Indexation collaborative : traces de lecture et constitution de communautés

Evelyne Broudoux

Indexation collaborative : traces de lecture et constitution de communautés

L'indexation collaborative ou « bookmarking social » est un service du « web 2.0 » qui, avec ses listes de mots-clés générés par les usagers et publiées en ligne sous la forme de « folksonomies »[1], a suscité l'intérêt de la recherche universitaire, des éditeurs scientifiques comme des bibliothèques, centres d'archives et musées, dès l'apparition des premiers services web dédiés au marquage de références (tagging) en 2004.
Un horizon s'est ouvert de connexion possible entre ces listes et les vocabulaires contrôlés, thésaurus, ontologies ou autres formes d'organisation hiérarchisée des connaissances. De nombreux travaux de recherches, engageant des équipes plurisdisciplinaires, ont investi deux champs en particulier [Broudoux, 2012] :

- d'une part, la conception d'algorithmes visant à améliorer le processus de marquage, la recherche d'informations, la connexion de personnes partageant les mêmes intérêts, ou encore la modélisation de branchements automatisables des folksonomies avec des thésaurus,
- d'autre part, l'intégration de dispositifs visant à recueillir les mots-clés des usagers dans les catalogues de bibliothèques, autour des expositions thématiques organisées par des musées ou l'enrichissement de collections dans des centres d'archives ; de nombreux projets socioculturels ont également vu le jour sur cette thématique.

Cet article se propose de s'intéresser plutôt aux usages qu'aux dispositifs du « bookmarking social » en analysant deux types de pratiques.
Le premier concerne les folksonomies créées par ceux qui thésaurisent des articles de recherche sur des gestionnaires dédiés de références en ligne [Besagni et al., 2012]. S'intéresser à ces traces de lecture d'articles permet, d'une part, de dégager les tendances (domaines de recherche) des projets en cours et, d'autre part, de disposer d'un complément qualitatif aux mesures de citation effectuées par les éditeurs de revues scientifiques.
Le second concerne la participation des usagers aux dispositifs collaboratifs initiés par les professionnels des bibliothèques, des centres d'archives, de documentation ou de musées. Quelle appropriation est faite de ces services 2.0 ? Quel apport aux dispositifs classiques de repérage de l'information peut-on identifier ?

Une représentation scientifique à travers l'activité des groupes de CiteUlike

Ce volet s'intéresse à la première génération de gestionnaires de références bibliographiques en ligne. Véritablement horizontaux dans leur ouverture, s'adressant à toute personne soucieuse de rassembler des références scientifiques en utilisant des pots communs, ces services imaginés par des universitaires (CiteUlike, Bibsonomy) dès 2004, rapidement adoptés par les éditeurs scientifiques (Springer pour CiteUlike puis Connotea développé par Nature, 2Collab par Elzevier) ont vu leurs fonctionnalités de tagging, recommandations automatisées, etc., intégrées dans des dispositifs qui aujourd'hui agrègent aussi des réseaux sociaux (ex : Mendeley, Zotero).

[1] « La folksonomie, concept basé sur l'auto-indexation, est l'affectation par l'auteur ou l'internaute de mots-clés par balisage (tagging) à un contenu caractérisant l'information ». [Broudoux et al. 2007, p. 191].

Une étude concernant la représentation des domaines et l'activité des communautés a permis de révéler les modalités de constitution des groupes et l'orientation scientifique des domaines représentés. En 2010, nous avons généré une carte de l'organisation thématique des groupes avec l'outil Neurodoc - au sein de l'Inist - qui applique la méthode de classification des k-means axiales (Lelu & François (1992) et Grivel & François (1995)). Initialement conçu pour traiter des notices bibliographiques, l'outil a été détourné en faisant correspondre le nom du groupe au titre du document et les tags aux mots-clés documentaires. Son application a été réalisée sur les groupes de CiteUlike et l'index de leurs tags associés en retenant uniquement les tags présents dans au moins cinq groupes. Ce filtrage a donc sélectionné 2 216 groupes[2] regroupés en vingt classes selon leur proximité thématique ; les intitulés des classes correspondent aux tags de poids le plus élevé pour chaque classe.
Nous retiendrons ici, pour les besoins de l'article, uniquement les cinq classes identifiant des domaines de recherche aux contours certains : *epidemiology*, *evolution*, *pliocene*, *information* et *nrf2*, qui est une protéine de liaison de l'ADN.

Remarquons que la représentation des domaines se divise en deux figures de concentration de références : soit un groupe fait converger plusieurs disciplines autour d'un thème central de recherche comme *bio-informatics* ou *epidemiology* ou *information,* soit un groupe se consacre à une thématique granulaire comme *C.elegans* en biologie moléculaire ou *nrf2* en épidémiologie.
Les domaines les mieux représentés dans les groupes sont ceux des sciences de la vie et de la terre :

- Pour les sciences de la vie, la classe *evolution* est la classe la plus active avec 777 tags et 241 groupes pour la plupart formellement identifiés. Les articles enregistrés régulièrement font ressortir une convergence de plusieurs disciplines autour de la bio-informatique. Le groupe comportant les références les plus nombreuses (27 742[3]) est nommé *C. elegans/WormBase* et étudie les nématodes (vers ronds). La production de ce groupe, créé en 2008, est caractéristique du devenir des entités sociales sur CiteUlike : la quasi-totalité des références a été listée par les deux premiers membres en un peu plus d'un an. En 2009, un troisième membre s'est inscrit et a ajouté une seule référence ; depuis, onze personnes se sont ajoutées mais seulement cinq sont du domaine, le reste relevant d'inscriptions « tests » ou de spams. *Wormbase* est un consortium international de biologistes et d'informaticiens qui tient à jour une base de données en accès libre dédiée plus particulièrement au *Caenorhabditis elegans*, un nématode utilisé comme modèle animal en biologie moléculaire, dans l'étude du développement embryonnaire et de l'apoptose (mort programmée des cellules).
- Pour les sciences de la terre, les premiers groupes de la classe *pliocene* traduisent bien les thèmes qui préoccupent les tagueurs : l'un concerne le changement climatique, l'autre la conservation de la biodiversité, un autre la géomorphologie tectonique, et le dernier est une bibliothèque des arbres à palmes. La classe se compose de 3 117 tags et de 85 groupes. Biodiversity_conservation (dont la description comporte *biodiversity conservation*, *conservation biology*, *conservation policy*) avec ses 6 477 articles et ses 79 inscrits est l'exemple typique d'un groupe

[2] Les statistiques chiffrées sont tirées des extractions réalisées sur CiteUlike en septembre 2009 : 341 498 articles étaient référencés par 68 522 tags différents.
[3] A la date du 17 janvier 2012.

actif de CiteUlike puisqu'il continue de gagner en nouveaux membres tout en veillant à ne pas se laisser envahir par les spammeurs (liste de membres exclus).

Les groupes pluri-disciplinaires concernent la classe *epidemiology* constituée par 177 tags et 43 groupes représentant deux dominantes d'intérêts. L'un regroupe des items médicaux (geriatrics, endocrinology, thyroïd, renal, vascular, healthcare, fiber, adrenal) alors que l'autre concentre spécifiquement les thématiques liées à l'environnement et au traitement de l'eau (hydrologie, groundwater, infiltration, climate change, microfluidics, sedimentation, geomorphology). A noter que la classe *nrf2* est composée de 1 459 tags et 98 groupes dont 70 sont des émanations de l'HEIRS[4], une organisation nord-américaine indépendante d'éducation à la santé, dont l'objectif est de fournir des informations issues de la recherche et des ressources sur les maladies liées à la détérioration de l'environnement.

La classe *information* et ses principaux tags (ontology, folksonomy, tagging, usability, trust, sociology, data mining, architecture, retrieval) rassemble 179 groupes dont *Philosophy_of_information* qui mérite d'être retenu, en tant que groupe actif depuis 2004 avec 235 membres, concentrant des références qui constituent un véritable cadre conceptuel visant à comprendre et interpréter la notion d'information.

Remarquons les enjeux sociaux qui se dégagent des traces de lecture : théories de l'évolution, changement climatique, épidémiologie liée à la détérioration environnementale, biodiversité, sciences environnementales, communautés coopératives, qui révèlent bien les tendances fortes des préoccupations des inscrits sur CiteULike.

L'activité des groupes de travail, formels et informels, publics ou privés, est remarquable sur CiteUlike. Bien que subissant un fléchissement fin 2009, 10 453 membres se distribuent à cette époque entre 2 871 groupes comportant de un à plusieurs centaines de membres[5] : un groupe sur deux est constitué par un seul membre et est donc soit un groupe « test » sans référence soit une entité individuelle qui réalise une bibliographie thématique et un groupe sur deux est fermé, ce qui nécessite l'autorisation de son créateur pour visualiser les références.

L'implication individuelle y semble forte : un créateur de groupes est souvent partie prenante dans plusieurs autres groupes, qu'il peut lui-même avoir créés. Peu de groupes semblent bénéficier d'un processus spontané d'apport régulier de références ; la vie des groupes est au contraire dépendante d' « animateurs » tagueurs qui en relancent régulièrement l'activité. Plus les usagers participent ou sont identifiés à une thématique précise, plus ils renseignent leur profil ; cette sortie de l'anonymat correspond aussi à l'institutionnalisation des groupes, repérables par des acronymes d'équipes de recherche, de groupes projets ou de laboratoires.

Le « tagging » est-il vraiment social ?

La seconde partie de cet article rend compte de plusieurs retours d'expériences dans lesquelles les usagers sont placés en situation de taguer des contenus.

Dès 2008, une enquête réalisée sur une formation « tout au long de la vie » (Saeed, Yang, 2008) a révélé que le « bookmarking social » était moins utilisé que le podcasting et le

[4] Health Education Information and Resource Services (HEIRS).

[5] A la date du 22 janvier, 5 804 214 articles sont référencés sur CiteUlike.

blogging, bien que ces éléments aient été combinés en un agencement d'éléments d'apprentissage à distance. Alors que l'unité d'enseignement concernée était spécialisée dans la programmation Web, le public ne s'est pas approprié la fonctionnalité, considérée apparemment comme superflue.
Signalons que le moissonnage des tags pour enrichir les collections de métadonnées est devenu une pratique recherchée : HarVANA (Harvesting and Aggregating Networked Annotations) répond à cet objectif en proposant un outil fusionnant annotations et tags (représentés en RDF) avec les indexations issues de vocabulaires contrôlés. Les annotations-tags provenant de différents serveurs distribués sont moissonnés par OAI-PMH puis agrégées, avec les métadonnées documentaires issues de différents répertoires institutionnels, dans un dépôt centralisé de métadonnées. Cette approche interopérable et extensible permet aux bibliothèques, archives, musées et répertoires de se connecter sur les communautés de tagueurs et d'augmenter leurs métadonnées tout en ouvrant leurs services (Hunter, Khan, Gerber, 2008). Le corpus d'HaVANA est constitué par la copie d'une base de données d'images (PictureAustralia[6]) stockée sur un serveur de l'Université de Queensland et interfacée avec la bibliothèque nationale d'Australie. Cette initiative a montré que la consultation du catalogue par les usagers était enrichie par les annotations entrées par les usagers « experts » du domaine, authentifiés par le système, chargés d'annoter et de taguer les images, selon une ontologie « allégée » pour la circonstance [Hunter J, 2008].

Les interfaces d'accès aux collections muséales ayant intégré tagging et blogging ont fait également l'objet d'études d'usages, comme celles réalisées par (Srinivasan et al., 2009) au Musée d'archéologie et d'anthropologie de l'Université de Cambridge. Deux types de public ont ainsi participé à *Blobgects* qui allie au catalogue traditionnel du musée un catalogue augmenté par un système de bookmarking et de blogging : un groupe d'étudiants de niveau Master du département *d'Information Studies* de l'UCLA (Université de Californie, Los Angeles) et un groupe d'étudiants Inuit du lycée Inukshuk d'Iqaluit, dans le territoire Nunavut. Les deux groupes sont considérés comme experts du domaine, l'un parce qu'il est intéressé par les nouveaux modes de partage des objets culturels avec le public, l'autre parce qu'il entretient des liens culturels avec les objets présentés en ligne, issus de sa propre communauté. Chaque groupe a été divisé en deux : un groupe expérimental et un groupe de contrôle qui interagissait l'un avec le catalogue augmenté, l'autre avec le catalogue original mais dont le design ressemblait à l'interface augmentée de *Blobgects*.
L'objectif principal était de vérifier si l'accès aux entrées du catalogue était facilité par le tagging et le blogging, considérés comme des moyens d'interaction supplémentaire avec le contenu. En dehors d'un déficit d'informations contextuelles notoire concernant les collections présentées et des difficultés d'appropriation de l'interface *Blobgects*, l'étude a indiqué que tagging et blogging n'ont pas réussi à engager le public plus avant, dans un catalogue aux métadonnées très spécialisées. Seule une mise en narration du contenu a permis son appropriation, d'où l'importance des tags à caractère narratif [Srinivasan Ramesh et al., 2009]. Dans ce cas précis, le tag n'est plus considéré comme aidant à constituer une description exhaustive d'un objet mais comme une trace de conversations entre usagers, la conduite d'une interaction entre internautes et objets, dans l'épaisseur de leurs différents contextes : historique, usages, etc.

---

[6] Ce projet qui vise à collectionner des images de la vie culturelle, économique et sociale de l'Australie est le fruit d'une collaboration entre la Bibliothèque nationale, le Musée national, les Archives nationales et les agences culturelles de l'Australie. http://www.pictureaustralia.org/

Le rôle des communautés « amateurs » dans la production de métadonnées sociales

L'intégration du tagging dans des portails de type web 2.0 permet, comme dans le cas décrit de CiteULike, la construction de réseaux sociaux scientifiques et la création de communautés d'intérêts interuniversitaires et interdisciplinaires. Mais l'exemple du portail MSI-CIEC[7], lancé en 2008, qui apparaît comme une coquille vide aujourd'hui, confirme que, sans interactions entre usagers, sans inscription réellement sociale dans la vie quotidienne - fût-elle scientifique, l'initiative reste à l'état exploratoire et ne s'ancre pas dans les pratiques.

Dans le cadre d'un programme d'enquête sur les apports de l'intégration des « métadonnées sociales » aux dispositifs de communication des bibliothèques, archives, musées, OCLC a publié fin 2011 les résultats d'une analyse des usages de métadonnées générées par les usagers au sein de 76 sites, réalisée par 21 membres des pays partenaires du projet (Etats-Unis, Pays-Bas, Australie, Nouvelle-Zélande et Royaume-Uni)[8].
L'objectif était d'établir un état de l'art des initiatives qui valorisent les collections en s'appuyant sur l'expertise de leur public pour enrichir les métadonnées, ce dernier terme étant d'ailleurs à entendre dans un sens très large puisqu'intégrant tagging mais aussi commentaires et recommandations, notations, etc. Deux cas de figures sont étudiés : soit les services de type web 2.0 sont intégrés aux dispositifs gérés par les bibliothèques, archives ou musées (catalogues, services web, sites d'expositions, etc.), soit les services sont implantés au sein de « plates-formes tiers » (Wikipédia, YouTube, Flickr, LibraryThing, etc.).

Deux objectifs principaux à la production de « métadonnées sociales » sont évalués en particulier :

- améliorer les métadonnées générées par les bibliothèques, archives et musées, de manière à accroître la qualité et la pertinence des résultats de recherche dans les catalogues et sur les moteurs de recherche ;
- contextualiser les contenus, de manière à faciliter leur compréhension et leur utilisation.

Les métadonnées produites par les usagers ont été rangées en sept catégories : métadonnées pour la description, métadonnées pour l'accès, marquage (tagging), construction des collections et des contenus, notations et appréciations, partage et facilitation de la recherche, travail et construction de communautés, promotion des activités hors site.

Toutes les interfaces d'interrogation des catalogues étudiés intègrent la fonction de tagging, qui permet aux usagers d'ajouter leur propre terminologie en complément des vocabulaires contrôlés.

---

[7] http://gf14.ucs.indiana.edu:8080/ MSI-CIEC est une collaboration entre l'American Indian Higher Education Consortium (AIHEC), l'Hispanic Association of Colleges and Universities (HACU), et la National Association for Equal Opportunity in Higher Education (NAFEO).
[8] Référence à préciser ?

L'exemple emblématique, abondamment cité, est celui de la bibliothèque de l'Université de Pennsylvanie qui est l'une des premières à avoir franchi le pas en 2005 en autorisant un enrichissement parallèle à son catalogue en ligne. P*ennTags* n'a – selon les auteurs de l'étude - pas encore réussi à construire sa propre communauté d'usagers, peut-être à cause de l'absence d'une campagne marketing incitative. Une expérimentation menée par une bibliothécaire et un professeur dans le cadre d'une bibliographie commentée sur l'Histoire du cinéma, copyright et culture a montré les limites de ce que les étudiants ont ressenti comme une contrainte supplémentaire à leur devoir. Bien qu'une bibliothécaire ait investi le cours et appris aux étudiants à se servir du dispositif, ceux-ci copiaient-collaient leurs ressources dans un fichier Word pour ensuite les réintroduire dans *PennTags* à la dernière minute afin de satisfaire la consigne.

De façon similaire, le projet *Mtagger*, lancé en février 2008 à la bibliothèque universitaire du Michigan, apparaît avoir suscité peu d'intérêt de la part des interviewés lors de l'enquête menée. Cependant, après 18 mois de vie, quelques 2 400 usagers avaient uniquement tagués 11 500 items avec près de 5 000 tags.

Inversement, LibraryThing[9] paraît réussir le pari des communautés avec ses 65 millions de tags ajoutés. Dans un mail du 25 mars 2009, le fondateur de LibraryThing estimait qu'il y avait plus de tags ajoutés en un jour sur LibraryThing que dans la totalité de *PennTags*. Rappelons que ce n'est qu'après un contrôle qualité et donc un filtrage que *LibraryThing for libraries* (LTFL) autorise leur ajout aux clients OPAC[10].

Autre projet lancé en 2008 pour une durée de trois ans, le *Steve Museum Social Tagging Project* qui rassemblait onze musées participants. La première phase a concerné 1 782 œuvres d'arts proposées par 2 017 usagers qui ont ajouté un total de 36 981 tags ; les professionnels des musées ont ensuite révisé les tags et ont considéré que 88% d'entre eux étaient utiles. Remarquons que la motivation de participation de la majorité des participants interrogés était d' « aider l'organisation », conformément à l'appel à participation sur la page d'accueil du projet.

C'est bien dans la construction des collections et des contenus que le marquage (tagging) et ses apports sont susceptibles de tenir toutes ses promesses :

- Apport de nouveaux matériels à une collection déjà organisée (ex : photographies).
- Amélioration de la description des contenus existants :
    - Identification de la source, du sujet et de l'année de publication de photographies anonymes, posters et autres documents.
    - Ajout de commentaires ou corrections de commentaires aux photographies, année de naissance et de décès, recensements, et autres objets.
    - Ajout de contributions non textuelles (téléchargement d'une image d'une page manquante d'un livre rare).

[9] à présenter succintement ?

[10] La plupart des Opacs (gestionnaires de catalogues en ligne) ont intégré les fonctionnalités des nuages de mots-clés générés par les usagers. L'intégration des tags de LibraryThing par les Opacs vise à diversifier l'accès aux documents des bibliothèques.

C'est ainsi que le projet *Paris-Normandie* de Patrick Peccatte et Michel Le Querrec, consistant à améliorer la description documentaire d'un fonds de 3 044 photographies historiques sur la bataille de Normandie (du 6 juin à la fin août 1944), est porté sur la plate-forme grand public Flickr depuis janvier 2007. La communauté d'une quarantaine de contributeurs (dont une quinzaine de membres « permanents » spécialisés) possède des compétences complémentaires (historiens, archivistes, documentalistes, enseignants, etc.) et connaît très bien l'histoire de la bataille de Normandie dont la plupart est originaire.
Le processus d'amélioration du contenu est différent de celui d'une accumulation de tags : lorsqu'une discussion s'établit entre les participants, elle s'achève par une validation collective des modifications proposées. Les rares désaccords se produisant lors des échanges concernent uniquement le choix des termes. Une seule personne – le co-initiateur du projet – se charge alors de la rédaction finale.
Le bilan documentaire réalisé par Patrick Peccatte[11] tient compte de plus de 6 900 contributions corrigées, complétées, mises à jour. La galerie reçoit près de 3 000 visites par jour et les 3 044 photos ont été vues plus de 8,3 millions de fois depuis la fin janvier 2007.

***

En conclusion, la constitution de communautés d'amateurs et de groupes de travail (pour CiteUlike) apparaît déterminante pour la vie des plate-formes de tagging et la réussite des projets. Car les résultats sont pauvres en termes d'usages lorsque les outils sont proposés sans que les acteurs puissent y investir leurs motivations.

Bibliographie

Besagni Dominique, Broudoux Evelyne, Fabry Cécilia, François Claire, Roussel Clotilde. Références scientifiques en ligne : folksonomies et activité des groupes. Conférence Isko-France des 27 et 28 juin 2011 à Lille, Hermès-Sciences (en cours de publication).

Broudoux Evelyne « Folksonomies et indexation collaborative. Rôle des réseaux sociaux dans la fabrique de l'information ». DocForum 24 nov. 2006. (Document en ligne sur <DocForum.
http://www.docforum.tm.fr/documents/23&24nov06SavResPar06InterBroudouxE.pdf
>)

Broudoux Evelyne et al. Auctorialité : production, réception et publication de documents numériques in La redocumentarisation du monde (Roger T. Pédauque). Cépaduès, janvier 2007.

Broudoux Evelyne. Tagging : modélisation, technique, usages (article en cours d'évaluation), 2012.

Grivel Luc, François Claire. « Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et

[11] Journée Fulbi du 18 janvier dernier : http://www.fulbi.fr/?q=content/2011.

technique ». *Solaris* n°2, 1995. Document en ligne sur http://biblio-fr.info.unicaen.fr/bnum/jelec /Solaris/d02/2grivel.html

Hunter Jane, Khan Imran, Gerber Anna. « Harvana: harvesting community tags to enrich collection metadata » in JCDL '08 : Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (2008), pp. 147-156.

Saeed N., Yang Y. « Incorporating blogs, social bookmarks, and podcasts into unit teaching »
in ACE '08: Proceedings of the tenth conference on Australasian computing education (2008), pp. 113-118.

Smith-Yoshimura, Karen and Cyndi Shein. 2011. Social Metadata for Libraries, Archives and Museums Part 1: Site Reviews. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-02.pdf

Srinivasan Ramesh et al. « Blobgects: Digital museum catalogs and diverse user communities » in Journal of the American Society for Information Science and Technology, Vol. 9999, No. 9999. (2009), NA.