.icio.us>), des sons (<http://freesound.iaa.upf.edu/>) ou en poussant à l'extrême le raisonnement l'adresse physique ou virtuelle et les données associées ou générées dynamiquement déposées sur une carte Google Maps (<http://maps.google.com>) ou Frappr (<http://www.frappr.com>)

fractal que l'on pourrait ainsi résumer (pour les tags) : il existe pour tout élément donné (texte, image, document) une série de mots et termes composant le plus petit lexique commun permettant de décrire l'objet ou le document.

A condition que l'observation de ces régularités et de ces motifs (patterns) soit confirmé par d'autres études, à condition également que la généralisation de ces systèmes de tags soit massive et qu'elle puisse se trouver centralisée sur un nombre restreint de services, à condition, enfin, que les communautés d'intérêt ainsi constituées puissent être perméables à d'autres selon certains critères, sur certaines requêtes ou selon certains « tags », à toutes ces conditions donc, ces folksonomies pourraient constituer un point de fuite complémentaire au web sémantique en cohabitant avec d'autres ontologies formelles. Ces deux volets (folksonomies à valider par l'usage et ontologies pour lesquelles la validation peut être de nature simplement formelle) s'inscrivent clairement dans la lignée des travaux définitoires du web socio-sémantique (ZAC 05)

CONCLUSION. La nouvelle donne documentaire : vers un nouveau big bang ?

Cet article exploratoire nous a permis d'esquisser les contours d'une approche documentaire en quête de positionnement dans ce nouvel espace du tout numérique en lien avec les travaux du collectif RTP-DOC et du web socio-sémantique. Nous voulons, pour terminer, en rappeler les enjeux.

Tout d'abord l'observation de l'émergence d'une nouvelle typologie documentaire, qui, en sus des documents dits primaires et secondaires, propose une hybridation de l'espace et des mécanismes d'inscription et de navigation. Cette étape est celle d'une tertiarisation documentaire.

Ensuite, et en écho à cette nouvelle tectonique, à ce nouveau maillage des unités documentaires qui sont dès à présent rassemblées au sein d'un espace unique « d'indexabilité », se pose, pour l'utilisateur, la question de la confusion des pratiques (désorientation de l'usager) : les mêmes outils permettent désormais sur d'immenses corpus hétérogènes (mails, documents privés, web public, images ...) d'effectuer des opérations cognitives très clairement distinctes : annoter, indexer, rechercher, partager, classer, s'orienter.

Pour ce qui est de l'activité dédiée aux « tags », la question posée est de savoir jusqu'où la stabilité observée dans les usages a ou non capacité à se propager à toutes les sphères informationnelles. Ainsi, très récemment, Gerry Mc Kiernann, en référence à l'initiative « Collib »²⁷ diffusait un message sur des listes de diffusion anglo-saxonnes pour dresser un inventaire des meilleures pratiques en terme d'indexation sociale dans le cadre d'archives ouvertes.

A l'échelle du volume des unités documentaires disponibles, le seuil critique au delà duquel le nombre de producteurs dépasserait celui de lecteurs est en passe d'être atteint, notamment pour cette partie du web que constitue la blogosphère²⁸. Il faut donc s'interroger sur les implications de ce phénomène pour les modèles traditionnels de diffusion et de réception.

Enfin, la préemption de cette masse documentaire, inédite dans son volume et dans sa nature (tertiaire), par un public « à l'échelle du corpus » amène à une renégociation des espaces et des typologies documentaires classiques. Les logiques ici à l'œuvre semblent être celles de désintermédiations et de réintermédiations successives. Ce type de boucles nécessite pour la communauté scientifique de continuer à s'interroger sur la nature de plusieurs processus

²⁷ <http://collib.info/>

²⁸ voir à ce sujet les billets (<http://www.sifry.com/alerts/archives/000298.html>) diffusés par Dave Sifry, directeur général du moteur Technorati (www.technorati.com)

afférents : processus de description documentaire, processus de validation²⁹, nature des objets documentaire pouvant être concernés, pérennité desdits objets.

Ces quelques pistes exploratoires paraissent engager, et ce n'est pas là la moindre de leur complexité, la redéfinition d'une série de modèles culturels aujourd'hui en lutte avec d'autres modèles, marchands cette fois. Le nouveau « big bang » documentaire évoqué dans le titre de cet article renvoie en effet aux débats actuels sur la nature de l'encyclopédisme (wikipedia) ainsi que sur la confrontation du modèle prônant un accès raisonné aux connaissances (modèle bibliothéconomique) face au modèle dérégulé d'accès, de mise à disposition et de marchandisation de biens culturels et patrimoniaux que les récents projets du moteur Google³⁰ ont porté à la connaissance de tous.

BIBLIOGRAPHIE

(BAT 77) - Bateson G., *Vers une écologie de l'esprit*, T. 1. Paris, Seuil, 1977.

(BER 89) - Berners Lee, Tim. « Hypertext and the CERN », 1989. En ligne :

<http://www.w3.org/History/1989/proposal.html>

(FOU 94) - Foucault M., *Dits et écrits - 1954-88*. Tome I (1954-69). Paris, Gallimard, 1994.

(GOL 05) - Golder Scott A. & Huberman Bernardo A., « The Structure of Collaborative Tagging Systems »,

Information Dynamics Lab, HP Labs , 2005. En ligne : <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>

(LEV 90) - Lévy P., *Les technologies de l'intelligence - L'avenir de la pensée à l'ère informatique*. Paris, La Découverte, 1990.

(RTP 03) - Roger T. Pédaque, « Le document : forme, signe et medium les reformulations du numérique »,

STIC-CNRS - Working paper - 8 juillet 2003. En ligne : http://archivesic.ccsd.cnrs.fr/sic_00000511.html

(SIN 05) - Sinha R., « A cognitive analysis of tagging. », billet posté le 27 Septembre 2005. En ligne :

http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html

(TIS 01) - Tisseron S., *L'intimité surexposée*, Paris, Editions Ramsay , 2001.

(ZAC 05)- Zacklad Manuel, « Introduction aux ontologies sémiotiques dans le Web Socio Sémantique ».

Communication, Ingénierie des Connaissances 2005. 03 juin 2005. En ligne :

http://archivesic.ccsd.cnrs.fr/sic_00001479.html

²⁹ voir notamment le fil de discussion intitulé « le web comme hégémonie de l'amateurisme, ou Wikipedia sous les feux croisés » sur les archives de la liste biblio-fr (<http://listes.cru.fr/www/arc/biblio-fr>)

³⁰ <http://print.google.com>